When a Few Words Are Not Enough: Improving Text Classification Through Contextual Information

Nujud Aloshban

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Engineering
College of Science and Engineering
University of Glasgow

May 2021

Abstract

Traditional text classification approaches may be ineffective when applied to texts with insufficient or limited number of words due to brevity of text and sparsity of feature space. The lack of contextual information can make texts ambiguous; hence, text classification approaches relying solely on words may not properly capture the critical features of a real-world problem. One of the popular approaches to overcoming this problem is to enrich texts with additional domain-specific features. Thus, this thesis shows how it can be done in two real-world problems in which text information alone is insufficient for classification. While one problem is depression detection based on the automatic analysis of clinical interviews, another problem is detecting fake online news.

Depression profoundly affects how people behave, perceive, and interact. Language reveals our ideas, moods, feelings, beliefs, behaviours and personalities. However, because of inherent variations in the speech system, no single cue is sufficiently discriminative as a sign of depression on its own. This means that language alone may not be adequate for understanding a person's mental characteristics and states. Therefore, adding contextual information can properly represent the critical features of texts. Speech includes both linguistic content (what people say) and acoustic aspects (how words are said), which provide important clues about the speaker's emotional, physiological and mental characteristics. Therefore, we study the possibility of effectively detecting depression using unobtrusive and inexpensive technologies based on the automatic analysis of language (what you say) and speech (how you say it).

For fake news detection, people seem to use their cognitive abilities to hide information, which induces behavioural change, thereby changing their writing style and word choices. Therefore, the spread of false claims has polluted the web. However, the claims are relatively short and include limited content. Thus, capturing only text features of the claims will not

provide sufficient information to detect deceptive claims. Evidence articles can help support the factual claim by representing the central content of the claim more authentically. Therefore, we propose an automated credibility assessment approach based on linguistic analysis of the claim and its evidence articles.

Contents

List of Tables

List of Figures

Acknowledgements

It has been a lengthy and arduous process to complete this thesis, however, it has been full of adventures, struggles, and learning opportunities. Not only about Affective Computing, but also about myself, my skills, and how I communicate with others. This thesis could not have been completed without the help and guidance of a variety of individuals, many of whom contributed in some way to the final product.

First and foremost, I would like to express my sincere gratitude to my supervisor Professor Alessandro Vinciarelli for his patience, motivation, great suggestions, absolute support and immense knowledge. He also taught me many of the statistical concepts that I put to good use in this work, and taught me new ways to look at the world. I could not have imagined having better advisors and mentors. I would also like to acknowledge Dr.Giorgio Roffo for insightful comments on technical parts.

I must also offer my gratitude to my colleagues as they are important part of this intellectual adventure. With a special mention to Xin Xin, Fatma Elsafoury and David Martin Maxwell who really helped me to "learn a trade" when I first started my PhD. I would extend my gratitude to Abeer Buker and Bashayr Aldawsari for playing such an important and inspirational role along the way.

Last but not the least, I would like to thank my parents for their unconditional support, love, and their believe in me. They have been the greatest inspiration and motivation for all the success I have achieved and endured difficult times in my life. I would extend my gratitude to my great brother, Dr.Yazeed Alashban, for the moral support during the hardest times, I cannot express in words how much I relied on you and how much you contributed to this thesis. Also, my deepest gratitude to my dearest husband, Ahmed Alkhathlan, for the endless love, unconditional understanding and support provided during these years. Without his infinite understanding, supports and encouragement, everything would not be easier for

me. I specially dedicate this thesis to my dearest children, I would not have been to this far without your love and patience.

Chapter 1

Introduction

1.1 Motivations

Over the last few decades, the number of computer users has significantly increased, especially due to the popularisation of the Internet and the possibility of automating office processes using dedicated software [214]. This induces a continuing growth in the amount of computer readable text produced, stored and handled, although the text information remains widely in hard copy format. Over the last twenty years, 'word processing' software of some form has been used to produce nearly most of the printed text information globally. Although text is an extremely rich source of information, extracting insights from text can be challenging and time consuming due to its unstructured nature. For example, 80% of the entity data, including person, place or thing, is provided only in unstructured form, such as email, views, news or interviews [151]. This unstructured data is considered a problem in most areas of data-intensive applications—business, universities and research institutions [151]. While it would be impossible to manually analyse these data, text analytics has become increasingly popular to automate this process.

Text analytics analyses the hidden relationships between entities to discover meaningful patterns that reflect the knowledge contained in the dataset. This knowledge is utilised in decision-making [48]. Text analytics typically employs various methodologies to process the text: one of the most important is natural language processing (NLP). It applies computational linguistics principles to analyse lexical and linguistic patterns [48]. Text classification, also known as text categorisation, is a classical problem in NLP that aims to assign labels or tags to textual units, such as sentences, queries, paragraphs and documents. It has different

applications, including question answering, sentiment analysis, misinformation detection and depression detection. The widely studied cases of text classification are binary text classification in which a textual text is classified into one of two mutually exclusive categories or classes. In 1960, Hans Peter Luhn [189] utilised the document-frequency method to automatically obtain literature summaries, which is also called the basis of text classification research. In 1970, [275] proposed a vector space model for text representation. Within the 1990s, machine learning became a new trend after the development of statistics, enabling researchers to apply machine learning algorithms to text classification. Recently, the increasing popularity of deep learning has induced the application of advanced methods to text classification.

Computers need a vast amount of common-sense and domain-specific world knowledge to understand natural language [84, 174]. However, the previous studies on semantic relatedness were purely based on a statistical approach that discarded background knowledge [23,87] or on lexical resources that incorporated limited knowledge about the world [49,144]. This is especially true when standard text classification approaches are applied to texts that have insufficient or limited words; thus, it may cause text brevity and feature space sparsity [103]. In this thesis, insufficient or limited words are defined as when text features alone underperform compared to their combination with other sources of data. Compared with paragraphs or documents, texts with a limited number of words are more ambiguous due to the lack of contextual information. Thus, simple text classification approaches based on words only may not properly represent the critical features of texts. One of the efficient solutions to overcome the mentioned problems is to enrich texts by using domain-specific information, which can be called the feature space augmentation method or, more specifically, feature enrichment [103]. In this thesis, we aim to study two different real problems in which their text inputs alone are insufficient for classification. The common solution is to enrich the text features with additional contextual features, and these additional features are based on the problem that we address. We comprehensively analyse and address the two problem scenarios separately.

The problem of insufficient text can occur in any communication, whether it happens face-to-face (F2F) or online. The most important means of human communication in F2F is the clinical interview. F2F clinical interviews are the foundation of all clinical activities in psychotherapy and are typically the first encounter between the mental health professional

and the patient. One of the more productive arenas for exploring text in clinical interview has been in the depression literature. More specifically, we studied depression detection in clinical interviews recorded in three Mental Health Centres (59 interviews). There were difficulties in obtaining enormous data, a problem that is inherent to depression detection due to ethical and practical concerns in recruiting depression patients. To effectively tackle this limited data availability, we segmented the transcription of interviews into clauses, i.e. to manually extract linguistic units that include a noun, a verb and a complement. These segmented clauses therefore have a limited number of words that may not provide sufficient contextual information. This problem of a limited number of words is imperative to study in depression domain because of several reasons. First, the literature provides evidence that depressed individuals tend to engage less in social interactions and, therefore, speak less than people that are unaffected by the pathology [44, 118]. Second, realistic application scenarios require one to tackle recordings that contain only a few words (e.g. the use of data collected at help lines [140]). Finally, when the speech data are obtained through interviews or other forms of interaction that involve medical personnel, reducing the amount of time necessary to gather enough information lowers the costs associated with depression diagnosis.

Major depressive disorder is a mental disease, and over 300 million people suffer from this disease globally [221]. Depression is considered a major cause of suicide and the second primary cause of death among teenagers [222]. Depression cases are increasing with an increase of around 18% between 2005 and 2015 [221]. According to the World Health Organization (WHO), less than half of depressed patients globally (in many countries, fewer than 10%) receive proper depression treatment. It can be difficult for the depressed to attain professional attention due to mobility, cost, motivation and hesitation to report since they are sometimes passive in contacting psychologists or psychiatrists to get treatment. Therefore, it is imperative to develop a computer-aided automatic depression assessment system that supports psychiatrists in the diagnosis of clinical depression and reduces subjective bias.

Depression certainly impacts the way people feel, think, and communicate [21]. Language reveals our ideas, moods, emotions, beliefs, behaviours and personalities [289]. The observed effect of depression on linguistic style is mainly explicated by cognitive mechanisms (e.g. studies in [29, 67]) in which depressed patients reveal increased negative emotions and self-focus. In line with these cognitive models, social integration/disengagement

theories (e.g. study in [95]) also study patterns in which suicidal patients become less socially engaged with community. These underlying mechanisms manifest themselves through language, indicating an increased self-focus, splitting from others and negative emotion [65, 310]. Therefore, studying language to detect and assess human mental health diseases is considered an appropriate mental health modelling. For example, a Russian speech study [298] found a more frequent use of all pronouns and verbs in the past tense among depression patients. This means that patients suffering from depression will reveal linguistic behaviours that vary from those of healthy individuals. Therefore, language reflects the mind [100].

Although studies have shown the strength of predictive factors of linguistic features for the depression status of individuals, no single feature on its own has enough distinctive power as a sign of depression due to the inherent differences in the speaking method [79]. This means that linguistic cues alone may not be sufficient to understand the mental traits and states of the person; thus, information from other modalities needs to be supplemented. Interview reveals the linguistic contents (what people say) and has paralinguistic/acoustic speech (how words are said) that show significant clues about the emotional, neurological and mental features of the speaker. Therefore, the recent speech technologies are suggested for the evaluation, diagnosis and monitoring of different mental disorders that affect the subject's voice [77]. Particularly, depression may induce cognitive and motor changes that affect speech creation, where decreases in verbal activity efficiency, prosodic speech impropriety and monotonous speech have all been revealed to be symptomatic of depression [300]. For example, spectral-based features of depressed people change remarkably in depressive states [228]. Considering the broad clinical outline of depression, it appears that a multimodal approach to identifying depression from collections of linguistic and paralinguistic/acoustic channels of communication yields significant benefits. The first part of this thesis aims to help clinicians and psychiatrists through the development of automatic approaches for identifying people affected by depression based on the automatic analysis of language (what you say) and speech (how you say it).

The other type of communication can happen in online websites. The innovative invention of the World Wide Web has enabled data sharing to the world very easy. People these days completely depend on news from the internet than the classic organisations. For example, a recent study showed that around 68% of U.S. adults get and share news using social media

applications and websites [54]. This explosive growth of the web, including online news and social media, has enabled the delivery of relevant content to the right users based on limited context information and implicit knowledge. Despite being a vast resource of valuable information, the spread of false claims has polluted the web. Therefore, we address the misinformation problem of exploring text on online websites.

According to the World Economic Forum, 'the rapid spread of misinformation online' is one of the top ten greatest challenges facing the world [111]. Recently, this rapid spread has widely emerged on online sites for different commercial and political influences. While this spreading of misinformation (also known as 'Fake News') deceives people to accept false beliefs and change the way they respond to the truth, it breaks the reliability of the entire information ecosystem [296]. During the 2016 U.S. presidential election campaign, misinformation was identified and became a severe risk to journalism, democracy, freedom and the public's trust in governments. The chance to mislead or to be misled increases during news production, dissemination and consumption, thereby necessitating many fact-checking websites, where people research claims, manually assess their credibility and present their verdict along with evidence, such as background articles and quotations [193]. However, human can detect deceptive claims just 4% better than chance based on a meta-analysis in over 200 studies [43]. This problem calls for credibility assessment tools that can automate the verification process of claims.

Individuals seem to employ their cognitive efforts to modify or hide information. This induces changes in behaviour, thereby inducing changes in verbal and written texts. For particular reasons, they attempted to change their writing style and to change their word choices to fabricate individual facts. This contains linguistic feature changes, and one may discover fabricated text by analysing these features. This challenge encourages researchers to consider several ways to detect deceptive texts [253]. Within this framework, writing misleading claims appears to be done by carefully selecting words because words are the richest and most distinguished way to communicate [98]. Also, to maintain 'cohesion' and 'coherence' in their claim, it is based heavily on lexicalisation and complex syntactic structures [55, 126]. Therefore, it produces more linguistic leakage to deception, meaning that linguistic patterns may leak information that people try to hide and indicate the claim's credibility.

Considering the structure or origin of the claim, it is relatively short and contains a very

limited context. Thus, analysing only textual claims will reveal limited clues that probably cannot sufficiently identify deception. Therefore, studies often combine this approach with other auxiliary features to improve detection, such as other linguistic or network analysis techniques (e.g. studies in [106] and [107]). Since any fact can be demonstrated as genuine with supporting evidence, gathering evidence is an ultimate step in assessing the credibility of claims or facts. Evidence articles, also referred to in this thesis as supporting articles, help to support the factual claim by representing the central content of the claim more authentically. The second part of this thesis aims to propose an automated credibility assessment that reduces the burden by assisting humans in verifying the veracity of the claim. More specifically, we linguistically analyse the claim along with its relative evidence articles to determine their opinions regarding the credibility of the input claim.

1.2 Thesis Statement

Binary text classification is becoming important in many problems, such as depression detection and misinformation identification. The classification of texts that include an insufficient or limited number of words is particularly challenging. This thesis asserts that enriching textual data with contextual information (domain-specific information) can help to impact the performance of text classification. Understanding the required contextual information will help build a more effective text classification for a problem. Also, the way how to leverage additional information to text directly influences the performance of text classification problems. Two different application scenarios—depression and misinformation—are studied to explore the effectiveness of leveraging additional information. Overall, the statements set forth by the thesis are as follows:

- **Statement 1**: Developing an objective, effective system that supports psychiatrists in their diagnosis of clinical depression was based on linguistic and acoustic/paralinguistic aspects of speech. We focus on estimating the likelihood that individuals could be considered depressed/non-depressed given their clauses. In this thesis, the clause is defined as a multimodal analysis unit that includes both speech signals and their transcription.

- **Statement 2**: Developing an objective credibility assessment system that reduces the burden by assisting humans in verifying the veracity of the textual claims that are ex-

pressed freely in Internet. The assessment is based on a linguistic analysis of the claims regarding evidence articles. We focus on estimating the likelihood that the claims could be considered credible/not-credible given the claims along with their evidence articles.

1.3 Contributions

This thesis's main contribution is the use of additional domain-specific information (contextual information) to enrich textual data for text classification in different forms of communication. In F2F communication, clinical interviews for depression are studied, while in online communication, misinformation is studied. We contribute a series of approaches to analyse the data in both depression and misinformation domains. More specifically, the work described makes the following contributions in each domain:
The main contributions and novel findings to the field of depression are as follows:

1. **Distinguishing between depressed and non-depressed participants, in the data of this work, was done by psychiatrists and not by administering self-assessment questionnaires.** Half of the participants have been diagnosed with depression by a professional psychiatrist, while the other half, referred to as control participants, have never experienced mental health issues. This is an important advantage because it increases the chances of the data being representative of the actual difference between depressed and non-depressed speakers. Alternatively, it ensures that the problem addressed in the work is depression detection and not the inference of self-assessment scores. This is important because self-assessment questionnaires are subject to multiple biases and, furthermore, the data show that they can be filled out inconsistently, especially by people affected by depression.

2. **Developing an objective, effective system that supports psychiatrists in their diagnosis of clinical depression from linguistic and acoustic aspects of speech.** The experiments show that the approach appears to be in condition to discriminate between cases that are sufficiently clear to be processed automatically and cases that require medical attention, thus allowing the system to potentially reduce by two-thirds the workload of the medical personnel while still keeping the accuracy above 90%.

3. **Structuring the input data by a clause which is a subject, a finite verb and possibly a complement that express part of a speech act such as narrating, explaining or interrupting.** Unlike the other works that utilise entire interviews, interviews are segmented into clauses. This methodological contribution is beneficial for tackling limited data availability.

4. **Manifesting conditions in non-depressed subjects tend to be much better in linguistic cues (what you say), while depressed patients seem to be better in manifesting their condition in speech (how you say it).** It means that people tend to manifest their condition either through what they say or through how they say it but not through both. This induces different types of errors in each modality; thus, the multimodal approaches benefit from these error differences as one modality compensates for the error of the other modality. This highlights the importance of utilising another source of information with text.

5. **Performing depression detection in less than 10 seconds (this equals less than eight clauses) can be possible without significant performance losses, especially for recall.** The experiment shows that the observed results do not depend on the protocol applied at the beginning of the interviews but on the amount of data. This finding can explain why depression patients tend to manifest their condition so consistently and that there is a high probability of correctly classifying any clause they utter.

The main contributions and novel findings to the field of misinformation are as follows:

1. **Developing an objective credibility assessment system that reduces the burden by assisting humans in verifying the veracity of the textual claims that are expressed freely in Internet.** The experiments show that the approach can reduce by four-fifths the workload of the trained journalists while still keeping the accuracy above 73.4%.

2. **Utilising complementary information beneficially classifies textual claims.** The experiments demonstrate that relying solely on claim inputs without enriching them with relevant articles is insufficient. This is because they underperform, to a statistically significant extent, compared with claims supplemented with relevant articles.

3. **Increasing the length of the supporting articles can capture all the key factors that contribute to identifying the claim identity.** This observation conforms with the actual process of manual fact-checking that entails reading the entire article to make a final decision towards a claim [45].

4. **Using multiple evidence articles for a claim constitutes an important source of information for improving the system's performance.** This finding conforms with the manual fact checking process since the journalists scan the web to investigate the claim identity. The more reliable articles the journalists read, the more confident the results are.

1.4 Organisation of Thesis

This section mainly discusses the remainder of the thesis with core ideas. The thesis is divided into three parts.

- **Part I Depression Detection by Linguistic and Acoustic of Speech**:This part comprises of Chapters 2, 3 and 4. Chapter 2 provides the background of depression from psychology, linguistics and acoustics aspects. Different diagnostic tools are also described, including clinical interviews and self-assessments. Also, the objective markers and indicators for depression, including speech, linguistic and the combination of them, are highlighted. It covers the relevant datasets used in the depression literature. Chapter 3 shows how efficiently linguistic and acoustic/paralinguistic features are combined for identifying depression. We proposed a model that uses network architectures combining textual transcriptions with speech signals through a wide spectrum of multimodal approaches that consider both what people say and how they say it. The effectiveness of the proposed approaches on real-world dataset are investigated. The efficacy of utilising advanced text embedding that considers the context of the word is analysed. In Chapter 4 detailed motivations and extensive experiments are provided.

- **Part II Assessment of the claim's credibility**: This part is divided into two different Chapters 5 and 6. Chapter 5 introduces the problem of misinformation by discussing the effects of misinformation on society and presenting different definitions of this

problem. Additionally, it overviews a language-based approach for distinguishing between fake and real claims. It also shows the importance of extracting external evidence alongside the claim to support the automatic verdict of the system. Furthermore, the importance of the interpretability of verdicts is highlighted to potentially help a reader in understanding the classification decision. Finally, different publicly available datasets are introduced that have been evaluated and used across this part. Chapter 6 presents the proposed approaches that are based on the self-attention mechanism for automated credibility assessment of claims. They obtain the signal from the textual claims and set of supporting articles, which act as evidence that captures higher-level semantics and mimics the human reading process. The effectiveness of the proposed approaches on real-world datasets are investigated. Following that, we empirically evaluated the effect of evidence articles regarding the number of articles required for a claim and their length. Also, the experiments illustrate several application scenarios where the proposed approach uses confidence measures that identify the cases likeliest to be correctly classified. Finally, we investigate the attention weights that highlight how much each word influences an article during the learning process.

- **Part III Conclusions**: This part includes only Chapter 7 with conclusions and future works.

1.5 List of Publications

Most of the materials presented in this thesis have been published in various international conferences and in a journal during the PhD programme. The following list various publications in chronological order:

1. **Aloshban, Nujud**. "ACT: Automatic Fake News Classification Through Self-Attention." In 12th ACM Conference on Web Science, pp. 115-124. 2020. (Full paper) **Part II**

2. **Aloshban, Nujud**, Anna Esposito, and Alessandro Vinciarelli. "Detecting Depression in Less Than 10 Seconds: Impact of Speaking Time on Depression Detection Sensitivity." In Proceedings of the 2020 ACM International Conference on Multimodal Interaction, pp. 79-87. 2020. (Full paper) **Part I**

3. **Aloshban, Nujud**, Anna Esposito, and Alessandro Vinciarelli. "What You Say or How You Say It? Depression Detection Through Joint Modeling of Linguistic and Acoustic Aspects of Speech." Cognitive Computation (2021): 1-14." **Part I**

4. **Aloshban, Nujud**, Anna Esposito, and Alessandro Vinciarelli. "Language or Paralanguage, This is the Problem: Comparing Depressed and Non-Depressed Speakers Through the Analysis of Gated Multimodal Units." In INTERSPEECH (2021). (Full paper) **Part I**

Chapter 2

The State of Depression: Depression Background

This chapter discusses relevant literature from psychology, linguistics and acoustics. Section 2.1 defines depression. Section 2.2 shows language and speech backgrounds, which overviews the language and speech production systems. Also, in Section 2.3, we describe the depression assessment, including different diagnostic tools and the objective markers and indicators for depression. In particular, speech indicators, linguistic indicators and their combination are analysed, and we present how these indicators are applied in the literature. Finally, this chapter covers relevant datasets used in the literature in Section 2.4.

2.1 Definition of Depression

Depression is one of the most common mood disorders worldwide (Figure 2.1) [224]. The most typical form of the pathology is major depressive disorder (MDD), commonly referred to *clinical depression*. It induces negative emotional, physical and psychological consequences [329], and its symptoms include anxiety, sadness, suicidal thoughts and self-hatred, which reduces physical function and sense of wellbeing. The diffusion of depression can be implied from the suicide rate—for example, in Canada, around 4,000 individuals committed suicide, and about 90% of them were diagnosed with some form of a mental illness [219]. It is more frequent among women and can happen at any age group and in any life condition [30, 105]. According to the World Health Organisation (WHO), "*at a global level, over 300 million people are estimated to suffer from depression, equivalent to 4.4% of the world's*

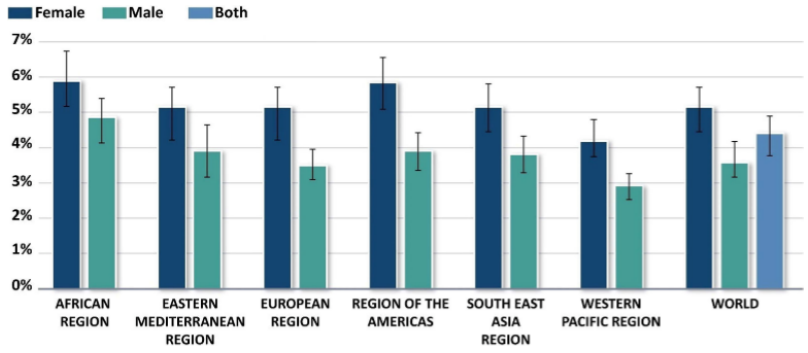

Figure 2.1: Prevalence of depressive disorders (% of population) by WHO region [223].

population [...] the single largest contributor to global disability (7.5% of all years lived with disability in 2015) [...] the major contributor to suicide deaths, which number close to 800,000 per year." [339]. These numbers may still be undervalued because of different factors, such as stigma and lack of available services, which reduce the patient's determination from seeking treatment [335].

The term 'being depressed' has commonly been used in everyday speech to describe altered mood or sadness. However, clinical depression differs from feeling depressed, which commonly results in misidentification by either under-diagnose or over-diagnose depression [186, 208]. Therefore, it is difficult to distinguish between them even for the most up-to-date classification patterns, such as the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), the most widely used resource in diagnosing mental disorders [332]. This may be attributed to the difficulty in psychiatric diagnosis, probably compounded by some factors—the time-consuming process of diagnosis, complicated medical conditions and physical problems that could overlap with the actual psychological illness. Also, emotional symptoms—sadness or hopelessness—are not always expressed by depressed patients [205, 286], possibly due to the lack of objective boundaries between healthy and depressed people, and it is often necessary to assess the past and the current psychosocial history of a possible patient [285]. Thus, this research highlights the importance of developing a computer-aided automatic depression assessment system that supports psychiatrists while diagnosing clinical depression and reduces subjective bias in the diagnosis.

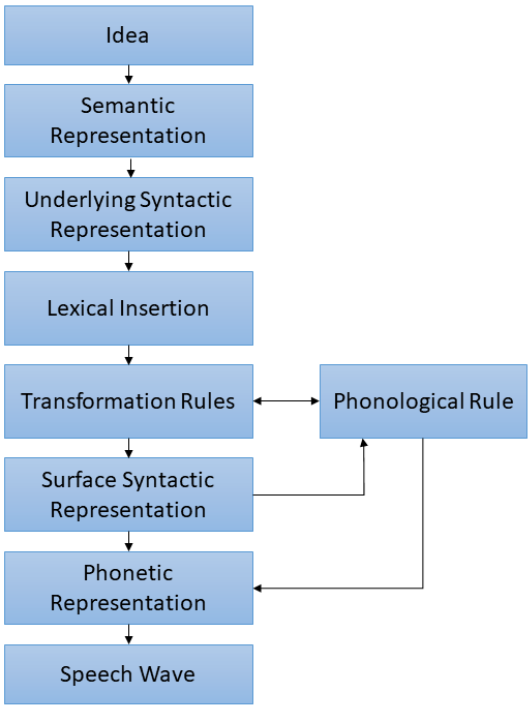

Figure 2.2: Speech production information flow, adapted from [72].

2.2 Language and Speech Backgrounds

Several researchers have proposed psychological theories of depression, such as cognitive, self-aware and social integration theories. Aaron Beck's cognitive theory of depression hypothesises that people prone to depression will usually have a depressive schema. This schema will cause them to have a negative outlook on life and is usually triggered by a very traumatic event. Once this happens, the depressed person will start having depressive thinking and will end up having an episode of depression [28]. Also, Pyszczynski and Greenberg proposed the self-awareness theory for depression, which theorises that people with depression tend to have an excellent opinion about themselves but a tough self-criticism personality [247]. Moreover, a social integration theory of suicide is posited by Durkheim with his own social model. It assumes that depression is a major suicidal key and is essentially initiated when someone believes that they are not accepted by their own society [96].

Given existing psychological theories of depression, both psychologists and linguists have investigated how these theories could manifest in language. Language is a medium that conveys the internal thoughts and feelings of people in a way that others can understand [310]. Thus, cognitive, personality, clinical and social psychologists study humans by language. De-

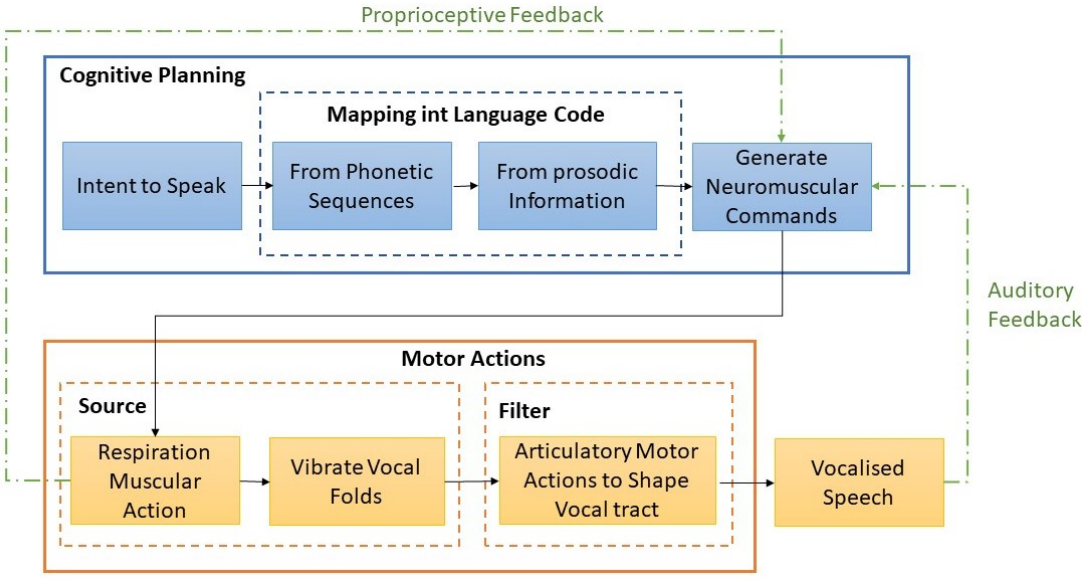

Figure 2.3: Speech Production System, It is adapted from [159, 226]

pression influences how people feel, think and communicate. To see how depression affects people's thoughts, language is scientifically analysed; however, it is complex to understand linguistic behaviour and patterns. Figure 2.2 shows the information flow in which the individual conveys messages into acoustic output during the speech production process.

From Figure 2.2, the information is passed to a different number of steps to generate a meaningful utterance. When a speaker produces an idea, it is interpreted as a linguistic representation, after which the grammatical representation of the utterance is developed. Consequently, the speaker selects one or more main lexical objects. When a fully developed basic structure has been expressed, it is presumed that the structure can subject to changes that delete or change constituents. The surface structure is the output of the transformation stage. Then, phonological rules of stress obligation may be set to the output of the surface structure. Lastly, phonetic representations are transported to the motor system, which creates the articulatory configurations of speech. These steps of semantics and syntaxes influence speech [72], and a multitude of psychological symptoms can affect language selection. Therefore, analysing language when studying depression is important [40].

The study of text analysis is one of the approaches for analysing language that shows the relationship between depression and language use. For example, Stirman and Pennebaker [305] studied suicidal and non-suicidal poets using the Linguistic Inquiry and Word Count (LIWC) dictionary. They found that suicidal poets use more first-person singular words (*I,*

me, my) and fewer words regarding the social collective (*we,us, our*). For medical reports, Poulin et al. [245] observed that certain words increased the probability of committing suicide. Similarly, exaggerated use of *"I"* word and more use of negative emotion words were found in the depressed group compared with non-depressed [270]. The increased utilisation of first-person singular pronouns may suggest a susceptibility of depressed patients to focus mostly on themselves. Their results showed proof coherent with both self-awareness and social integration theories. Additionally, the experiments in [298] observed that the affected group frequently utilised all pronouns and verbs in past tense. Recently, Al-Mosaiwi et al. [9] found that affected group's forums contained more absolutist words than control forums, such as completely, absolutely and nothing. In related work, specific LIWC categories were indicators of some depression symptoms, such as sadness and fatigue [215]. They found a clear discrimination between depressed and control groups regarding their writing styles, content and latent topics [215].

Moreover, linguists strongly believe in the critical relationship between syntax and prosody [62]. Some have even debated whether prosody can be straight projected from the syntactic tree configuration of a sentence [331]. Hence, studying word use and syntax (linguistic domain) alongside prosodic and phonetic features of language usage (acoustic domain) is imperative. The human speech system is complicated, and during speech, over 100 individually innervated muscles are synchronised in the tongue and mouth. Therefore, speech is a sensitive mechanism. Some scientists strongly believe that minor physiological and cognitive variations can cause acoustic alterations in speech [277]. This study presumes that depression can cause cognitive and physiological alterations that affect speech creation, and this alteration can then be detected and assessed. Figure 2.3 describes the speech production system.

According to the schematic diagram, when people communicate, they cognitively scheme a message. After that, they create the phonetic and prosodic details of the message that will be kept in their temporary memory. Consequently, the details will be converted into phonetics and prosodic descriptions, accomplished by sequences of neuromuscular orders. These orders start with the motoric movements needed to produce speech. Motor movements comprise the source and filter. The source is represented by air created by the lungs, which goes through the filter that forms the sound. The filter (Figure 2.3) characterises the vocal tract.

The articulators of the vocal tract change the sound created, depending on their position. Research has examined the cognitive properties of speech creation, and they concluded that cognitive weaknesses are strongly related to depression and affect the working memory. The phonological loop (Figure 2.3) is a significant factor in the speech creation system and is a segment of working memory; the loop regulates the articulatory system. Thus, a cognitive weakness in working memory may disturb this segment of the speech creation mechanism. Many psychological studies [64] concluded that depression strongly impacts the phonological loop, producing articulation and phonation faults. Therefore, this will affect the speech creation mechanism, which makes speech an attractive candidate for an objective marker of depression.

As described above, the relationship between speech, specifically non-verbal paralinguistic cues, and psychological depression has been established. It distresses speech production and cognitive processes, and psychological depression may highly affect speech motor control [79, 278]. This disease can be detected by prosodic irregularities and articulatory and acoustic faults [147]. The speech quality has been influenced [3, 137, 272, 278, 297] regarding the glottal pulse form, breathiness degree, jitter and shimmer. Recent studies have shown a link between depression and changes in the neurophysiological system. This change can target and alter the laryngeal control and its dynamics—the performance of the vocal folds [51,79,248,297,300]. After these studies, many voice characteristics, such as jitter and glottal flow, have been suggested as speech-based biomarkers for detecting depression-related problems [228, 248].

In this study, while we refer to any techniques inducing the understanding of natural language (i.e. syntactic and semantic analysis) to linguistic analysis, those techniques inducing the understanding of speech are referred to as acoustic analysis.

2.3 Depression Assessment

This section presents the most common diagnostic tools for depression. It also discusses several objective signs of depression, emphasising linguistic, acoustic and multimodality signs, and how they are applied in previous works.

2.3.1 Diagnostic Tools for Depression

Commonly used assessment tools for depression include clinical interviews and self-assessments questionnaires. The standard approach to diagnosing depression is using a clinical interview assessment tool that assesses the presence of DSM-5 criteria. Detecting the presence and intensity of depression symptoms is usually assisted by scoring scales filled out by a well-certified psychiatric specialist. For example, the Hamilton rating scale for depression (HRSD) [128]) is one of the most common scales in clinical practice. The HRSD is clinician-administered, includes 21 questions, and takes 20 to 30 minutes to complete. This instrument evaluates the intensity of 17 symptoms associated with depression (such as depression mood swings, suicidal thoughts, sleep disorder, anxiety and irritability) and gives a patient a score, which relates to their depression level. Each question has 3 to 5 possible responses ranging in severity (i.e. scored between 0–2, 2–3 or 4–5), depending on the importance of the symptom. All scores are then summed, and the total is arranged into five categories (from normal to severe). However, HRSD and DSM-5 clinical criteria are unreliable [24, 61], due to their inconsistency in diagnosing MDD [158].

Self-assessments tools can also contribute to the clinical diagnosis of depression by providing scores on self-assessment scales and inventories (Self-RIs). The most widely used self-reported measures of Self-RIs are the patient health questionnaire (PHQ) [160], which has several versions of 2, 8 or 9 items, and the Beck depression inventory (BDI) [31]. The BDI comprises 21 items and takes 5 to 10 minutes to complete. The question items aim to cover important cognitive, effective and somatic symptoms observed in depression. Each question has a score on a scale of zero to three based on how severe the symptom was over the previous week. Like HRSD, all scores are summed, and the final score is classified into four different levels, ranging from minimal to severe. Although this type of assessment is practical and affordable, it has some drawbacks. It may not be reliable due to a lack of adjustment to individual differences biases in self-assessments, and attempts to conceal the pathology to escape treatment [238, 258, 306]. While the cost-effectiveness of widespread screening practices for improving the quality of depression care is debated [149], practical issues related to the aforementioned limitations of Self-RIs raise questions regarding the overall utility and effectiveness of this practice for population-based mental health.

These types of questionnaires are different in some items and scoring points that cause

inconsistencies in diagnosing depression between psychiatrists. Also, given that only one test tool is performed, the subjective views and experiences of psychiatrists could produce differences on the diagnosis. Alternatively, two psychiatrists performing the same diagnosis test have a strong possibility of producing a different score for the same subject. Hence, recently, different objective methods exist for diagnosing depression [10,11,50,120,145,355].

2.3.2 Objective Markers for Depression

Theoretically, machine learning algorithms for depression detection should access the same amount of evidence as a clinician requires during the diagnostic process. Consequently, the classifiers should use features that represent each communicative modality: language and speech. This section reviews each modality (i.e. focusing on linguistic, acoustic and multi-modalities) and highlights the successful markers in depression detection systems.

Linguistic Indicators

As previously mentioned in Section 2.2, many theories of depression persist, such as the 1987 Pyszczynski and Greenberg's self-awareness theory, the 1967 Aaron Beck's cognitive theory of depression and the 1951 Durkheim's social integration model. All these theories have inspired empirical studies of depressed language and concurrently supported their validity. Some studies used the LIWC tool to examine patterns in word choice. LIWC shows success in text analysis research; thus, other different studies have used this instrument for depression detection with promising results.

Adding to LIWC, many other approaches have shown accomplishment in modelling word usage. Coppersmith et al. [73] models were also applied on Facebook posts for identifying depression degree in individuals [288]. The study in [17] developed a large-scale quantitative study on the discourses. A set of discourse features was built to measure the correlation of different linguistic aspects of conversations with their performance. The results support Psyzczynski and Greenberg's depression theory, in which writers with a minimum of self-focus were the more successful in counselling conversations.

Adding to word usage, several studies have explored syntactic aspects of language for depressed people. Zinken et al. [366] analysed the narrative style of depression patients and found that specific structures were related to patients' possibility of finishing a guided self-

help treatment. Their observations confirmed the promise in studying syntactic characteristics of an individual's language use. Part-of-speech (POS) tags also have been discovered as the POS's occurrences were useful in detecting depression from writing [210].

Researchers find social media to be an interesting domain to investigate depression. An SVM classifier estimating the onset of postpartum depression (PPD) with an accuracy of 70% has been built [86]. Behavioural indicators, such as linguistic style, social engagement and medication history, were determined. They found that depressed users tend to be slightly engaged in social media especially at morning [86]. Recently, Trotzek et al. [314] addressed the early detection of depression using a single layer of convolutional neural network (CNN) and logistic regression with user-level linguistic metadata, such as text length and POS tags. Another study in [215] investigated linguistic language in online communities by sentiment analysis, ANEW lexicon (affective norms for English words), LIWC and mood tags were applied. They found a clear discrimination between depressed and control groups regarding their writing styles, content and latent topics. Overall, the main issue with social media platforms is the difficulty of testing whether the post claim is true.

Researchers have also investigated the content of the transcriptions of individuals communicating with a human-controlled avatar. In the literature, there are two types of modelling settings for analysing spoken content: context-aware and context-free modelling. The former approach relies on feature engineering (i.e. topic modelling) to extract question-answer pairs. For example, patient's answers to closed questions are selected when they are related to specific symptoms associated with psychoanalytic aspects of depression, such as *'Do you have a history of depression?'*. The label of the sentences reflects the presence or absence of specific words/phrases related to the selected symptoms [340]. Similarly, the approach in [354] selected questions related to the symptoms associated with psychoanalytic aspects of depression, such as sleep disorder. However, this approach requires topic modelling to formulate each answer to its topic, and it is limited to known symptoms.

Additionally, context-free modelling involves responses without prior knowledge of the structure of the interview. This model is data-driven and disregards a priori knowledge of interview structure. For example, Alhanai et al. [6] applied bidirectional long short-term memory network (Bi-LSTM) models to detect depression from human-machine interviews. In related work, Dinkel et. al [92] proposed a text-based multitask Bi-LSTM for modelling

text with attention pooling to visualise the sentences likeliest to account for depression. They experimented with three different text embedding—Word2Vec, ELMo and BERT—and they found that ELMo and BERT have robust and stronger performance than Word2vec in their work [92]. Similarly, semantic and syntactic transcripts were analysed, and they observed that depression severity is related to occupation's and sleep's lexicon [210].

Speech Indicators

Using speech as a diagnostic and monitoring aid is effective for depression [74, 79, 279, 341]. The human speech production framework is exceptionally complex; therefore, slight cognitive or physiological changes can deliver acoustic changes in speech. This idea has driven research on using speech as an objective marker for depression. Depressed speech depends on a wide extent of prosodic, source, formant and spectral indicators.

In patients with depression, several changes in prosody have been attributed to vocal-source and vocal-tract [6, 131, 301]. These prosodic features, including pitch [250], speech intensity [251], loudness [357], energy [302], speaking rate [303], speech pauses [16], voice quality [280] and formant measures, are effective for classifying depression due to increasing tension in the vocal tract associated with depression [110, 112]. Also, reduced loudness variability, repetitious pitch inflections and stress patterns, alongside monotonous pitch and loudness, are good indicators. Alghowinem et al. [13] inspected some features for detecting depression from spontaneous speech and found that the most discriminative features are loudness, root mean square and intensity features. Psychomotor retardation in depressed people has been studied, and they found that individuals suffering from depression are likelier to have a reduction in the second formant range (F2) and slower rate of speech compared with individuals without mental illness [1]. Similarly, the experiments in [78] showed that the first three formants produced high classification performance. Specifically, many researchers have observed correlations between a reduced F0 range and a reduced F0 average with increasing levels of depression. Nevertheless, this contrasts with several studies that showed no significant correlations between F0 variables and depression levels. The disagreement in the results might be based on heterogeneousness for representing depression in individuals [69].

Spectral-and energy-based features are suitable for classification because the depressed signal can carry more information in the higher energy band compared to the neutral sig-

nal [112, 147]. The study in [188] inspected several acoustic features—spectral, cepstral, prosodic, glottal and a Teager energy operator. The best performances were 87% and 79% accuracies for males and females, respectively. In [78], spectral features, particularly MFCCs, were found to be useful with an accuracy of 80% in a speaker-dependent configuration. Recently, the study in [12, 291] investigated spontaneous speech and found that MFCC, energy and intensity features are the most discriminative. Williamson et al. [342] studied two vocal tract representations, which are formant-frequency tracks to encode the vocal tract resonant frequencies and MFCC features to encode spectral shape dynamics. These feature sets cause changes in the coordination of vocal tract motion associated with MDD. Then, Gaussian mixture model (GMM)-based multivariate regression scheme was applied for final prediction.

Many available toolkits are widely applied to extract combinations of low-level indicators—OpenSmile [101], COVAREP [88], SPTK [141], KALDI [246], YAAFE [197], and OpenEAR [102]. Each existing toolbox is generally due to a single laboratory's work. Each researcher considers different features from his or her own viewpoint and suits their dataset. Yet, these hand-crafted statistical properties are extracted based on prior knowledge about speech perception and speech production and may not contribute to the improvement. Likely, no consent set of features exists that may be considered the most useful for depression analysis.

Manually extracting features needs to understand the domain knowledge; thus, deep learning could better capture useful information from signal. Combining hand-crafted features with deep-learned features shows to be effective. For example, the authors [192] applied an audio-based approach for depression classification using CNN, followed by a long short-term memory network (LSTM) on a log Mel filter-bank and magnitude-spectrogram features. Similarly, LSTM on 279 features extracted from the COVAREP tool were investigated [5]. He et al. [131] also adapted CNN to learn deep-learned features from spectrograms, state-of-the-art LLD and raw signals. However, some researchers have tried to extract features directly from sound waves for depression detection [230, 273].They found that using MFCC features yields better performance than using raw sound waves directly. Similarly, the experiment in [256] assessed the severity level of depression from speech using MFCC features as input for the LSTM network, and they achieved an accuracy of 76.27% on the DAIC-WOZ dataset.

Multimodal Indicators

Recent experimental works have explored the automatic analysis of depression from multimodal approaches. In [200], they used a late fusion approach that trained visual and acoustic models separately, and their decisions were combined using the weighted sum rule. They found that combining these modalities at the decision level gained a better improvement for depression detection. The relationship between facial actions and vocal prosody for depression detection has been investigated. Facial movement, head movement and vocal prosody were studied, and their combination achieved good results [69].

Adding to combining voice and visual-based markers, researchers have also provided empirical support for the existence of a relationship between depression and language. By extracting numerous hand-crafted features from text and speech, the experiment in [114] observed that a multimodal system combining these two sets of features induces the best performing system (accuracy of 65.8%). Morales and Levitan [210] used automatic speech recognition (ASR) to automatically transcribe speech and found that audios with their transcriptions enhance the performance. The experiments in [6] modelled audio and text sequences with an LSTM network, and the best fusion performance reached 77% F1 measure. Lam et al. [166] analysed depression levels using a transformer for text feature modelling and CNN for audio feature modelling. Moreover, in [229], they investigated the depression level by extracting hand-crafted features, such as sentiment analysis on the participant's responses to the interviewer's questions, the speaking rate and the average length of the utterances during this answer.

The combination of text, audio and video has also been explored. Gong et al. [122] proposed an approach to predict depression levels by combining topic modelling of question/answer of the interviews with hand-crafted features from text, audio and video features. The study in [354] also extracted hand-crafted features from each modality, which were then fed as an input into a CNN. The learned features were fed to a feed-forward network to predict the severity of depression. Similarly, the approach in [249] combined acoustic, visual and text modalities, and different combinations of these modalities were fed to an attention-based neural network to predict depression level. They found that text modality, regarding accurate prediction, outperforms other single modalities. Similarly, [255] proposed multiple layers of the attention model and found that the text modality had the highest weight, and almost

equal weights were assigned to audio and video modalities. Also, different fusion techniques, including early, late and hybrid fusions, have been deeply studied [209]. This research also proposed a syntax-informed fusion approach to leverage syntactic information to obtain more informative aspects of the speech signal. However, the overall results do not seem to obtain any statistical evidence of this finding. Rohanian et al. [265] also proposed word-level multimodal fusion with feed-forward networks as a gating mechanism from visual, speech and text. They found that combining text and audio modals achieved the best result of 81.0% F1. Furthermore, fusing audio, text and video features to a decision tree algorithm has obtained satisfying results in predicting the PHQ-8 score over the benchmark dataset, DAIC-WOZ. The researchers in [91] proposed a decision-level fusion approach for predicting the depression scale with features extracted from the provided DAIC-WOZ dataset. Gaussian staircase model was applied in [340] to produce the final regression result by combining facial actions, vocal prosody and text features.

2.4 Existing Datasets

The availability of empirical data is critical for developing and evaluating methods for automatic depression detection. Various datasets are reported in relevant works, which are summarised in Table 2.1. The table includes total number of participants, ground truths, research questions and the availability to the third parties. The most common practices to encourage collaboration are to conduct challenges and to release public data and code. The advantage of it is to promote research, spur interest and build connections across the research community. The examples of such challenges are the computational linguistics and clinical psychology (CLPsych) shared task (2013–2017) and the audio-visual emotion challenge (AVEC, 2013-2016). This section briefly describes the most common existing datasets applied in previous studies, including Pittsburgh, BlackDog, ORYGEN, AVEC and DAIC-WOZ.

Overall, these datasets differ in assessing depression, which can be clinical assessment or self-assessment tests. DAIC-WOZ and AVEC datasets are used depression self-assessment tests, such as PHQ. While Pittsburgh, BlackDog and ORYGEN datasets are used clinical-assessment depression tests, such as DSM-IV and HRSD. Moreover, the objectives of each dataset differ, ranging from comparing depressed to control subjects to evaluating depression

Table 2.1: Datasets employed by the reviewed studies for depression research

.

Corpus	**Total (Affected group/Control group)**	**Ground Truth**	**Research Question**	**Availability to third parties**
Pittsburgh	49 subjects	Clinical Assessment	Severe/Low depression Detection	Visual and Audio Recordings
BlackDog	80 subjects (40/40)	Clinical Assessment	Severely Depressed/Healthy Control Detection	Visual and Audio Recordings
ORYGEN	30 subjects (30/30)	Clinical Assessment	Depression prediction	Visual Recordings
DAIC-WOZ	189 subjects	Self-assessment	Depressed/Healthy Control Detection and prediction	Visual, Audio Recordings and Transcripts
AVEC	292 subjects	Self-assessment	Severe/Low depression Detection and prediction	Visual and Audio Recordings

severity. Concerning data availability, because of the confidential nature of the data and privacy issues, the depression dataset is difficult to gain. Among the existing datasets, only the AVEC and DAIC-WOZ datasets could be shared under a privacy agreement. However, they are not used clinical depression assessment tools, which might affect the scale of depression and its automatic evaluation. The lack of this standardised dataset introduces challenges, such as results replication, and increases the difficulties of developing a generalised system that recognises depression symptoms.

2.4.1 Pittsburgh Dataset

The Pittsburgh dataset was collected at the University of Pittsburgh during treatment sessions of depressed patients. It is a clinically validated depression dataset used to determine the relationship between vocal prosody and changes in depression severity over time. The participants were recruited from a clinical trial, where the participants were diagnosed with depression according to DSM-IV. A total of 49 depressed patients were evaluated at seven-week intervals using a semi-structured HRSD clinical interview for assessing severity of depression [357]. The dataset included visual and audio recordings.

2.4.2 BlackDog Dataset

BlackDog corpus was collected at the Black Dog Institute, a clinical research institute focusing on mood disorders, including depression and bipolar disorder. Audio-video recordings from 40 depressed subjects with DSM-IV scores exceeding 15 and 40 age-matched controls were obtained. The interview was conducted by asking particular questions in which the subjects were asked to describe events that had provoked significant emotions. The dataset had many components, such as reading sentences and interviews [291].

2.4.3 ORYGEN Dataset

ORYGEN dataset was acquired due to research cooperation with the ORYGEN Youth Health Research Centre. It contained video recordings of discussions conducted between parents and their adolescent children. It was built to predict whether initially non-depressed adolescents would develop depression at the end of a two-year follow-up period [220].

2.4.4 AVEC Dataset

The AVEC challenges were organised competitions aimed at comparing multimedia processing and machine learning methods for automatic audio, video and audio-visual emotion and depression analysis, with all participants competing strictly under the same conditions. It included audio and video recordings of interviews in German, conducted by an animated virtual interviewer. Overall, the dataset included 292 videos with duration ranging from 6 seconds to 4 minutes. In addition, it comprised various vocal exercises, including free and read speech tasks with their answers. Depression was estimated for each recording using BDI-II. Different versions of the AVEC dataset are available, including AVEC2013 [322], AVEC2014 [321] and AVEC2016 [320]. Both the AVEC 2013 and 2014 corpora are available to download [1].

2.4.5 Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WOZ)

Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WOZ) was built by conducting interviews in the English language to diagnose several psychological distress conditions, such as anxiety, depression and post-traumatic stress disorder. The interviewer was an animated

[1] https://avec2013-db.sspnet.eu/

virtual interviewer called Ellie, and she asked the interviewees a series of open-ended questions to identify clinical symptoms. Each subject was assigned a single depression value using PHQ-8 [89, 124]. DAIC-WOZ dataset served as the benchmark dataset in AVEC'17 and AVEC'16 depression sub-challenges. For audio recordings and transcripts, raw data were available, while only features extracted with OpenFace were available for video recordings. Despite this limitation, several interesting approaches were presented.

2.5 Conclusion

This chapter reviewed the clinical definition of depression, symptoms and diagnostic assessment approaches, including current diagnostic questionnaires and the objective markers for depression, especially linguistic and speech aspects. A brief background on speech and language was also given in this chapter, highlighting language and speech production systems. It provided a better insight into how the process of language and speech involve separate but coordinated actions, any of which can be interpreted by psychogenic illnesses, such as depression. The chapter also showed that psychologists and linguists have proven that depression influences how a person communicates; thus, the theories and studies motivate the building of a multimodal system. Further, existing datasets were identified. In the next chapter, we present our methodology for building a depression detection system.

Chapter 3

What You Say or How You Say It? Depression Detection Through Joint Modelling of Linguistic and Acoustic Aspects of Speech

3.1 Motivation

Depression certainly impacts the way people feel, think, and communicate [21]. The influence of depression on linguistic style is primarily explained by cognitive models (e.g., research in [29, 67])), which indicate that depressed patients have higher levels of negative emotions and self-focus. Social integration/disengagement hypotheses (e.g., research in [95]) also study patterns in which depressed patients become less emotionally involved with community. These underlying mechanisms manifest themselves through language, indicating an increased self-focus, splitting from others and negative motion [65, 310] (see Chapter 2). Therefore, studying language to detect and evaluate human mental health conditions, such as depression has been observed as an appropriate mental health model. In this research, while we study the linguistic aspects and examine its utility to capture depression, we explore whether a more advanced word embedding methodology (e.g. BERT) that considers meaning and represents the same word differently depending on its context can contribute to the result. Particularly, we aim to develop a high-quality contextualised embedding for the interview transcriptions from BERT-based model to capture the linguistic properties of speech, where it achieves the state-of-the-art performance in many NLP tasks.

Because of inherent differences in the speech system, no particular cue has sufficiently distinctive influence as a symbol of depression on its own [79]. This means that linguistic cues alone may not be sufficient to understand the mental traits and states of a person; thus, input from other modality must be supplemented. Interviews concurrently reveal linguistic contents (what people say) and paralinguistic/acoustic speech (how words are said), which can reveal about a person's emotional, physiological and mental characteristics. Therefore, modern speech techniques are proposed to assess, diagnose and monitor the various mental disorders that affect the voice of the person in question [77]. Particularly, depression interferes with the neural processes underlying language and communication (see, e.g. [81, 290] and Chapter 2, Section 2.2), thus leaving detectable traces in both what people say and how they say it. Also, the use of speech has several advantages from an application viewpoint, including the possibility to detect depression via phone [140], typically the means through which people contact counselling services or using ordinary laptop microphones in an informal setting like it happened for the data used in this work. Hence, this chapter proposes a multimodal approach designed to detect depression based on linguistic and acoustic aspects of speech. Particularly, it uses network architectures combining speech signals and their transcriptions through various multimodal approaches that consider both *what people say* and *how they say it*.

Recently, depression detection has attracted significant attention in the computing community. However, it is frequent to observe that the people involved in the experiments have not been diagnosed with depression by professional psychiatrists, like in this work, but have simply performed a self-assessment with questionnaires, such as BDI and PHQ (see Chapter 2 , Section 2.3.1). In this respect, the actual task being addressed is not depression detection, but inference of the scores resulting from the questionnaires. This is the case, for example, of the benchmarking campaigns conducted in the framework of the AVEC [261, 320] and of other works that have used the data at the core of such campaigns [8, 80, 353, 363, 365]. Therefore, the experiments of this work were performed over a corpus of 59 participants, including 29 persons diagnosed with depression by a psychiatrist and 30 that never experienced mental health issues (see Section 3.2 for the dataset). During the experiments, the approaches were applied to *clauses*, that is, to manually extracted linguistic units that include a noun, a verb and a complement. Given that the average number of clauses per participant is 114,

this allows one to perform, for every person, numerous clause level decisions, and these can be aggregated through a majority vote. This is important because it shows that effectively tackling the limited amount of available data is possible—a problem inherent to depression detection due to ethical and practical concerns in recruiting depression patients.

To the best of our knowledge, this is one of the first depression detection works involving Italian speakers. This is important because it shows that depression detection technologies can be effective not only for English speakers, the most common case in the literature, but also for people that belong to different cultures. Furthermore, unlike other works in the literature (see Section 3.3), the distinction between depressed and non-depressed participants has been made by psychiatrists and not through the administration of self-assessment questionnaires. This is an advantage because it increases the chances of the data to be representative of the actual difference between depressed and non-depressed speakers. Alternatively, it ensures that the problem addressed in the work is depression detection and not inference of self-assessment scores. This is important because self-assessment questionnaires are subject to multiple biases [232] and, furthermore, the data show that they can be filled inconsistently, especially by people affected by depression (see Section 3.2). Additionally, this work presents one of the first comparative unimodal and multimodal effects on depressed and non-depressed individuals.

The research questions and subsequent novel contributions of this work are the following:

1. Does replacing static word representation with contextualised word representation (BERT) induce a significant improvement?

2. Is it 'what it is said (linguistic)', 'how it is said (acoustic)' or the combination of them?

This chapter first describes the data used in the experiments. Then, a survey of previous work related to the problem is highlighted. Further, data preprocessing is described. After that, the proposed approach is then presented. Finally, we report our experiments and discuss the result.

3.2 The Data

As described in the Chapter 2, Section 2.4, the existing datasets have some limitations. Therefore, a new dataset was collected and used for the experiments in this research. The dataset

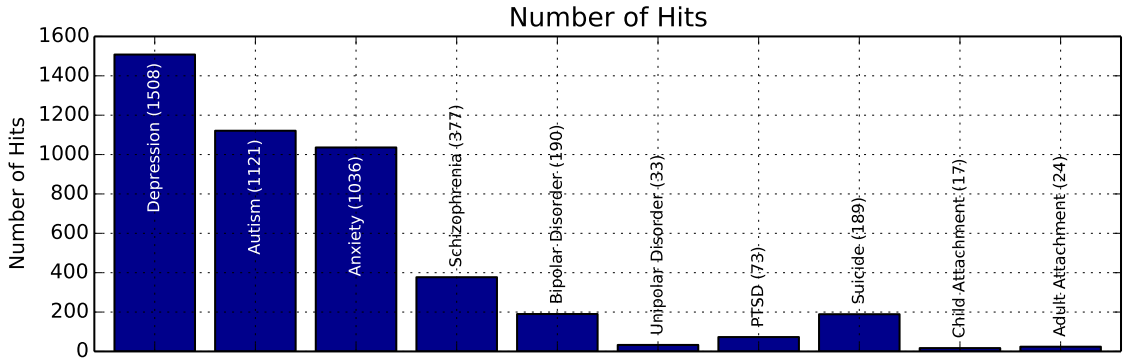

Figure 3.1: The chart shows the number of hits returned when submitting queries related to mental health issues to IEEEXplore (`https://ieeexplore.ieee.org/Xplore/home.jsp`). The queries have been submitted with the constraint of returning material published after 2009.

contained 59 interviews recorded in three Mental Health Centres in Southern Italy. Every interview involved a different participant, but the protocol remained the same. In particular, the interviewers posed always the same questions (e.g., "*What have you done during the last week-end?*") and always in the same order. Out of the 59 participants, 29 were diagnosed with depression by a professional psychiatrist, while the remaining 30, referred to as *control* participants, have never experienced mental health issues. The interviewers were instructed to speak as little as possible and, on average, they were speaking 10.0% of the interview duration. When considering separately depressed and control participants, the fractions were 5.1% and 14.7%, respectively. The difference was statistically significant ($p < 10^{-5}$ according to a two-tailed t-test) and one possible explanation is that control participants tend to involve the interviewers in interaction, while depressed ones simply tend to answer the questions.

Table 3.1 provides demographic information. The gender distribution is the same for both depressed and control groups with 2.47 times more females than males. This follows the observation that, despite cultural and national differences [19], women tend to develop depression roughly two times more frequently than men [148]. Concerning age, the range is roughly the same and, according to a two-tailed t-test, no statistically significant difference exist between the average ages (45.7 for depressed and 44.0 for control). This range is chosen because depression tends to be less frequent for children [117], adolescents [149] and people older than 65 [199]. Thus, the experiment participants should be representative of the popula-

Table 3.1: The table shows the demographic information available about the participants. According to a t-test, no difference exists between depressed and control participants regarding age. Similarly, according to a χ^2 test, the distribution of gender and education level is the same for both groups.

	F	M	Avg. Age	Age Range	Primary	Superior
Depressed	21	8	45.7	23-69	16	13
Control	21	9	44.0	23-68	12	18
Total	42	18	44.4	23-69	28	31

tion prone to depression. Finally, the table reports the distribution across the education levels of the Italian system, namely *Primary* (up to 8 years of education) and *Superior* (between 13 and 18 years of education). According to a two-tailed χ^2 test, the difference between the two distributions is not statistically significant, which means that, overall, the two groups differ regarding mental health condition (depressed or control), but not regarding the other factors (gender, age and education). This should ensure that the approach proposed in this work detects depression and not other factors that probably induce linguistic or acoustic differences in speech.

The upper chart of Figure 3.2 shows how durations distribute across the participants. On average, every interview lasts 242.2 seconds, but statistically significant difference ($p < 0.05$ according to a one-tailed t-test) was observed when considering separately depressed and control participants (the averages were 216.5 and 267.1 seconds, respectively). Every interview was manually transcribed and segmented into clauses, i.e. basic linguistic units that include a noun, a verb and a complement. The clauses were the analysis unit of the experiments, meaning that they were analysed and recognised individually before a participant was classified as depressed or control (see Section 3.4 for more detail). Hence, the lower chart of Figure 3.2 shows the distribution of clauses and the average number of words they include. Overall, the average number of clauses was 114.0 but 100.8 and 126.8 when considering depressed and control participants, respectively (the difference was statistically significant with $p < 0.05$ according to a one-tailed t-test). However, statistically significant difference was unobserved regarding average number of words per participant (429.7 and 463.9 for depressed and control participants, respectively), thus suggesting that depressed participants tend to use

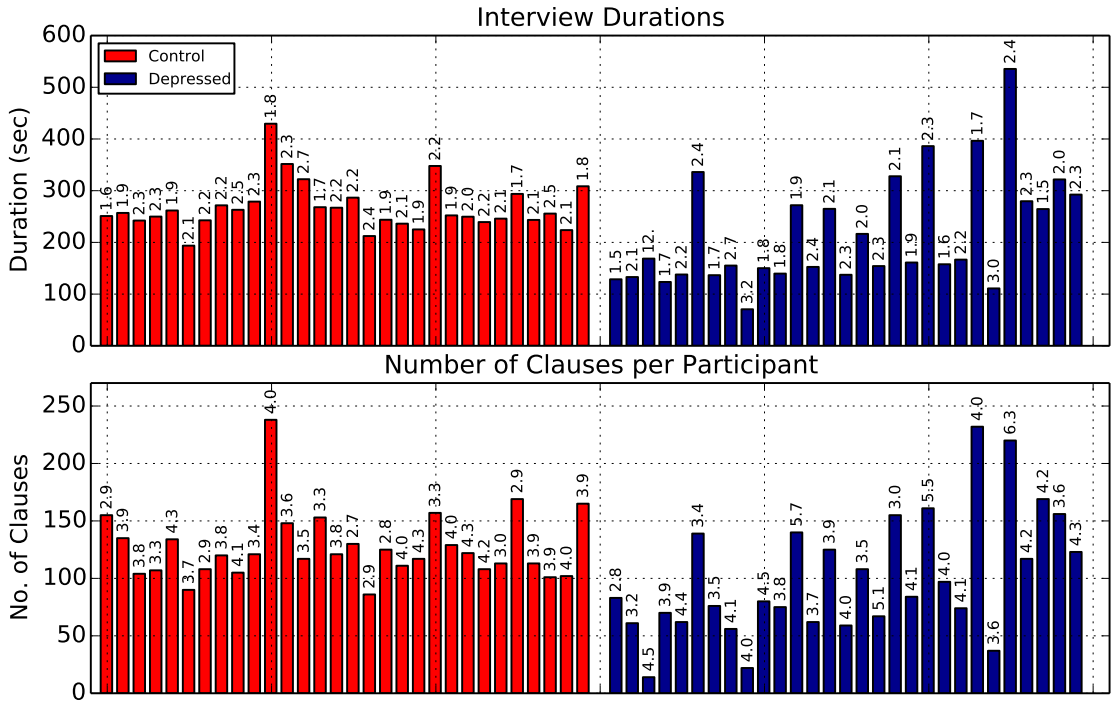

Figure 3.2: The upper chart shows the interview durations for all participants and the number at the top of each bar is the average duration (in seconds) of each clause. The lower chart shows the number of clauses for each participant, and the number at the top of each bar is the average number of tokens per clause (the tokens are sequences of characters enclosed between two consecutive blank spaces and typically correspond to words). In both charts, depressed and control participants are shown separately.

Table 3.2: The table shows the distribution of the score across the four conventional ranges used to interpret the Beck Depression Inventory II scores, namely *minimal* (0-13), *mild* (14-19),*moderate* (20-28), severe (29-63).

Condition	Minimal	Mild	Moderate	Severe
Depressed	7	2	11	6
Control	24	2	1	2
Total	31	4	12	8

more words per clause. Difference in duration and number of clauses agreed with previous observations, showing that people affected by depression tend to display lower involvement in conversations [44, 118].

Out of the 59 participants, 55 accepted to fill the BDI-II [32], one of the self- assessment questionnaires most commonly used to support depression diagnosis. The result of the questionnaire was a score that, on average, was proportional to the severity of the depression condition, which is the state of depression with regard to its symptoms. Table 3.2 shows the distribution of the scores across the four conventional ranges used to interpret the BDI-II scores, namely *minimal* (0-13), *mild* (14-19), *moderate* (20-28) and *severe* (29-63). The data shows that, on average, the scores account for the actual condition of the participants (the average scores are 21.7 and 9.7 for control and depressed participants, respectively). However, roughly one third of the participants diagnosed with depression had scores that fell in the minimal and mild ranges, those considered non-pathological. This suggests that the BDI-II scores, at least in the data used for this work, cannot be considered fully reliable, especially for depression patients. One possible explanation is that self-assessment questionnaires are sensitive to multiple biases and '[...] accuracy is not the only motive shaping self-perceptions [...] the other powerful motives are consistency seeking, self-enhancement, and self-presentation" [232]. Alternatively, the data of Table 3.2 suggest that several depressed participants could not fill the questionnaire or, possibly, they have tried to conceal their condition, probably to avoid the stigma associated to mental health issues.

Regarding the ethical implications of the work, we make sure that the data is stored in a password protected repository and the name of the participants is never known (only an ID). All participants also have signed an agreement where they accepted to have their data

shared. Furthermore, every person that will get access to the data will have to sign an EULA (End User Licence Agreement). The ethical clearance at the origin of the project is the ethical committee of the Department of Psychology at Universita' degli Studi della Campania "Luigi Vanvitelli", responsible for the data collection, provided the ethical clearance with protocol number 09/2016.

3.3 Survey of Previous Work

Chapter 2, Section 2.1 shows that depression mainly impacts both the life of patients and the entire society. Correspondingly, Figure 3.1 shows that, when submitting the query "*depression psychiatry*", *IEEEXplore* returns more hits than for any other mental health issue. Depression has been the subject of at least four benchmarking campaigns organised in the last decade, including two based on a corpus that show 292 people performing a human-computer interaction task [318, 319] and two based on a corpus where over 200 individuals interact with an artificial agent [260, 317]. In all cases, the task addressed by the participants is the inference from behaviour of scores resulting from the administration of self-assessment questionnaires, such as BDI-II [32] or different versions of the PHQ [119].

The corpora collected for the challenges above were used to investigate different approaches, including the use of facial behaviour [7, 361, 364], the analysis of paralinguistics [75] and the multimodal combination of multiple cues [352]. The experiments presented in [7] were based on how the activation of individual facial muscles changes over time, while in [361] the goal was the identification of *markers* of facial behaviours that explain the presence of depression. For the approach proposed in [364], there was no attempt to analyse the way depression leaves traces in facial expressions and the focus was on the sole inference of the BDI-II scores. The methodology investigated in [75] included two main steps—the inference of the BDI-II range in which the score of a particular person is and the inference of the exact score of a person in such a range.

The results obtained in the works above can be compared (they were obtained over the same data) and range between 8.2 and 9.8 regarding root mean square error (RMSE). Given that the BDI-II scores can be between 0 and 63 (see Section 3.2), such RMSE values do not necessarily allow one to distinguish between people in the range 0–13 (corresponding to

minimal or null depression) and people between 14 and 63 (corresponding to mild to severe depression). Also, differences between facial expression utilization and paralinguistic utilization seem unobserved. A different approach was made in [352], where a multimodal approach (based on facial expressions, paralinguistics and manual speech transcriptions) achieved an F1 score of 75% in discriminating between people below and above the PHQ-8 score threshold corresponding to depression.

While the works mentioned so far have focused on the inference of self-assessment scores, others have addressed the problem of detecting people diagnosed with depression by professional psychiatrists (like it happens in this study) [10, 11, 50, 120, 145, 355]. Such a task is performed with accuracy up to 90% in [50] using electroencephalograms (EEG), with 88% accuracy through the multimodal combination of paralinguistics, head pose and gaze in [10] (following up on a previous approach presented in [11]), and with F1 measure up to 80% by analysing body movement combined with head pose and facial expressions [145]. While being incomparable (they have not been obtained over the same data), such results suggest that replicating the judgment of professional psychiatrists around 4 times out of 5 is possible.

Regarding the modalities used in this study (linguistic and acoustic aspects of speech), several works propose experiments aimed at investigating specific aspects of depression. In [140], the focus was on using short utterances collected through mobile phones (a setting typical of counselling services accessible through the phone). The results show that detecting people above the PHQ-9 threshold corresponding to depression with accuracy up to 72% is possible. The experiments proposed in [187] addressed the problem of adolescent voices that, not being fully formed, are more challenging to process automatically. The results show that energy, accounting for how loud people speak, is the best depression marker, especially when measured with the Teager [282]. Similarly, the result presented in [74] showed that the main difference between depressed and non-depressed speakers is phonetic variability, with depressed people tending to be less variable.

Adding to the above, several works have addressed the problem of combining speech and its transcription (like in this work). While some works have suggested, based on experimental evidence, that linguistic and paralinguistic aspects of speech should always be modelled jointly [211], others have shown that this is not necessarily the case and better results can be achieved, for example, using the sole speech transcriptions [343]. Furthermore, other works

have suggested that the multimodal combination of speech and its transcription improve over the individual modalities only when considering that a sentence has been uttered during an interaction [5] or using models that include attention gates capable to identify, for every sample, the modality more likely to produce the best results [264]. Alternatively, it is unclear whether depression relevant information is carried more effectively by linguistic or acoustic aspects of speech.

Overall, the brief state-of-the-art presented in this section suggests that no form of behavioural evidence (speech, facial expressions, gestures, etc.) clearly outperforms the others. Furthermore, the use of similar approaches (e.g., the joint modelling of linguistic and acoustic aspects of speech [5, 211, 264]) over different data does not necessarily induce the same conclusions about how effective using a certain modality is regarding the others. One possible reason of such state-of-affairs is that several works ignore the problem of identifying people diagnosed with depression by a doctor, but the problem of inferring self- assessment scores. These are affected by different biases (see Section 3.2) and, therefore, can induce ambiguous results. Furthermore, depression is a complex phenomenon involving varying factors (e.g., physiology, socio-economic status, age, gender, etc. [142]) that result into individual differences in the way people manifest the pathology.

3.4 Data Preprocessing

Data needs to be preprocessed before being analysed, either by removing noise, reducing the high dimensionality and/or segmenting interesting parts for feature extraction. It is the first step towards building a robust model. In this research, the raw data underwent several preprocessing steps—segmentation and data cleaning.

Segmentation It is the process of identifying the discrete units occurring in a sequence of sounds. Before the segmentation, interviewer is separated from pure subjects' speech. The data is then segmented into meaningful parts for further analysis. In this study, every interview was manually transcribed, and the audio and its transcriptions were segmented into *clauses*. Clauses is the basic linguistic units that include, following the definition of the Cambridge Dictionary, a noun, a verb and a complement. The clauses are the analysis unit of the experiments, meaning that they are analysed and recognised individually before

a participant is classified as depressed or control. The rationale behind this segmentation is that the clause is a smallest unit of speech that express part of a speech act, such as narrating, explaining or interrupting. Also, the experiments involved 59 participants, considering that we have an average of 114 clauses per participant (see Section 3.2); thus, by applying this segmentation, loads of clause level decision can be achieved.

Data Cleaning Each speech signal and its transcription contain the participant's answers and the interviewer's questions and responses. The interviewer's questions and responses were discarded. The participant's answers can be verbal cues, which is the spoken words, and nonverbal cues (e.g. laughing and coughing). In this study, the verbal clauses of the participants were only extracted and considered. Given the transcriptions, no stemming or stopwords removal were done on the extracted clauses, and the resulting tokens of each clause were lowercased. Finally, the transcriptions were aligned with the audio clips.

3.4.1 Preprocessing for BERT Model

For BERT model, the data is required to be processed further in a specific format to feed it to the model.

Tokonizer For each word in a clause, WordPiece tokenizer was applied. It breaks the words down to their prefix, root and suffix to handle unseen words better. This eliminates the need for stemming or other out-of-vocabulary (OOV) word handling.

Special Tokens Addition Special tokens were inserted to the start of each clause ([CLS]) for classification embedding and to the end of each clause ([SEP]) for denoting the end of the clause.

3.5 The Approach

The proposed approach comprises four main steps—*unimodal recognition*, *multimodal recognition*, *clause classification* and *aggregation*. Figure 3.3 shows the unimodal recognition approach used in the experiments. The feature extraction maps the clauses into sequences of feature vectors that are then fed to a Bi-LSTM. This latter outputs a representation that is

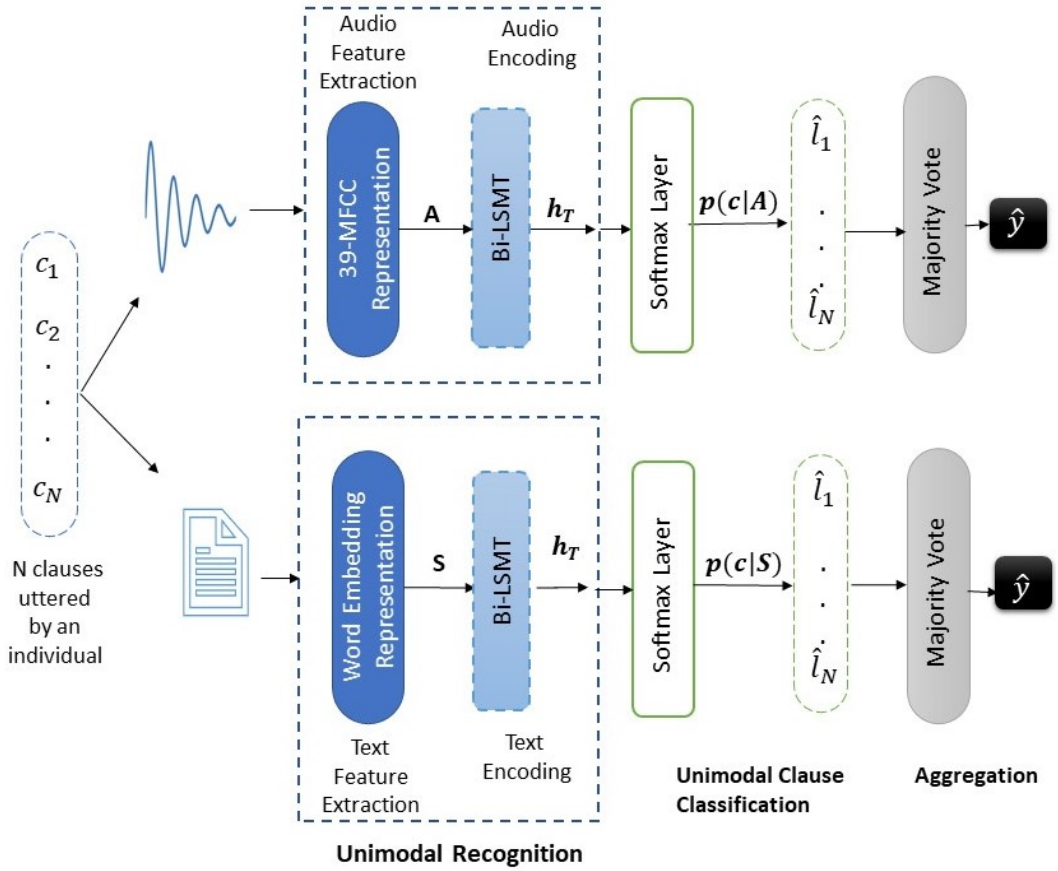

Figure 3.3: The figure shows the unimodal recognition approach. Speech signal and textual transcription corresponding to every clause are converted into sequences of feature vectors (A and S, respectively) that are fed to a Bi-LSTM followed by a softmax layer. The output of this latter can be thought of as the a-posteriori probability distribution of the classes (a clause is assigned to the class with the highest a-posteriori probability). The classification outcomes of the individual clauses are aggregated through a majority vote (a participant is assigned to the class her or his clauses are most frequently assigned to).

given as input to a softmax layer that estimates the posterior probabilities of the two possible classes (*control* and *depression*).

Figure 3.4 shows the different strategies of multimodal recognition approaches. In particular, the unimodal representations output by the Bi-LSTMs (h_T corresponding to text and audio models) are combined through different intermediate fusion strategies including feed-forward (FF–TF), logistic regression (LR–TF) and gated multimodal unit (GMU) [20]. In addition, the output of the unimodal classifiers serve as input to *Sum Rule* [153] multimodal approach. Using different combination approaches ensures that the conclusions of this work result from actual properties of the data and not from using a particular methodology.

In both unimodal and multimodal cases, the input corresponds to the N clauses $\{c_1, c_2, ..., c_N\}$

that a given participant utters (the value of N changes from one participant to the other). Each clause c_i is then assigned to the class with the highest a-posteriori probability, resulting into N individual outcomes $\{\hat{l}_1, \hat{l}_2, ..., \hat{l}_N\}$, where $\hat{l}_j$ is one of the two possible classes, i.e., *depression* or *control*. The classification outcomes corresponding to the N clauses uttered by a particular individual are obtained by aggregating $\hat{l}$ through a majority vote. Alternatively, an individual is assigned to the class her or his clauses are most frequently assigned to.

The rest of this section provides more detail about unimodal recognition including feature extraction process and unimodal encoding (see Section 3.5.1), multimodal recognition (see Section 3.5.2), classification approach (see Section 3.5.3) and aggregation (see Section 3.5.4).

3.5.1 Unimodal Recognition

The unimodal recognition component includes two main steps—*feature extraction* and *unimodal encoding*. Since every clause includes both an audio signal and its transcription, two distinct feature sets are extracted, one from the audio and the other from the text. In both cases, the result is a sequence of feature vectors, a suitable format while using Bi-LSTMs, which acts as an encoder.

Audio Feature Extraction

The vocal tract spectral magnitude information is extracted by applying MFCCs (see Appendix A, Section .5.1). Temporal characteristics of mel-cepstral features that capture coordination in vocal tract spectral shape dynamics are obtained. In detail, the signal is segmented into 25 *ms* long analysis windows (corresponding to 551 samples) that start at regular time steps of 10 *ms* (corresponding to 220 samples) and span the entire clause (two consecutive windows overlap by 15 *ms*). This 25 *ms* length is large enough to capture enough information and yet the features inside this frame should remain relatively stationary. The Hanning windows are used as window function to remove edge effects. The values of both window length and step are standard in the literature and no other values have been tested. After the segmentation, the signal interval enclosed in every window is mapped into a feature vector where the components are *MFCC* coefficients [282]. The first 13 coefficients contain the most salient information needed for speech recognition and to represent dynamic nature of the audio, and the first- and second-order derivative of first 13 coefficients are extracted as

well. Thus, each frame in an audio signal has a total of 39 features comprising 13 MFCC coefficients, 13 first-order derivatives of MFCC and 13 second-order derivatives.

Such a representation, based on the physiology of hearing, is widely applied in the literature, and it accounts mainly for energy (how loud someone speaks) and phonetic content (what sounds someone utters). The main motivation behind its use is that it has been effective in various approaches aimed at inferring social and psychological phenomena from speech, including emotions [99], personality [324] and depression (see Section 3.3).

Correspondingly, each clause is converted into a sequence of vectors $A = (a_1, a_2, \ldots, a_{T_A})$, where a_i is the F-dimensional vector extracted from the i^{th} window. The number of frames are changed based on the length of the audio clauses; thus, the input features are truncated at its T_A^{th} element where T_A is the number of vectors allowed in A and when the input features are shorter than T_A, the input features are padded with zero vectors. The value of the hyperparameter T_A has been set through hyperparameter optimisation technique during the experiments (see section 3.6.2). Hence, every clause is mapped into a two-dimensional matrix $A \in \mathbb{R}^{T_A \times F}$.

Text Feature Extraction

Recently, transfer learning as pretrained language models has become ubiquitous in NLP and has contributed to the state-of-the-art on varying tasks. For extracting linguistic features, word embedding is utilised, where words from the vocabulary are mapped to vectors of real numbers (see Appendix A, Section .4). In this chapter, we experiment with two types of embedding for the feature extraction phase: one is based on static word representation and the other is based on contextualised representation.

- **Static Word Representation:** Conventionally, supervised lexicalised NLP approaches take a word and convert it to an index, which is then transformed into a feature vector f using a one-hot representation. Given a word $w_i \in W$, where W is a fixed-sized word vocabulary and f_i is a $|W|$-dimensional vector, where all elements are set to 0, except element k that sets to 1. The one-hot representations are then transformed into word embeddings by defining an *embedding layer*. Considering a sequence $w_1, w_2, ..., w_{T_S}$, each word $w_i \in W$ is embedded through embedding layer into a D-dimensional vector

space using the following formula:

$$s_i = E f_i. \tag{3.1}$$

Here the matrix $E \in \mathbb{R}^{|W| \times D}$ indicates all the word embeddings that are learned in this layer, the same as the other parameters of the network. Practically, we applied a look up table to substitute this computation with a simpler array indexing operation, where $E_{w_i} \in \mathbb{R}^D$ corresponds to the embedding of the word w_i. This look up table operation is then used for each word in the sequence, where the sequence in our case is the clause. Finally, the resulting word embeddings are then concatenated where

$$S = [E_{w_1}; E_{w_2}; \ldots; E_{w_{T_S}}] \in \mathbb{R}^{T_s \times D} \tag{3.2}$$

where $S = (s_1, s_2, ..., s_{T_S})$ and T_S is the maximum number of vectors allowed in the sequence S. If the number of vectors is shorter than T_S, then the corresponding sequence S is padded with zero vectors, otherwise, it is truncated at its $T_S{}^{th}$ element. The value of D is set to 100 (no other values have been tried), while the value of T_S is set through hyperparameter optimisation during the experiments (see Section 3.6.2). Consequently, S is represented as as a two-dimensional matrix $S \in \mathbb{R}^{T_S \times D}$.

- **Contextualised Word Representation:** Unlike static representation that maps every word always into the same vector, irrespectively of the different contexts in which it appears, contextualised word representation overcomes such a limitation. In this chapter, the pretrained BERT model is used, specifically the smaller BERT-base (uncased) with 12 transformer blocks and a hidden size of 768, which has 110M trainable parameters in total (see Appendix A, Section .4.2).

  To compute word features based on BERT, different methods exist for extracting the features. BERT layers capture different information; thus, the word embedding for each token can be extracted from any of these layers. The outputs from either the embedding layer, the second-to-last hidden layer or the last-hidden layer, are commonly used in the literature. However, the last layer is too closed to the target functions (i.e. masked language model and next sentence prediction) during pretraining, which is probably

biased to those targets, while the output of embedding layer may preserve the very original word information (with no fancy self-attention). Therefore, we chose to extract the features from the second-to-last hidden layer.

We obtain BERT word vectors from an open toolkit bert-as-service[1]. It is a sentence encoder service using BERT pretrained models for generating the BERT embeddings. We spin up BERT as a service server and create a client to get the embeddings. We use the uncased L-12 H-768 A-12 pretrained BERT model to generate the embeddings. Finally, the same word can have different embeddings according to the context. In this study, we average the representations of the same word to generate an approximate vector for that word. Lastly, each clause is fed to the BERT model to generate a sequence of word representations $S = (s_1, s_2, ..., s_{T_S})$, where s_i is embedded into a q-dimensional vector space and q is 768 dimensional size.

Unimodal Encoding

After the feature extraction process, the clauses are mapped into sequences of feature vectors $X = (x_1, x_2, ..., x_T)$, where X corresponds to A or S, and T corresponds to T_A or T_S, depending on whether the features have been extracted from the audio signal or from its transcription (see previous subsections). The main motivation is that the input data is sequential and, in particular, audio vectors a_t correspond to different points in time of the speech signal, while vectors s_t correspond to different words in a text. However, the vectors do not carry sequential information, i.e., they do not encode possible relationships between feature vectors extracted at different points in time. Hence, the X sequences are fed to Bi-LSTMs [125], well known to capture such relationships, if any (see Appendix A, Section .2.4 and .2.5 for LSTM and Bi-LSTM frameworks respectively).

Such a model produces the creation of the network's internal hidden state h_t to model the time series pattern. This internal hidden state is updated at each time step with the input data x_t and the hidden state of the previous time steps h_{t-1} as follows:

[1]https://github.com/hanxiao/bert-as-service

$$\begin{aligned} \overrightarrow{\mathrm{h_t}} &= \overrightarrow{\mathrm{LSTM}}(x_t, \overrightarrow{\mathrm{h}}_{t-1}) \\ \overleftarrow{\mathrm{h_t}} &= \overleftarrow{\mathrm{LSTM}}(x_t, \overleftarrow{\mathrm{h}}_{t-1}) \end{aligned} \tag{3.3}$$

where $\overrightarrow{\mathrm{h_t}}$ and $\overleftarrow{\mathrm{h_t}}$ represent the hidden states at t^{th} time step for forward and backward directions, respectively, and x_t represents the t^{th} MFCC features in a sequence A for audio model, and for text modal is the t^{th} embedded word in a sequence S. Then, we concatenate $\overrightarrow{\mathrm{h_t}}$ and $\overleftarrow{\mathrm{h_t}}$ to obtain h_t as follows:

$$h_t = [\overrightarrow{h_t} \oplus \overleftarrow{h_t}], \tag{3.4}$$

where $\oplus$ denotes a concatenating operation for forward and backward pass outputs. After concatenating the output vectors of each time step for T vectors, a matrix H is generated with the shape of $[T, 2U]$,where U denotes the hidden dimension of the LSTM network.

$$H = (h_1, h_2, ..., h_T). \tag{3.5}$$

After encoding the input sequence X with Bi-LSTM, the last hidden state, h_T, is extracted to be the representative vector that contains all of the sequential input data. This unimodal representation (h_T) is then classified by feeding h_T to a softmax layer (see Section 3.5.3).

3.5.2 Multimodal Recognition

The multimodal combination approach builds upon the unimodal representations introduced in Section 3.5.1 and implements different strategies for combining lexical and paralinguistic information extracted from the data. For more details about multimodal representation see Appendix A, Section .6. The rest of this section presents every multimodal combination approach in detail.

Late Fusion (LF)

The classification of the unimodal representations occurs by feeding the output of the encoders to a softmax layer trained to minimise the cross-entropy (see Section 3.5.3). The output of such a layer can be thought of as the *a-posteriori* probabilities $p(c|X)$ of the classes

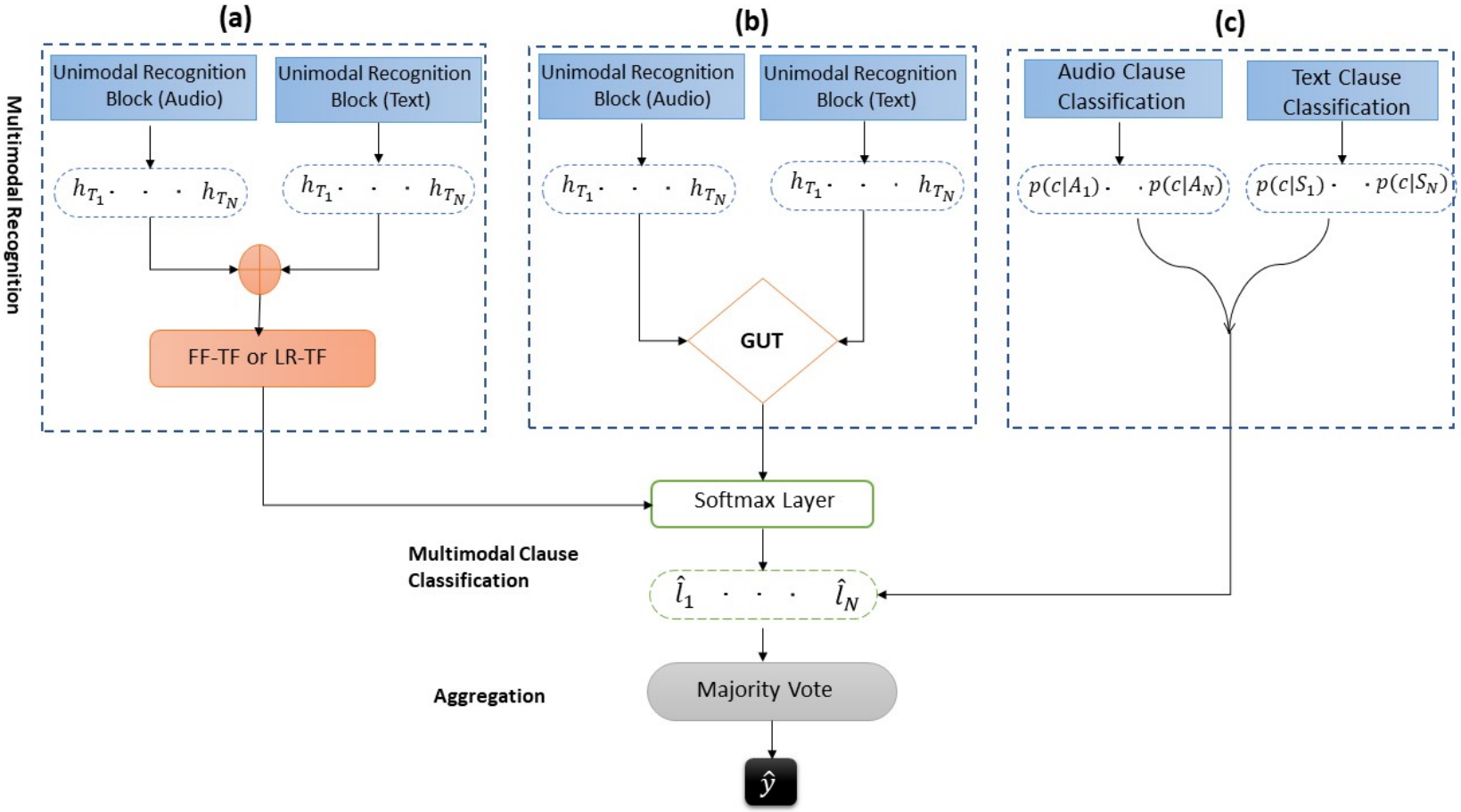

Figure 3.4: The figure shows the three strategies for the multimodal combination of linguistic and acoustic aspects of speech. (a) The Feed Forward Intermediate Fusion (FF-TF) or Logistic Regression Intermediate Fusion (LR-TF) 'fuses' the unimodal recognition (see Section 3.5.1) through a 4-layer network that takes as input the concatenation of H_T for both text and audio models or 'fuses' the unimodal representations through a logistic regression. (b) The Intermediate Fusion with Attention Gate (ATT-TF) which uses the Gated Multimodal Unit (GMU) to weight the unimodal representations according to how likely they are to induce the right classification outcome. Finally, (c) The sum rule (or late fusion) uses the unimodal posteriors as a criterion to assign a clause to a given class.

(see Appendix A, Section .6.2). Figure 3.4 (c) represents the LF approach. Based on the assumption that both modalities used in this work are equally important and that the feature vectors extracted from the different modalities are statistically independent given the class, it is possible to apply the *sum rule*, probably the most widely applied approach for the late fusion of multiple classifiers, corresponding to multiple modalities [153]:

$$\hat{l} = arg\max_{c \in \mathscr{C}} \{p(c|A) + p(c|S)\}, \tag{3.6}$$

where $\hat{l}$ is the class assigned to a clause, $\mathscr{C}$ is the set of all possible classes (*depression* and *control* in the experiments of this work), while A and S are the sequences extracted from the speech signal and its transcription, respectively (see section 3.5.1).

Early Fusion (EF)

The other typical approach for multimodal recognition is *early fusion*, i.e. the concatenation of feature vectors extracted at the same moment from multiple modalities (see Appendix A, Section .6.1). The problem is that, in the experiments of this work, there is a significant difference in the rate at which the vectors are extracted from the data. For speech signals, one vector is extracted every 10 ms, thus inducing a rate of 100 Hz, while for texts, there is one vector per word, thus inducing a rate of roughly 2 Hz (the average number of words per second). In such a situation, the application of the early fusion requires one to downsample the sequence where the rate is greater to discard information. However, it is possible to avoid such a problem by obtaining a joint representation through intermediate fusion technique as described in the next subsection.

Intermediate Fusion (TF)

Section 3.5.1 shows that the feature vector sequences extracted from the speech signal and its transcription are encoded using unimodal Bi-LSTMs that learn a representation capable to consider relationships between the vectors in the sequence, possibly accounting for temporal patterns in the data. For intermediate fusion, we can extract the internal data representation from any layer and use it as an input of another network (see Appendix A, Section .6.3).

Recall that the input of the text model is the raw data and the word embedding is used as its first representation. Then, Bi-LSTM network is applied on the embedding to produce a

second internal representation of the input text. However, the feature extraction by MFCCs is used as a first representation for the audio input signal. Then, Bi-LSTM network is applied on the extracted features to produce a second internal representation of the signal. Since the neural networks learn representation with each successive layer [34], we extracted the latent features from the last layer before the softmax layer, which is the last hidden state of Bi-LSTM (h_T), as a learned representation for each input sequence of text and audio.

The two vectors resulting from such a process are L2-normalised due to the different nature of the two modalities. This makes features from different modalities compatible to be combined. and the bias towards some features rather than others is reduced. The normalised features are then fused according to multiple strategies. The first, referred to as *Feed Forward Intermediate Fusion (FF–TF)* in the following, corresponds to concatenating the unimodal recognitions and feeding them to a feedforward network with four hidden layers (128, 64, 32 and 16 neurons, respectively). The expected effect of the hidden layers is to embed the encodings in a new, multimodal space more suitable for discriminating between depression and control participants.

Likewise, the second fusion strategy, referred to as *Logistic Regression Intermediate Fusion* (LR–TF),works by feeding the concatenation of the unimodal recognitions to a Logistic Regression function trained to maximise the classification accuracy. Both FF–TF and LR–TF are represented in Figure 3.4 (a).

In both cases above, the assumption is that both modalities are equally effective at discriminating between depressed and control participants. However, this is not necessarily the case and, therefore, the last intermediate fusion strategy uses a *Gated Multimodal Unit* (GMU), and it is referred to as *Intermediate Fusion with Attention Gate* (ATT–TF) as described in Figure 3.4 (b). The GMU is a processing block that weighs the different modalities through a self-attention mechanism [20]. It learns to weigh the representations of the two modalities and, in particular, to increase the weight of the modality that appears to carry depression-relevant information. If h_a and h_s are the encodings of speech signal and its transcription, respectively, the fusion is performed through a non-linear transformation that works according to the following equations:

$$x_a = tanh(W_a h_a) \tag{3.7}$$

$$x_s = tanh(W_s.h_s) \tag{3.8}$$

$$z = \sigma(W_z.[h_a \oplus h_s]) \tag{3.9}$$

$$h_T = z * x_a + (1-z) * x_s, \tag{3.10}$$

where W_a, W_s and W_z are learnable parameters and $\oplus$ is the concatenation operator. The values of z and $1-z$ can be thought of as weights that account for the contribution of the different modalities to the final classification outcome.

3.5.3 Clause Classification

All representations, whether unimodal (see Section 3.5.1) or multimodal (see Section 3.5.2), are fed to a fully connected *softmax* layer that implements the following equation:

$$\hat{l} = \sigma(Wh_T + b), \tag{3.11}$$

Where σ is the softmax function, h_T is the unimodal/multimodal representations, W is the weight matrix, and b is a bias vector. Both W and b are learned through a training process aimed at the minimization of the cross-entropy between groundtruth and classification outcome [70]:

$$\mathscr{L}(\mathscr{X}) = -\frac{1}{M}\sum_{m=1}^{M}[l_m \log \sigma(\hat{l_m}) + (1-l_m)\ \log(1-\sigma(\hat{l_m}))], \tag{3.12}$$

where $\mathscr{X}$ is the training set, M is the total number of samples in $\mathscr{X}$, l_m is the ground truth of training sample m, and $\hat{l}_m$ is the classification outcome for the same sample. The training takes place through back-propagation using gradient clipping to alleviate the exploding gradient problem [231].

3.5.4 Aggregation

As mention before, an individual has uttered many clauses. The clause classification step (see section 3.5.3) processes each of them independently so that, for an individual that has N clauses, there are N independent classifications $\hat{l_1},..,\hat{l_N}$. These are aggregated into a single classification outcome $\hat{y}$ through a majority voting, meaning that the individual is assigned to

the class most frequently represented among the $\hat{l}_k$ values.

$$\hat{y} = \arg\max_{c \in \mathscr{C}} n(c), \tag{3.13}$$

where $n(c)$ is the number of clauses assigned to class c and $\mathscr{C}$ is the set of all classes (depressed and control in the experiments of this work).

3.6 Experiments and Results

The goal of the experiments is to build an objective multimodal approach that can help psychiatrists to distinguish between the cases. To understand more about different depression aspects, we developed several fusion approaches to investigate whether it is what people say, how people say it or both of them. We analysed deeply the performance of the system by conducting several experiments to answer the research questions proposed in Section 3.1. This section presents the setting of the hyperparameters of the experiments and then the experimental results.

3.6.1 Issues with Existing Datasets

There are many existing datasets used in the literature yet they differ in assessing depression. The available datasets are used self-assessment tests which lack of the clinical judgment of the doctors. This is different from our research's objective thus might affect the scale of depression and results. Moreover, the presence of a control group is uncommon in the available datasets. The matching between depressed and control groups in terms of age, education level and gender is also captured. For these reasons, we use our data that assesses depression based on clinical interviews where the distinction between depressed and non-depressed participants has been made by psychiatrists and not through the administration of self-assessment questionnaires. This is an advantage because it increases the chances of the data to be representative of the actual difference between depressed and non-depressed speakers. Due to these differences we do not compare our results with the existing works.

3.6.2 Hyperparameter Setting

The experiments were performed according to a k-fold experimental design (for more details see Appendix A, Section .3). The participants were randomly split into $k = 5$ disjoint subsets, and the clauses uttered by all participants in $k-1$ groups were used as training set. Correspondingly, the clauses uttered by the participants in the left out subset were used for test. The process was repeated k times and, at each repetition, a different subset discarded, making it possible to perform experiments over the whole corpus at disposition, while still keeping separated training and test set. Another advantage of the setup is that the experiments are person independent, meaning that the same participant is never represented in both training and test set. This excludes that what the approach recognises is the identity of the participants and not their condition.

Every time a fold is used as a test set, the union of the remaining four is split into training set (90% of the material) and validation set (10% of the material). This latter is used to select the value of the hyperparameters through *hyperparameter optimisation* (for more details about different hyperparameter optimisation see Appendix A, Section .3). Automatic hyperparameter tuning approach with Bayesian optimisation [225] was applied to conduct a guided search for the best hyperparameters. The combination showing the highest accuracy over the validation set was retained and used to classify the samples of the test set. The spaces of the hyperparameters that were searched are *learning rate* α_0 (a factor that influences the size of the parameter updates during training), *number of neurons in the hidden layer* U, *number of training epochs* P (the number of cycles through which the network is trained), *batch size* B (number of training samples used at any training epoch) and *padding* T (length of the vector sequences fed to the networks).

During the experiments, the predefined sets used for the different hyper-parameters are as follows: for the learning rate, the values are 10^{-3}, 3×10^{-3}, 10^{-2}, and 10^{-1}. For the hidden layers, the number of neurons is 32, 64 or 128. The training epochs are 30, 50 or 80 and the samples in the batch size are 32, 64 or 128. Finally, the padding values for speech are 40, 50, 60, 70, 80, 100 and 120, while those for text are the integers between 9 and 14. In all cases, the main motivation behind the choice of the values is that they are considered standard in the literature. The only exception is padding that depends on the type of data being used. The training has been performed through backpropagation using the Adam optimiser and

Table 3.3: The table reports accuracy, precisioN and recall for the two embedding techniques used in the experiments, at the level of both individual clauses and participants. The values in the table are the averages obtained over 30 repetitions of the experiment.

Embedding	**Level**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**
Wikipedia2vec	Clause	60.4 ± 0.003	56.1 ± 0.005	46.5 ± 0.007
Wikipedia2vec	Participant	74.1 ± 0.023	100.0 ± 0.000	47.4 ± 0.047
BERT	Clause	60.4 ± 0.006	56.0 ± 0.008	47.4 ± 0.008
BERT	Participant	73.3 ± 0.021	100.0 ± 0.000	46.0 ± 0.031

categorical cross-entropy as a loss function[2].

For the unimodal approaches, according to a practice common in the literature, the initial learning rate was progressively reduced over successive training epochs using the expression $\alpha = \alpha_0 \beta^{\phi/\delta}$, where β= 0.96 is the decay rate, ϕ is the step and $\delta = 500$ is the number of decay steps. For the text model, the highest validation accuracy was obtained for $\alpha_0 = 0.003$, $P = 80$, $B = 64$, $U = 128$ and $T = 10$. For the static word embedding, the experiments of this work are based on a version of *Wikipedia2vec* pre-trained word embedding on a corpus of Italian texts, *'itwiki'*, including Wikipedia articles written in Italian, which is based on a 100-dimensional embedding space (see Appendix A, Section .4.1). However, for contextualised word embedding, uncased multilingual-L-12-H-768-A-12 pretrained BERT is used with 12 transformer blocks and a hidden size of 768, which has 110M trainable parameters in total. Regarding the audio model, the hyperparameter values inducing the highest validation accuracy were $\alpha_0 = 0.001$, $P = 80$, $B = 32$, $U = 128$ and $T = 40$.

For the multimodal approaches (FF-TF and ATT-TF), the hyperparameter values maximising the validation accuracy are $\alpha_0 = 0.003$ and $B = 128$. For FF-TF, the number of neurons in the 4 layers of the network is 128, 64, 32 and 16 (the values have been set a-priori and not through hyperparameter optimisation). For ATT-TF, the size of the hidden layer in the gate is 27.

Given the small number of observations used in my research, the network could risk overfitting to the training set. The overfitting issue makes it hard to generalise to bigger datasets. Therefore, we applied several techniques that reduce the overfitting problem such as cross-validation, early-features and regularization, which comes down to adding a cost to the loss function for large weights.

[2]All models and training methodologies were implemented with Tensorflow.

Table 3.4: The table shows the performance of unimodal and multimodal approaches used in the experiments, at both clause and participant level. The values are reported regarding the averages obtained over 30 repetitions of the experiments and their standard errors.

Approach		**Level**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**	**AUC (%)**
Unimodal	Text	Clause	60.4 ± 0.003	56.1 ± 0.005	46.5 ± 0.007	51.0 ± 0.005	59.0 ± 0.003
		Participant	74.1 ± 0.023	100.0 ± 0.000	47.4 ± 0.047	64.1 ± 0.045	73.7 ± 0.023
	Audio	Clause	70.0 ± 0.006	65.1 ± 0.008	65.0 ± 0.008	65.0 ± 0.007	69.0 ± 0.006
		Participant	73.0 ± 0.021	76.0 ± 0.031	66.3 ± 0.031	71.0 ± 0.024	73.0 ± 0.021
Multimodal	LF	Clause	64.0 ± 0.004	60.0 ± 0.006	54.3 ± 0.008	57.0 ± 0.005	63.0 ± 0.004
		Participant	83.0 ± 0.036	94.0 ± 0.032	69.4 ± 0.068	80.0 ± 0.049	83.0 ± 0.036
	FF-TF	Clause	64.0 ± 0.004	59.3 ± 0.006	55.2 ± 0.008	57.2 ± 0.005	62.7 ± 0.004
		Participant	83.0 ± 0.027	93.1 ± 0.030	71.0 ± 0.043	80.1 ± 0.034	83.0 ± 0.027
	LR-TF	Clause	68.0 ± 0.006	64.3 ± 0.008	60.0 ± 0.010	62.0 ± 0.008	67.0 ± 0.006
		Participant	78.4 ± 0.021	85.0 ± 0.029	68.3 ± 0.033	76.0 ± 0.025	78.2 ± 0.021
	ATT-TF	Clause	63.0 ± 0.004	58.1 ± 0.005	54.5 ± 0.010	56.2 ± 0.007	62.0 ± 0.004
		Participant	83.5 ± 0.031	95.0 ± 0.025	70.3 ± 0.058	80.5 ± 0.042	83.2 ± 0.031

3.6.3 Recognition Results

Tables 3.3 and 3.4 show the performance at the level of both individual clauses and participants, i.e. after that a majority vote was applied to all clauses uttered by a given participant. Given that the initialisation of the network weights takes place through a random process, every experiment was repeated 30 times and, therefore, the results were reported regarding average and standard deviation of the different performance metrics. The limited variance across the 30 repetitions, suggests that the models are sufficiently robust to changes in the initialisation and, therefore, the averages can be considered realistic estimates of the performance. According to a two-tailed t-test with Bonferroni correction, the accuracy is always better than chance to a statistically significant extent, at the level of both clauses and participants.

Table 3.3 shows accuracy, precision and recall for the text unimodal approach. According to a binomial test, the accuracy is better than chance to a statistically significant extent in all cases ($p < 10^3$). We conducted an experiment to compare between contextualised and static embedding to see if the contextualised sophisticated word embedding can contribute in such a problem. Using BERT rather than Wikipedia2vec for the word embedding (see Section 3.5.1) does not induce performance improvements, possibly because there is a mismatch between the dictionary used during the currently available versions of BERT for Italian texts and the dictionary of the data used in this work. This could be because the version of multilingual BERT-base has limited vocabulary size compared to the English one. Another probable reason is that the clauses are short (the average length is 3.9 words) and, therefore, the context might not carry sufficient information. Thus, the experiments rely solely on using Wikipedia2vec.

Table 3.4 shows the depression detection results obtained with unimodal and multimodal approaches at the level of both clauses and participants. For unimodal approaches, the audio-based classifier outperforms the text-based at the clause level and, according to a two-tailed t-test, the difference is significant ($p < 0.05$). This disagrees with other works of the literature [264, 343], acoustic aspects of speech appear to be more effective than linguistic in conveying depression relevant information (despite the clauses having been transcribed manually). One possible explanation is that approaches based on language are difficult when tackling short linguistic units like clauses (the average number of words is 3.9). However,

another possible reason is that *paralinguistics* (how things are said) might be a cue more honest than *lexical choice* (what people say), at least regarding the features used in this work. This follows the observation of social psychology that nonverbal behaviour, being displayed outside conscious awareness, tends to convey more reliable information about the inner state of an individual [235].

However, from an application viewpoint, the most important metrics are those at the participant level and, in this case, the difference between audio and text is not statistically significant. At the clause level, multimodal approaches perform roughly like unimodal, but for participant level, multimodal approaches outperform unimodal which achieves an accuracy of 83.5% and F1 measure of 80.5% (for the best approach), meaning that it correctly distinguishes between depressed and non-depressed speakers roughly 4 times out of 5. In particular, the best multimodal approach improves over the best unimodal system by 9.4 points. This means that supplementing linguistic cues with their paralinguistic cues substantially improves the performance. The difference concerning the best unimodal approach is always statistically significant ($p < 0.05$ according to a binomial test), except for LR-FT. One possible explanation is that the unimodal encoders (see Section 3.5.1) capture temporal patterns in their respective input data, but represent them in a space where the difference between depression and control participants does not emerge with sufficient clarity. In this respect, the multilayer network used in FF-TF to embed the unimodal encodings in a space where there is more difference between depression and control participants appears to induce higher person level accuracy. ATT-TF disregards the four layers network, but it still achieves the same person level accuracy as FF-TF. In this case, the probable explanation is that the GMU effectively identifies the modality likelier to carry information inducing the correct classification.

It is worth noting that several techniques have been also applied to the data such as CNN, RNN and Logistic regression. Like LSTM, these are proven techniques and have been used extensively in the literature. There were no statistically significant difference exist between our results and other techniques' results therefore, we do not mention them.

3.7 Conclusion

This chapter presented experiments on depression detection for combining speech signals and their transcriptions. The experiments were performed using varying approaches aimed at fusing multiple modalities, including synergising unimodal classifiers through the sum rule, one of the most traditional approaches for combining multiple classifiers, and network based approaches for the intermediate fusion of multiple modalities, one of the most recent trends in multimodal behaviour analysis.

Experiments presented herein explored a new set of data involving 29 depression patients and 30 persons that have never experienced mental health issues. The distinction between depressed and control participants was made by professional psychiatrists and not through the administration of self-assessment questionnaires. This is an important advantage because it makes it more likely that the proposed approaches actually learn to detect depression.

The experimental results show that acoustic aspects of speech (how people say) appear to be more effective than linguistic (what they say) in conveying depression relevant information. However, the combined aspects achieve the best accuracy of 83.5% (F1 measure of 80.5%) at the person level, meaning that it correctly distinguishes between depressed and non-depressed speakers roughly 4 times out of 5. This chapter also discussed about advanced text embedding (BERT) that considers the context of the word. The results show that it did not improve the data used for the experiments. In the next chapter, we present our comprehensive analysis for depression detection.

Chapter 4

A Comprehensive Analysis for Depression Detection System

4.1 Motivation

The previous chapter introduced the state-of-the-art methodologies for joint modelling of linguistic and acoustic aspects of speech (corresponding to what people say and how they say it, respectively). The results show that, to a statistically significant extent, the multimodal approaches outperform unimodal approaches. However, using similar approaches (e.g. the joint modelling of linguistic and acoustic aspects of speech [5, 211, 264]) over different data may not necessarily induce the same conclusions about how effective using a certain modality is regarding the others. This is true since the state-of-the-art is uncertain in identifying the best way to detect depression (see Chapter 3, Section 3.2). One possible reason is that several works ignore the problem of identifying people diagnosed with depression by a doctor, like in this work, but the problem of inferring self-assessment scores. These are affected by different biases (see Chapter 3, Section 3.2) and, therefore, can induce ambiguous results. Furthermore, depression is a complex phenomenon involving various factors (e.g. physiology, socio-economic status, age, gender, etc. [142]) that induce individual differences in the way people manifest the pathology. Hence, it is important to analyse deeply the performance of unimodal and multimodal methodologies, showing to what extent the majority vote is more beneficial for a certain approach and how and when the synergised multiple modalities can be addressed by investigating the possible differences between depressed and non-depressed speakers in the modalities through which one's condition is manifested.

Adding to the above, the results presented in the Chapter 3 suggest that the proposed approaches can make the right decision about an individual around 4 times out of 5, but it is unclear whether this can be considered satisfactory regarding possible clinical applications. Therefore, we illustrate several application scenarios where the proposed approaches use confidence measures that identify the likeliest cases to be correctly classified. Thus, the work of psychiatric and counselling services can be supported by possibly allowing doctors to focus on ambiguous and difficult cases while leaving the machines to tackle the most evident ones. This is important because it can increase the efficiency of screening services and, correspondingly, reduce the costs associated with depression diagnosis.

Computing efforts made so far have targeted mainly the improvement of the detection performance and have addressed, only to a limited extent, if at all, the problem of how much data is necessary to make a reliable decision about an individual (see Section 3.3). Such a problem is mainly important because realistic application scenarios require one to tackle recordings that contain only a few words (e.g. the use of data collected at help lines [140]). This is especially true as the tendency of depressed individuals is to avoid social interactions and to speak less than non-depressed people [44, 118]. Furthermore, when the speech data is obtained through interviews or other forms of interaction that involve medical personnel, reducing the amount of time necessary to gather enough information lowers the costs associated with depression diagnosis. Hence, this chapter also investigates the relationship between performance and number of clauses (amount of time). In particular, the chapter presents the effectiveness of the recall measure when considering less than 10 seconds of speech (less than 8 clauses) compared to the one obtained using the whole data at disposition. This is imperative because recall measures the effectiveness at recognising all depressed individuals as such, i.e. at avoiding type II errors (classifying a depressed individual as healthy), those inducing the most negative consequences from a clinical viewpoint. Concerning a type I error (a control individual classified as depressed), the consequence is that a healthy individual will be examined more thoroughly by doctors, but such an extra medical attention will be harmless. In contrast, for type II error, a depressed individual will go undetected and will not undergo proper treatment, thus joining the estimated 79% of depression patients without appropriate care [148], a major issue in nowadays psychiatry.

In this chapter, the research questions and subsequent novel contributions are the follow-

ing:

1. To what extent is the majority vote more beneficial for a certain approach? (see Section 4.2.1)

2. In which channel does the depressed people manifest their pathology more clearly? And is it the same for healthy people? (see Sections 4.2.2 and 4.2.3)

3. Are the results considered satisfactory regarding possible clinical applications? (see Section 4.3)

4. Is it possible to detect depression in less than 10 seconds and, if yes, what is the impact of speaking time on depression detection sensitivity? (see Section 4.4)

The rest of this chapter is organised as follows: Section 4.2 analyses multimodal recognition, Section 4.3 describes different application scenarios, Section 4.4 presents the depression detection based on number of clauses, and the final Section 4.5 concludes the chapter.

4.2 Experiment 1: The Analysis of Multimodal Recognition

In this section, we develop a set of experiments that deeply analyses the performance of unimodal and multimodal methodologies. Specifically, it shows to what extent the majority vote is more beneficial for a certain approach. It also discusses how and when the combined multiple modalities can address and how effective using a certain modality is regarding the others. We further analyse GMU to identify the modality that contributes most to depression detection. The experiments in Sections 4.2.1, 4.2.2 and 4.3 are presented in the published work[1]. The experiments in Section 4.2.3 are presented in the published work[2].

4.2.1 The Application of Majority Vote

The results in Chapter 3, Section 3.6.3 show that the application of the majority vote allows one to achieve high participant level accuracy, especially for multimodal approaches. This

[1]Aloshban, Nujud, Anna Esposito, and Alessandro Vinciarelli. "What You Say or How You Say It? Depression Detection Through Joint Modeling of Linguistic and Acoustic Aspects of Speech." Cognitive Computation (2021): 1-14."

[2]Aloshban, Nujud, Anna Esposito, and Alessandro Vinciarelli. "Language or Paralanguage, This is the Problem: Comparing Depressed and Non-Depressed Speakers Through the Analysis of Gated Multimodal Units." In INTERSPEECH (2021).

result, probably, is mainly because the average number of clauses per participant is greater than 100 for both depression and control participants (see Chapter 3, Section 3.2). Therefore, a limited accuracy at the clause level is sufficient to increase the probability of at least half of the clauses being classified correctly, the condition for a participant being assigned to the right class. According to Table 3.4 in Chapter 3, multimodal approaches outperform unimodal regarding person level accuracy, the metric that matters from an application viewpoint. However, it is the speech-based unimodal approach that shows the highest clause level accuracy. Overall, this means that multimodal approaches considerably benefit from the majority vote. Figure 4.1, showing the individual clause level accuracies in descending order, possibly explains observation. In particular, the figure shows that, concerning the multimodal approaches, correctly classified clauses tend to distribute more uniformly across participants. This induces numerous cases in which the accuracy is above 50% (the condition for the majority vote to work).

While possibly explaining why the majority vote is more beneficial for certain approaches, the observations above did not show to what extent the benefit can be considered satisfactory. One way to do it is to consider the accuracy gain $\Delta\alpha$:

$$\Delta\alpha = \frac{\alpha - \alpha_{min}}{\alpha_{max} - \alpha_{min}}, \tag{4.1}$$

where α is the person level accuracy actually observed after the majority vote, and α_{min} and α_{max} are minimum and maximum person level accuracy a majority vote can induce. Person level accuracy α_{min} can be observed when all correctly classified clauses concentrate in the smallest possible number of participants. In contrast, the maximum value α_{max} can be observed when the clause level accuracy is the same for all participants. Given that the clause level accuracy α_c can be consider the probability of making the right decision about a clause, α_{max} can be estimated as the probability of having more than half of the clauses classed correctly:

$$\alpha_{max} \simeq \sum_{k=M/2+1}^{M} \binom{M}{k} \alpha_c^k (1-\alpha_c)^{M-k}, \tag{4.2}$$

where M is the average number of clauses per participant (114 in the data of this work).

Table 4.1 shows the results for the different approaches and, in particular, it shows that

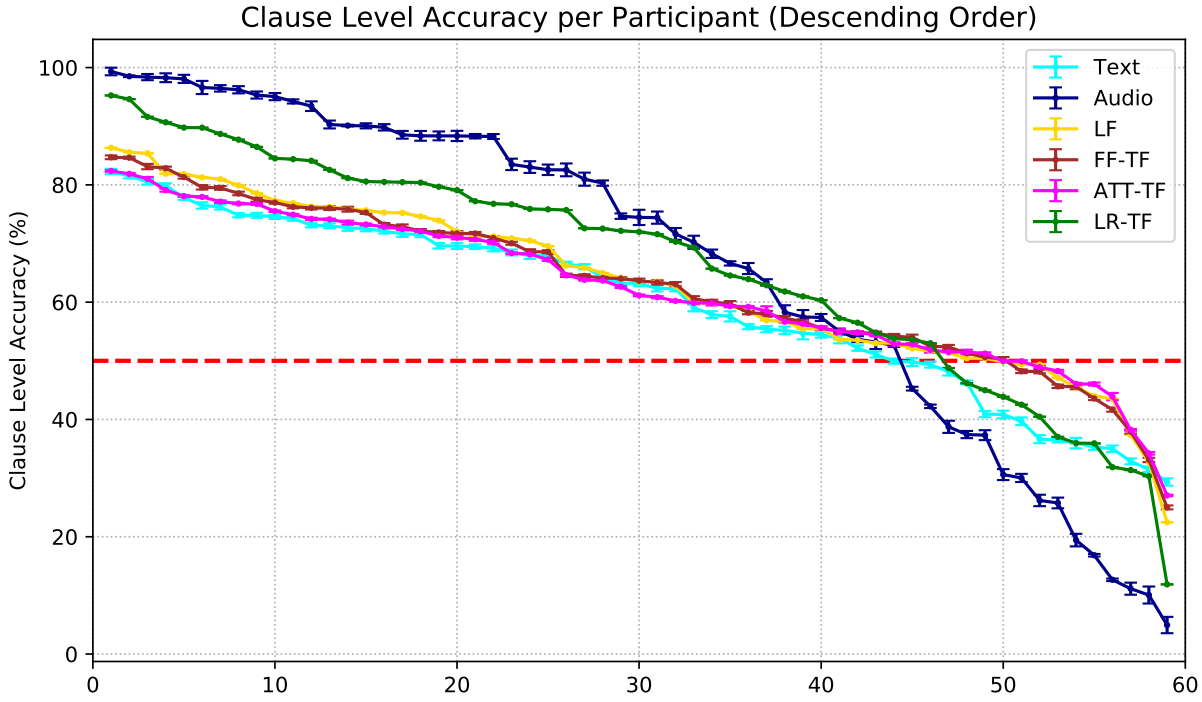

Figure 4.1: The figure shows, in descending order, the clause level accuracy per participant. The curves corresponding to the multimodal approaches intersect the 50% horizontal line later. This means that correctly classified clauses tend to be distributed across a greater number of participants and, consequently, there is a greater number of cases in which the majority vote induces a correct person classification. The acronyms LF, FF-TF, LR-TF and ATT-TF stand for *Late Fusion*, *Feed Forward Intermediate Fusion*, *Intermediate Fusion with Logistic Regression* and *Intermediate Fusion with Attention Gate*, respectively.

multimodal ones tend to obtain higher accuracy gains. Alternatively, achieving high clause level accuracy is insufficient to correctly classify the participants. It is also necessary that the distribution of correctly classified clauses allows one to achieve a clause accuracy exceeding 50% for the largest possible number of participants. This is important because the networks are trained to maximise the clause level accuracy, but not to perform uniformly across participants. Therefore, there is a possible misalignment between the way the models are trained and the actual goal of the approaches. The next subsection shows how and when the combined multiple modalities can address, at least to a partial extent, such a problem.

Table 4.1: The table shows the accuracy gain $\Delta\alpha$ for the different approaches used in the experiments. The values α_{min} and α_{max} are minimum and maximum accuracy that can result from the application of the majority vote, respectively.

Modality	α_{min} (%)	α_{max} (%)	$\Delta\alpha$ (%)
Text	47.4	98.4	52.3
Audio	55.9	100	41.5
LF	49.1	99.5	63.5
FF-TF	49.1	99.7	67.0
LR-TF	49.1		
ATT-TF	49.1	99.7	67.0

4.2.2 The Combination of Multiple Modalities

The previous section revealed that the unimodal approaches tend to concentrate correctly classified clauses for few participants. One possible explanation is that some of the participants tend to consistently manifest their condition through at least one of the modalities. In this way, they leave detectable traces of their condition in many clauses, thereby enabling high accuracy to be easily achieved at the clause level. The participants tending to do this only through one modality are likely to inject *diversity* [252], i.e. to lead the unimodal classifiers to make different mistakes over different participants. This is advantageous because a multimodal approach can be beneficial mainly when unimodal approaches disagree and, hence, one of these has a chance to compensate for the errors of the other.

Following the above, one way to measure the diversity is to compare N_d, the number of times the two unimodal approaches classify differently the same participant, with its upper bound, i.e. with the number N_{max} of disagreements expected when the two unimodal approaches are statistically independent. According to the data, $N_d = 21$, while N_{max} can be estimated as follows (the accuracy can be considered the probability of making the right decision about a participant):

$$N_{max} = [\alpha_1(1-\alpha_2) + \alpha_2(1-\alpha_1)]N, \tag{4.3}$$

where α_1 and α_2 are the person level accuracies of the two unimodal approaches and $N = 59$ is the total number of participants. From the results of Table 3.4 in Chapter 3, $N_{max} =$

Table 4.2: The table considers the 21 cases (out of the total 59) for which there is disagreement between the two unimodal approaches. When the audio-based approach is the correct one, the classified participant is always depressed. In contrast, when it is the text-based approach to be the correct one, the distribution of the participants across the classes is roughly uniform. One explanation is that, whenevr depressed people tend to manifest their condition through only one modality, they tend to do it through audio, i.e., through the way they speak.

Correct Modality	Depressed	Control
Audio	11	0
Text	4	6

23, meaning that N_d is 91.3% of its upper bound and the unimodal approaches appear to be highly diverse.

The results above suggest that a significant fraction of participants (N_d corresponds to 35.6% of the total) tend to manifest their condition either through one modality or the other. In particular, Table 4.2 shows that depressed participants tend to manifest their pathology rather clearly through the way they speak, while doing it more ambiguously through the words they use (hence the high recall of the audio-based unimodal approach). All multimodal systems show significantly higher person level accuracy probably because of the resulting diversity of the unimodal approaches (see Table 3.4 in Chapter 3.4). In particular, the best multimodal approach improves over the best unimodal system by 9.4 points. Furhermore, the person level accuracy difference between multimodal and unimodal approaches is always statistically significant ($p < 0.05$ according to a two-tailed t-test).

The problem left open in the above is to what extent the improvement resulting from the application of multimodal approaches can be considered satisfactory. Therefore, it is possible to estimate how close is the performance of the multimodal approaches to α_{max}, the upper bound of the accuracy that can be estimated as follows (it is the probability of at least one of the two unimodal approaches making the right decisions and, hence, giving the combination a chance to make the right decision too):

$$\alpha_{max} = 1 - (1 - \alpha_1)(1 - \alpha_2), \tag{4.4}$$

where α_1 and α_2 are the person level accuracies of the unimodal approaches. The value of α_{max} is 93.0% and, therefore, the person level accuracy of the multimodal approaches ranges

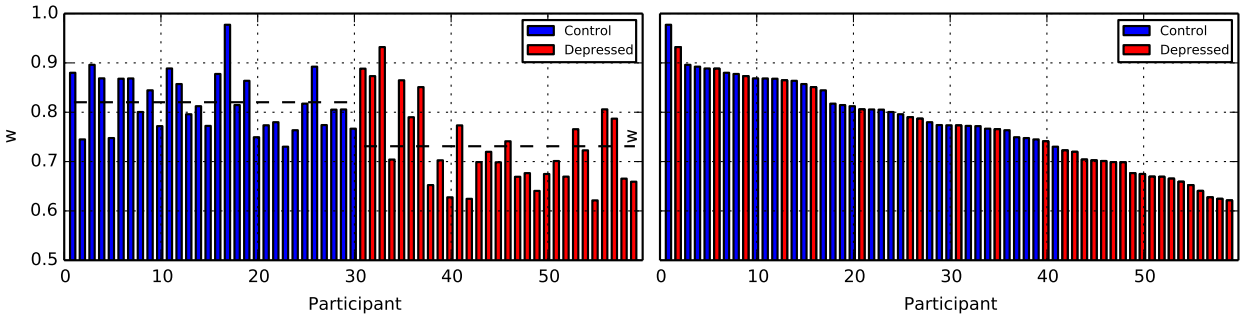

Figure 4.2: The left chart shows the w ratio for all participants (the horizontal dashed lines correspond to the average w values for control and depressed participants). The right chart shows the same w values in descending order.

between 89.2% and 90.0% of α_{max}, the maximum that can be obtained with the two unimodal approaches at disposition. In particular, given that the best unimodal approach achieves an accuracy of 74.1%, the improvement by 9.4 points (see above) corresponds to 54.3% of the maximum improvement that can be achieved. Alternatively, roughly half of the times, the two modalities disagree, the one inducing the correct classification compensates for the error of the other.

4.2.3 The Analysis of Gated Multimodal Units

Previous Section 4.2.2 revealed that the diversity occurs because participants belonging to a given class tend to manifest their condition through one modality, while those belonging to the other class tend to do it through the other modality. Hence, we analyse GMU that trained to weigh modalities according to how effectively they account for the condition of a speaker (depressed or non-depressed). The left chart in Figure 4.2 shows, for every participant, the value of the ratio $w = w_l/w_p$, where w_l and w_p are the weights that the GMU assigns to language and paralanguage, respectively. The higher such a ratio, the more the GMU considers language to convey reliable information and vice versa. The value of w is always below 1, thus suggesting that paralanguage tends to play a more important role than language in depression detection (at least for the data of this work). However, the average w value for control participants is 0.82, while it is 0.73 for depressed. Such a difference is statistically significant ($p < 10^{-5}$ according to a two-tailed t-test), suggesting that, on average, language plays a more important role for control participants compared with the depressed, and this confirms our finding in the previous Section 4.2.2.

The right chart of Figure 4.2 shows the w values in descending order and further confirms the observations above. In particular, the chart shows that the lowest 18 values correspond to depressed participants, thereby suggesting that roughly two thirds of these latter (18 out of the total 29) can be correctly identified by simply finding the speakers for which w is below or equal to a threshold corresponding to the 18^{th} value from the bottom. Alternatively, the w value can possibly be used as a confidence score when a speaker is classified as depressed. The remaining 11 depressed speakers distribute roughly uniformly across the rest of the chart. However, it can be observed that 15 of the speakers corresponding to the top 20 w values are non-depressed, thus confirming the tendency of the GMU to assign higher weights to language for control participants.

This experiment aims to identify the modality that contributes most to depression detection, and the results show that, at least for the data used in this work, it is paralanguage to consistently be assigned the higher weight, possibly because the proposed approach is based on the recognition of clauses, sentences that include only a few words (less than 10, on average). Therefore, the input texts might be too short for text modelling approaches to achieve their best results. However, the most interesting observation is that the ratio w between the weights of language and paralanguage is higher, to a statistically significant extent, regarding non-depressed speakers. This suggests that the role of language is likely to be more important for control participants than depressed ones.

To summarise the experiments of Section 4.2, the main differences between unimodal and multimodal methodologies is that the latter tend to have more uniform clause level accuracy across the participants. This is important because it induces higher person level accuracy, the metric that actually matters from an application viewpoint (see Figure 4.1). Such a result stems from the tendency of certain participants, in particular depressed ones, to manifest their condition either through what they say or how they say it, but not through both. However, control participants seem to manifest their condition much better in language through the word they use. To our knowledge, this observation remains unknown in the literature. But it is an important aspect of this work because it is a source of diversity across the unimodal approaches, and it is thanks to such a property that these disagree about a participant roughly

one third of the times. In this way, the correct unimodal approach can compensate for the error of the other, the key assumption underlying multimodal methodologies.

The observations above suggest that it is the behaviour of the participants, at least to a certain extent, that determines the conditions for the approaches to work. This is important because it might explain why the state-of-the-art is uncertain in identifying the best way to detect depression (see Chapter 3, Section 3.3). The experiments of this section suggest that the modality carrying the most reliable information can be different for people belonging to different classes. This further confirms that the best strategy is not necessarily looking for the best modality but for a set of modalities that cover all groups of people appearing in the data. The way people manifest depression can change significantly from one individual to the other, depending on numerous social, psychological, economic and cultural factors [142]. Thus, none of the behaviours considered in the literature (facial expressions, paralinguistics, body movements, etc.) appear to clearly outperform the others.

4.3 Experiment 2: Application Scenarios

The results presented so far in Chapter 3 suggest that the proposed approaches can make the right decision about an individual around 4 times out of 5, but it is unclear whether this can be considered satisfactory. In this section, the experiment is conducted to identify the cases in which the outcome of a system can be trusted, while leaving the others to medical attention. The experiment illustrates several application scenarios where the proposed approaches use confidence measures that identify the likeliest cases to be correctly classified.

One possible benchmark for comparison is the performance of *General Practitioners* (GP), the doctors who are the first line of intervention against depression, especially for convincing patients to seek for treatment. According to a meta-analysis of the literature, *sensitivity*[3] and *specificity*[4] of GPs are in the ranges 41.3% to 59.0% and 74.5% to 87.3%, respectively [204]. This corresponds to an accuracy between 57.9% and 73.1% for the data used in the experiments of this work.

Following the above, all approaches proposed in this work appear to perform comparably to an average GP, especially about sensitivity (the name of recall in medical domains). Such

[3]Percentage of depressed individuals actually diagnosed as such (equivalent to recall).
[4]Percentage of non-depressed individuals actually diagnosed as such.

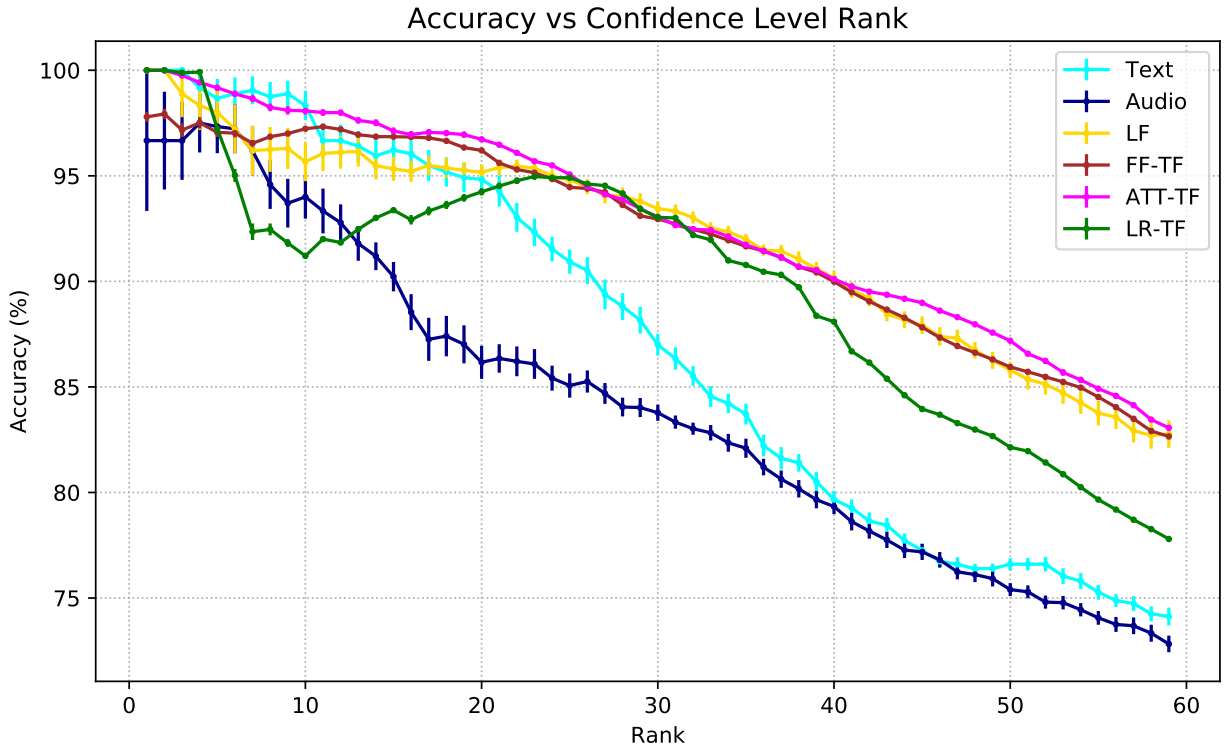

Figure 4.3: The plots show the accuracy when considering only the r persons with the highest confidence values. On average, multimodal approaches appear to have higher accuracy for every value of r and, in particular, they appear to have accuracy at least 90% when considering the 40 top ranking participants. Alternatively, it is possible to automatically isolate two thirds of the participants for which the system decides correctly 9 times out of 10. The acronyms LF, FF-TF, LR-TF and ATT-TF stand for *Late Fusion*, *Feed Forward Intermediate Fusion*, *Intermediate Fusion with Logistic Regression* and *Intermediate Fusion with Attention Gate*, respectively.

a measure is particularly important because Type II errors (classifying a depressed person as non-depressed) are those with the most negative consequences and, therefore, should be as limited as possible. This suggests that one possible approach for the application of depression detection technologies is to identify the cases in which the outcome of a system can be trusted, while leaving the others to medical attention. This appears to follow recent trends suggesting that AI-driven technologies should collaborate with their users and not simply replace them [83].

One way to address the problem above is to consider only those participants for which the two unimodal approaches agree with each other. The rationale is that agreement between multiple modalities might correspond to higher confidence and, correspondingly, to higher

performance. In our experiments, the unimodal approaches agree 38 times out of 59 (corresponding to 64.4% of the total participants), 24 times over a control participant and 14 over a depressed one. In 33 of the 38 cases, both approaches are correct (corresponding to an accuracy of 86.8%). In the remaining 5 cases, the participants are always depressed, thus inducing a recall of 64.3%. This means that filtering the participants according to agreement between modalities increases accuracy while keeping the sensitivity at the level of an average GP. Consequently, at least in our experiments, it is possible to process automatically roughly two thirds of the participants, while leaving to the doctors only the remaining third (without accuracy or sensitivity losses compared to the doctors considering all participants).

The above approach is disadvantageous that it can be applied only to the unimodal approaches that, according to Table 3.4 in Chapter 3, have the lowest performance. Hence, it is necessary to define a confidence measure that is independent of the particular approach being used. One possibility is to consider the following:

$$c = \frac{\max(n_D, N - n_D)}{N} \tag{4.5}$$

where N is the total number of clauses a participant has uttered and n_D is the number of clauses that, for a given participant, have been assigned to class *depression*. The *rationale* behind the definition above is that the higher the fraction of clauses the approach assigns to a given class, the higher the confidence of the system.

The measure above allows one to rank the participants according to the value of c (from largest to smallest) and to consider the accuracy at position r. If higher values of c correspond to correct decisions, the accuracy should be high when considering only the top positions of the ranking. Figure 4.3 appears to confirm this expectation and, in particular, it shows that the multimodal approaches have an accuracy exceeding 90% when considering the top 40 ranking participants (roughly two thirds of the total) ,except for LR-TF that exceeds 88%. In this respect, the approach appears to be in condition to discriminate between cases that are sufficiently clear to be processed automatically and cases that require medical attention, thus allowing the system to potentially reduce by two thirds the workload of the medical personnel while still keeping the accuracy above 90%. This is important because it can increase the efficiency of screening services and, correspondingly, reduce the costs associated with depression diagnosis.

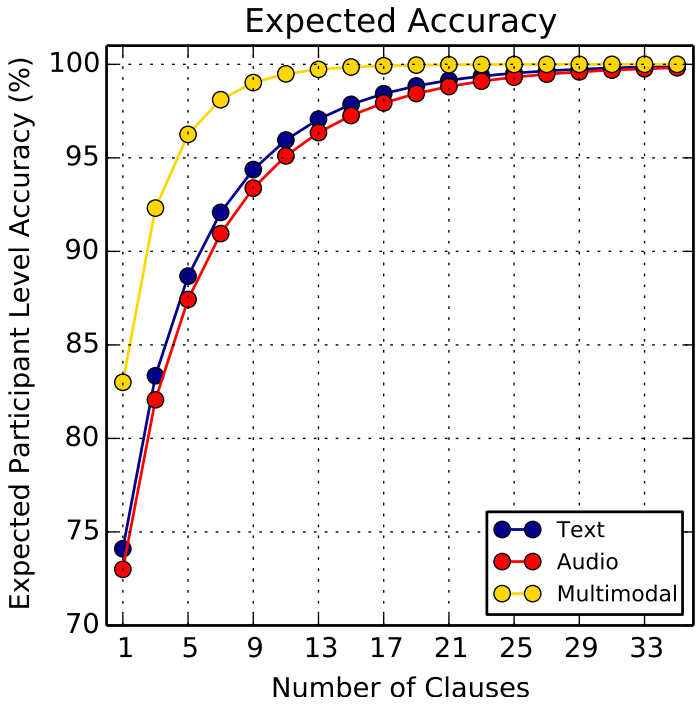

Figure 4.4: The plots show the expected accuracy of unimodal approaches and FF when using only a limited number of clauses. The expected accuracy is based on Equation (4.2) and it is based on the assumption that correctly classified clauses distribute uniformly across speakers.

According to the Gartner Group, one of the most important strategic consulting companies in the world, the detection of mental health issues is one of the most promising areas of Social and Emotion AI [5], the AI areas concerned with the inference of effective phenomena from observable data. The main reason is the increasingly greater number of people affected by mental health issues [339] and the resulting pressure on healthcare services. In such a context, our approaches can support the work of psychiatric and counselling services, possibly allowing doctors to concentrate on ambiguous and difficult cases, while leaving machines to tackle with the most evident ones. This agrees with all observations showing that the best way to implement AI is to use it for supporting humans and not for replacing them [83]

[5] www.gartner.com/smarterwithgartner/13-surprising-uses-for-emotion-ai-technology/

4.4 Experiment 3: The Analysis of Time: Based on Number of Clauses

Previous computing efforts, to our knowledge, have largely concentrated on improving detection efficiency and have addressed only to a limited extent, if at all, the problem of how much data is necessary to make a reliable decision about an individual (see Section 3.3 in Chapter 3). The experiments introduced in Section 4.2 shows that a limited accuracy at the clause level is sufficient to increase the probability of at least half of the clauses being classified correctly, the condition for a participant being assigned to the right class. This probability can be estimated using the equation 4.2 (by assuming that the clause level accuracy is the same for all participants). Figure 4.4 shows that such a probability increases significantly with the number of clauses and, therefore, the greater the number of these latter, the higher the expected participant level accuracy.

One of the main consequences of the considerations above is that it takes a substantial amount of time before the number of clauses is sufficiently large to ensure high performance. This is a problem for at least two reasons—the tendency of depressed people to speak less than the others (see Chapter 3, Section 3.2), and the need to shorten the interviews to lower the costs associated with depression diagnosis. Hence, this section investigates the relationship between performance and number of clauses. In particular, the analysis focuses on the two unimodal approaches and on FF-TF, the approach with the highest participant level recall. This experiment is presented in the publish work [6].

Figure 4.5 shows how accuracy, precision and recall change as a function of the number of clauses used to make a participant level decision. The reason for considering only odd numbers is that this makes it possible to apply the majority vote without the risk of a tie. For unimodal approaches, the plot shows that the accuracies of both audio and text unimodal approaches after one clause are within a statistical fluctuation regarding the accuracies obtained while using the whole corpus. However, there are statistically significant differences for precision and recall. For both modalities, after the first clause, the precision is lower, but the recall is higher. For what concerns FF-TF, the pattern is similar, with the recall that has a

[6] Aloshban, Nujud, Anna Esposito, and Alessandro Vinciarelli. "Detecting Depression in Less Than 10 Seconds: Impact of Speaking Time on Depression Detection Sensitivity." In Proceedings of the 2020 ACM International Conference on Multimodal Interaction, pp. 79-87. 2020

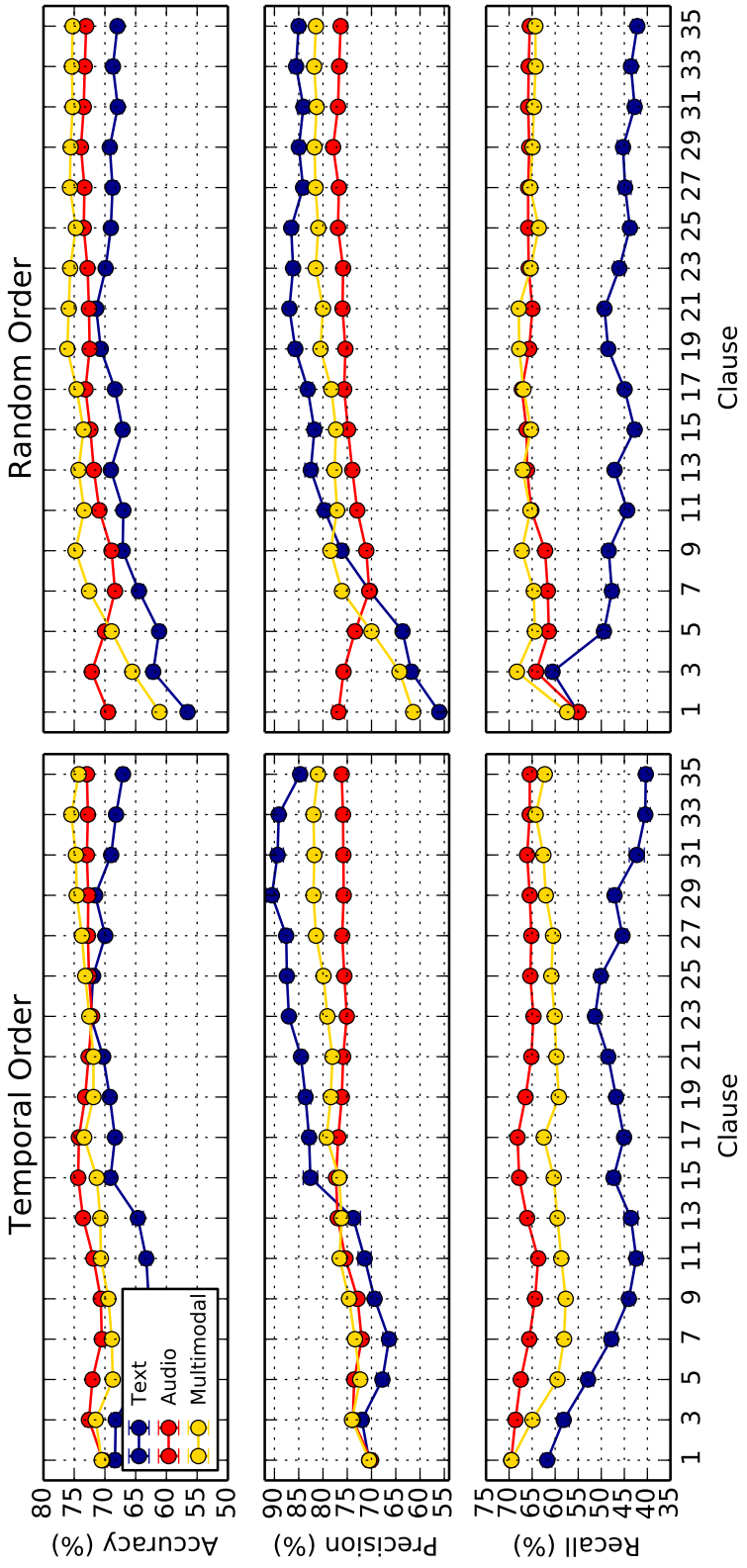

Figure 4.5: The plots show accuracy, precision and recall as a function of the number of clauses. The left column shows the results when the clauses are added in the same order as they appear in the interviews, while the right column shows the sam results when the clauses are added randomly.

small decrease (from 71.0% to 69.5%).

As the number of clauses increases, the pattern remains roughly the same for both unimodal and multimodal approaches. Therefore, the recall seems to improve or remain stable (for FF-TF) when considering a limited amount of material. In this respect, using a limited number of clauses appears to ensure that more depressed patients are recognised as such. Even if this comes at the cost of more control participants being classified as depressed, this result can be considered positive because the consequences of type II errors (classifying a depression patient as control) are significantly more negative than those of type I (control participants classified as depression patients).

The effectiveness of the approaches after the first few clauses, especially for recall, can induce the interpretation that the depressed patients tend to manifest their condition more clearly at the very beginning of the interview. Similarly, it can be argued that the results stem from the particular questions asked at the beginning of the interaction. Hence, the same experiment was conducted after shuffling the order of the clauses (see plots in the right column of Figure 4.5). It can be seen that the pattern is similar, suggesting that the clause order is irrelevant. Furthermore, it confirms that using a limited amount of material appears to induce the same recall level as when using the whole interview. Given that the average length of a clause is 1.2 seconds, the results above mean that such a time is sufficient to identify as many depressed patients as those that get detected when using the whole material at disposition. Alternatively, it is possible to perform depression detection with less than 10 seconds without significant performance losses, especially for recall. Furthermore, the results show that such a result can be observed whether the clauses are recognised in the order as they appear in the interviews or randomly. This suggests that the observed results do not depend on the protocol applied at the beginning of the interviews, but on the amount of data.

One possible explanation of the results above is that depression patients tend to manifest so consistently their condition, that there is high probability of correctly classifying any clause they utter. Not surprisingly, the clause level accuracy is well above chance for all approaches considered in the experiments. This result agrees with previous observations showing that limited amount of audio, possibly captured in naturalistic settings like the one in our experiments, is sufficient to perform depression detection, especially when the approach is based on paralanguage [140]. However, our results seem to contradict the finding in [5]

that depression detection can improve by considering when a given sentence is uttered during a conversation.

These observed results are significant since fast depression detection addresses several issues in clinical practice. The first is the tendency of depressed individuals to avoid social interactions and to speak less than non-depressed people [44, 118]. The possibility to detect depression with limited material canhelp handle such a tendency and to obtain good results for people that cannot sustain an interview like those used in this work. The second is to spot actually depressed people among the many individuals that call counselling services because they are momentaneously in distress but are unaffected by a pathology. In this respect, approaches like those presented in this work can help to quickly dispatch callers among operators more or less qualified to handle depressed individuals.

4.5 Conclusion

This chapter has presented several experiments to analyse comprehensively depression problem. It discusses several important contributions to the research community as they serve as great resources for understanding the effectiveness of the multimodal combination, alongside the differences between unimodal and multimodal aspects when designing a depression detection system. The experiments show that one of the main differences between unimodal and multimodal methodologies is that these latter tend to have more uniform clause level accuracy across the participants. This is significant because it contributes to higher person level accuracy. This is due to the tendency of certain participants, particularly those who are depressed, to express their condition either through what they say or through how they say it but not through both. Additionally, GMU weights were analysed, and it shows that higher weights are assigned to language regarding control participants.

This chapter also conducted experiments to show that it is possible to measure the 'confidence' of the approach and automatically identify a subset of the test data in which the performance is above a predefined threshold. This is important because it shows that the systems can reduce the workload of the doctors by up to two thirds while still ensuring a desired level of performance (above 90% accuracy). Additionally, this chapter investigated the performance as a function of the time to see how many materials the system needs to

detect depression. The experiments show that it is possible to perform depression detection with less than 10 seconds without significant performance losses, especially for recall. The results are based on the amount of data not on the procedure used at the beginning of the interviews. This finding may explain that depressed patients' states appear to be expressed consistently, making it easier to classify any clause they utter. The next chapter will present the misinformation problem.

Chapter 5

The Landscape of Misinformation: Misinformation Background

This chapter presents a background of misinformation problem. Section 5.1 differentiates fake news from a series of related terms. We also discuss in Section 5.2 the effect of misinformation in society and introduce different methods to address the problem of misinformation. Section 5.3 overviews language-based approach for distinguishing between fake and real claims. External evidence approach is also discussed in Section 5.4 to enrich the linguistic features of the claims by representing the central content of them more authentically. The importance of interpretability of verdicts is highlighted in Section 5.5, in which it potentially helps a reader in understanding the classification decision. Finally, different datasets are introduced for misinformation problem that we will used in our experiment in Chapter 6.

5.1 A Taxonomy of Misinformation

Supported by social media, misinformation or fake news has reached the public and caused more serious social damage. Misinformation detection has been studied extensively by both academic communities and the industry. Nevertheless, consensus about the definition of misinformation among many existing studies remains absent. Thus, we begin by discussing the definitions of misinformation that have been widely used in previous studies. Then we present our definition of misinformation, which will be adopted for the rest of this study. We use the term fake news or misinformation interchangeably in this study.

Generally, the term *'news'* indicates all types of claims, statements, speeches, posts via

social media or mainstream channels. Fake news was exclusively used in the satire context [26, 46, 268]. Balmas et al. [26] stated that fake news is intended to be seen as fictional, whereas conventional news is intended to be seen as rational. The study in [68] described fake news or misinformation broadly, stating that it includes everything from malicious articles to political propaganda. They discovered that many articles are published by journalists who rely on online searches for information but do not verify it. Also, 53.8% of journalists use microblogs, such as Twitter and Facebook, to collect facts and report on news stories [344]. According to a recent analysis of misinformation in the 2016 election [14], misinformation or fake news is described as any news that is deliberately and verifiably inaccurate and may deceive readers. This definition has been widely adopted in some of previous research [71, 155, 213, 243]. Following this definition, misinformation has been linked to several terms and concepts, including satire, parody, fabrication, manipulation, propaganda and advertising [362].

News satire is considered irony and humour than information delivering. News fact is exaggerated to attract audience to their posts or shows. However, it is harmful because it will reach different kinds of audience with different levels of cognitive abilities. This leads to inability to discover the satirical cues; therefore, this news may be believed easily [309]. *News parody* is similar to news satire, involving entertainment in news. The difference between satire and parody is that the latter relies on unreal information to inject sarcasm while the former relies on facts but presents it in a diverting format [309]. Also, *fabrication news* is unreal fact lacking factual basis but are published in the style of news article to create legitimacy. Unlike the previous ones, there is no implied understanding between the producers and receivers to know that the contents are fake. Its goal is to deliberately convey the misinformation to deceive audience in their own interests.

Different from textual information, *photo manipulation* is based on photos and videos rather than text to illustrate visual news. The difficulty in photo manipulation detection ranges from minor adjustments (i.e. change pictures' colour saturation) to complex adjustments (i.e. add or remove some items in photos). *Advertising and public relations* is another type that aims to entice the public to click on specific link to transfer them to the commercial marketing site, this is called clickbait headlines [309]. Finally, *propaganda* is related to political news that are originated by the political entities to affect the audience perspectives. This concept

has flourished during the election for a president in the United State 2016. Numerous fake news were propagated through social networking to deceive voters or gain a profit [309]. In the US election of a president 2016, it was alleged that fake news might have been pivotal in the election of President Trump [14].

The scope of this thesis is to detect the credibility of information on web based only on the textual information in the English language. We formally define misinformation as any claim or statement that intentionally and verifiably not credible. The term *'credibility'* instead of *'truth'* should be used since undisputable truth is frequently elusive and ill-defined. We focus on tackling the problem as a text classification problem, i.e. attempting to automatically detect whether a particular claim is fake/not credible or not. By *'fake / not credible'* it means unverified or untrue claims, or attempts to disseminate information that is not accurate.

5.2 The State of Misinformation

The great discovery of the World Wide Web has made distribution of information around the world easy. The increase growing of the web use has caused substantial increases in the use of internet. Recently, people are gradually turning to the internet for watching news rather than conventional news sources. A recent study shows that around 68% of adults in the USA watch news on the web rather than in TV [54]. Although internet is a fast source of useful information, there is a large growth in the spread of false information on the web [161, 294]. In [169], the velocity of misinformation was studied and it has shown that posts containing false information reach people on six times faster than truthful posts. Given the high spread of false information, words like 'Post-truth' and 'Fake news' are called as word of the year by Oxford dictionary in 2016 and by the American Dialect Society in the 2017, respectively. This high spread of false information on the internet has negatively affected the society in general, such as influencing stock market [4, 42], manipulating political decisions [26, 46], defamation of personalities [336] and creating bias to change real world event outcomes [104]. Studies on the false information consequences have also shown weakening in human memory rising after experiencing false information [185, 212]. Figure 5.2 exemplifies true news and its fabrication.

The societal challenges mentioned above have increased efforts to limit the distribution

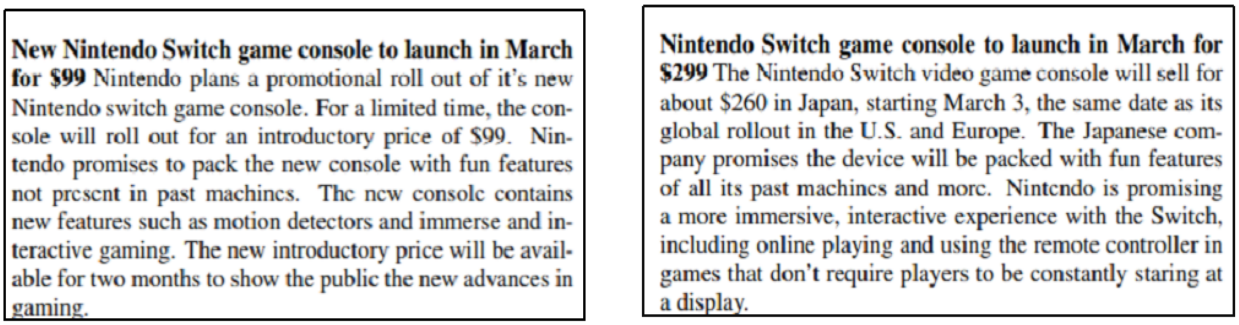

Figure 5.1: The Figure shows a sample of legitimate and fabricated information in the technology domain. The Figure (a) presents the fabricated information of the original article in the Figure (b). The sample is taken from [236].

of false information. Specifically, with this spread of misinformation, fact-checking and debunking websites have increased, including Snopes[1], PolitiFact[2] and FullFact[3]. *Fact-checking* is the process of verifying information to determine the claims veracity and correctness, where qualified journalists manually assess such controversial claims, analyse their credibility and provide analysis along with the supporting evidence (e.g. background articles, trustworthiness of the information source and quotations). In the fact-checking process the *claim* is defined as the statement that is being fact-checked, *verdict* is the conclusion on the veracity of the claim according to the fact-checkers, and *supporting or relevant evidence* is defined as the relevant documents or articles that are extracted from reliable sources to supports a claim.

The main advantage of fact-checking is that it can correct a person's misconceptions. A long-term study of the 2014 election showed that being exposed to fact-checking substantially enhanced the accuracy of people's beliefs [218]. Recently, a cooperation between Facebook and fact-checkers resulted in fact-checked articles valued false to spread to 80% fewer people; though, further data about the extent and the relative coverage of fact-checks has not been shared to public and with researchers [18]. Nonetheless, the study in [15] showed that sharing wrong information happened 60% less on Facebook than on Twitter due to Facebook's implementation of high proactive stance, which limits the spreading of false information on their platform. Moreover, after the increases of fact-checking websites, political candidates have been more aware of fact-checking concept. When government representatives believed

[1] www.snopes.com
[2] www.politifact.com
[3] www.fullfact.org

that their speech were being fact-checked, they decrease the number of imprecise statements they announce [217]. The fact-checking of political statements are used to certify politicians' statements and devoting financial and personnel resources.

Despite the benefits of fact-checking process, it has some drawbacks. First, fact-checking process may induce into counterproductive in which fact-checking an article could produce familiarity with it, and this familiarity produces acceptance rather than rejection [37, 97, 234, 308]. Lewandowsky et al. [308] suggested that debunk rumour should not be broadcasted, and the credible facts should rather be reported without revealing the false information. This is because changing people's opinions of what is true is enormously difficult [116]. Second, the same as in education sector, the manual fact-checking process is an individual's responsibility. Lazer et al. [171] indicated that people would probably not fact-check an article that aligns with their previous beliefs. In [177, 195, 195], they demonstrated that fact-checkers frequently do not allocate the same truth scores to the same claim, and the scores rarely overlap between fact-checkers. This is especially true when a fact-checker has a bias towards a subject or an organisation. Third, multiple researches in social psychology and communications showed that humans can identify deception slightly better than chance. The accuracy rates ranged in 55%-58%, with a mean accuracy of 54% over 1,000 participants in over 100 studies [315]. Finally, manual fact-checking may not scale well due to the high volume of new information on the web. Therefore, it is intellectually challenging and time-consuming [358] and based on the difficulty of the claim, the process may last from hours to few days [129].

Based on the above shortcomings, objectively fact-checking a claim is required by automating the manual assessment process. The objective of automated credibility assessment is to decrease the burden by supporting human in validating the veracity of the information. By considering the cruelty of the problem, the main advantage of performing automatic claim verification is that it can be conducted on a large scale. For example, the experiment in [66] converted Wikipedia into a network of knowledge graphs therefore, unverified claims can be checked against this network. A claim known to be true in Wikipedia will appear as an edge of the knowledge graph or will have its subject and object linked via a short path in the graph. Otherwise, the false claims will not appear as a connection in the graph. Jaradat et al. [143] also proposed ClaimRank, which is a computational framework that distinguishes claims that may require checking. The claims that need to be checked can then be submit-

ted to fact-checking websites for manual verification or to the automatic systems. The study in [207] proposed a framework that searches for relevant documents to a given claim and snippets of evidence. Although using the system may not be fully automated, it can support human verification experts.

5.3 Language-Based Text Analytics

People seem to harness their cognitive efforts to alter or conceal information, which induces changes in behaviour and, thus, changes in verbal and written texts. This induces linguistic feature changes, and by examining these features, fake texts may be discovered. In this context, writing a false information tends to be a matter of carefully choosing words because words are the richest and most distinctive form of communication. Despite regulating what they are writing, language 'leakage' happens in such linguistic aspects that are involuntary behaviour and difficult to control, such as frequencies and patterns of pronoun, conjunction and negative emotion word usage [107]. This challenge inspires researchers to investigate various methods for detecting deceptive texts [253].

Detecting misinformation based on its text content is an intuitive and straightforward approach adopted by many existing studies. The experiment in [267] showed that the total word count in fake texts is greater than the legitimate texts. Besides, self-oriented pronouns in fake texts are used less often than other-oriented pronouns, and sensory-based words are used more frequently. Moreover, several researchers have attempted to identify linguistic features, such as semantic features (e.g. category, entities, keywords), sentiments features (e.g. subjectivity) and syntactic features (e.g. part-of-speech tag, punctuation marks, spelling errors) [52,56,242,316]. Similarly, those who write or spread misinformation need to capture the attention of readers thus the text style becomes distinct. Style can be seen as a set of language features that includes lexical choice, syntactic complexity, organisation and flow of information. Some of these features such as lexical choice, are easier to capture with computers than the others [167]. The study in [239] noticed that the article's language style is critical in assessing its credibility. Therefore, they use language stylistic features, such as assertive verbs, factive verbs and implicatives to evaluate the credibility of web claims.

To predict, however, the linguistic methods first need enough text content. Hence, they

cannot detect fake claims with either short or no text content. Textual claim is also diverse regarding topic, style and platform. Thus, content-based features that work well on one dataset of fake news might not work well on another [295].Several studies have used additional cues to enrich the linguistic cues, such as temporal spreading patterns [184], network structures [293,326,346] and users' feedbacks [292,327,328]. However, limited studies have utilised external webpages that can provide complementary context to the claim.

A recent trend of researchers is to develop more objective tasks and evidence-based verification solutions, which concentrate on using evidence collected from more trustworthy sources, such as encyclopaedia articles [311]. The study in [108] utilised news headlines as evidence to predict whether a claim would be supported, debunked or dismissed. In the Fake News Challenge, the body's article is used as evidence to reveal the stances regarding the claim presented in the headline. The experiments in [312] formulated the Fact Extraction and VERification (FEVER) task to verify the claim by gathering evidence from Wikipedia and synthesising information from multiple documents. Other study introduced an evidence-aware neural attention mode named DeClarE, which extracts salient words from related articles as the main evidence to verify a claim [241].

5.4 External Evidence

In the present era of big data, the most challenging issue for automated assessment is to gather related and enough evidence to confirm facts. A fact is something that can be proved with evidence to be true. Hence, collecting evidence is a crucial step in evaluating the veracity of any claim or information. Automated assessment of such claims requires machines to automatically collect the related evidence; thus, this automated system will be confined to the online evidence's repository in a machine-readable format.

A massive quantity of textual data is posted online every moment, and the majority of this data is unstructured. Approaches such as knowledge base depositories (e.g. YAGO [307], WikiData [330]), information extraction and semantic web transform the unstructured text into a machine-readable design. Such repositories are out dated and their coverage is somehow inadequate. Therefore, they can be considered useless in presenting evidence to validate contested facts, especially those emanating from current world affairs.

An alternative approach for gaining evidence online is by search engines. An automated method for credibility assessment can use these search engines and perform an online search to obtain the related evidence, although search engines usually list online pages that are related to the textual search query. These related online pages have multiple designs with multiple structure, including news stories, a group of question answers or an online discussion panels. Extracting relevant evidence from this disorganised clutter is a major challenge for automated credibility assessment.

Although online web has a massive resource of information, not all the knowledge posted there is credible, and not all the web sources are trustworthy. The trustworthiness of the data sources immediately impacts the credibility of the information [109]. For example, a fact stated in The New York Times[4] is probable to be credible—strictly examined by the expert journalists. However, some information from The Onion [5] is most likely unreliable because it is a satirical news institution. Therefore, assessing the trustworthiness of the data sources is very critical for evaluating the credibility of the information. The most common conventional methods for assessing the quality of online sources are PageRank [47] and authority-hub analysis [156], which depend on the format of the hyperlink of the Web- graph, though such methods detect only the authority and reputation of the online sources and not their trustworthiness from the information credibility standpoint. For example, 'The Onion' website has a high page rank mark, which is 7 out of 10.

Recently, the study in [240] suppressed the previously mentioned limitations for evaluating the quality of the retrieved evidence. They retrieved diverse evidence articles about the claim and assessed them based on identifying the language of the reporting articles (i.e. bias and subjectivity), the articles' stance towards the claim (i.e. whether it supports or refutes the claim) and the reliability of the web sources generating the articles. In details, number of articles were retrieved by firing the textual claim as a query, and they assessed the most related articles based on different factors mentioned above. First, they assumed that the reporting articles should be reported in an objective and unbiased language to be considered credible. Therefore, language stylistic features of the retrieved articles were analysed, including assertive and factive verbs (e.g. 'claim', 'indicate'), hedges (e.g. 'may') and reporting verbs

[4] www.nytimes.com
[5] www.theonion.com

(e.g. 'deny'). Second, the stance of the articles is captured by their proposed determination classifier to assess whether the articles reporting the claim are supporting it or not. For example, an article from a reliable source such as 'truthorfiction.com' refuting the claim will make the claim less credible. The detailed method for stance determination is outlined in Algorithm 1 in [240]. Finally, the reliability of the web-source hosting the article significantly impacts the credibility of the claim. Unlike PageRank and authority-hub, a web source is considered reliable if it contains articles that refute false claims and support true claims. In this study, extracting and evaluating relative evidence are beyond the scope; thus, we follow the same approach in [240] for retrieving evidence.

5.5 Interpretable Machine Learning

In the age of artificial intelligence, machine learning models have started to be the first option for resolving serious issues in different fields such as in finance, healthcare and justice. Because of this prominent importance, it is vital to understand how and why these models create certain decisions. Interpretability is the degree to which a person can realise the reason of the decision [203]. The majority of existing machine learning models are not interpretable because they do not explain their decisions. Overall, interpretability can assist in identifying hidden biases in machine learning models. This problem has concerned the research community [180, 345]. A comprehensive discussion on the motivation of interpretability and multiple methods to achieve it can be found in [182].

Multiple traditional machine learning models, such as regression, Naive Bayes, decision tree and random forest are naturally interpretable. The coefficient weights in regression, for example, indicate the significance of the features. Likewise, the traditional feature selection approaches also assist in clarifying model decisions by highlighting the importance and contribution of each feature [127, 356]. Studies in [164, 165, 175, 333] suggested methods for creating decision lists that enhance interpretability over decision trees.

The study in [263] represented a method for clarifying predictions of black-box models for individual instances. Likewise, the study in [22] proposed a different method for describing the decision of arbitrary nonlinear classifiers. Another study in [259] also proposed another method to demonstrate the predictions of any classifier by estimating the black box

model locally around the prediction. Furthermore, Samek et al. [276] suggested procedures to clarify predictions of deep learning models.

In contrast, few approaches have concentrated on automatically gathering evidence that supports factual claims raised in a debate. Supervised learning approaches to achieve this for claims on social media and debate platforms are illustrated in [2,262]. Likewise, the studies in [2, 33] proposed methods that tackle this issue from an information retrieval standpoint using varying documents to retrieve evidence in support of an assurance claims. The study in [240] pursued a similar direction to the evidence retrieval approaches. Having a claim, their models thoughtfully pull snippets from the relevant articles and that will assist in understanding their automated valuation (see Section 5.4 for more details).

5.6 Exiting Datasets

There are several datasets available online for misinformation problem, especially for social media domain. However, limited number of datasets are intended to an open-domain setting. In this section, we will examine three most common datasets that are available for open-domain setting and are used in the literature. These datasets are used in our experiments in Chapter 6.

5.6.1 Snopes-A Dataset

Snopes is a fact-checking and debunking website that validates rumours, hoaxes, urban legends, e-mail forwards, and other stories of unknown or questionable origin. 300,000 visits a day usually call the website. The false information is typically collected by users from Facebook, Twitter, Reddit, news websites, e-mails and from any platforms. The credibility of a claim is manually verified by performing a contextual analysis. The verdict of a claim is true or false and sometimes can be mostly true or mostly false. This verdict is followed by a summary of how the editor(s) found the claim (e.g. it was gathered from social media or obtained from a user's email), a section explaining the origin of the claim, and a section of analysis supporting the verdict.

The Snopes dataset was created by Popat et el. [240], which contains fact-checking claims from Snopes published until February 2016 (referred to as Snopes-A dataset). Each claim was

used to query the Google search engine, and several relevant articles to a claim were retrieved. Relevant articles refer to the articles that support the claim based on several factors mention in Section 5.4. Ultimately, any search results coming from the Snopes domain are discarded to avoid any kind of bias. At the end, the dataset includes the claims, labels, relevant articles and articles' sources. Each claim has more than one supporting articles that may be retrieved from different web sources. The dataset is publicly available to download [6].

5.6.2 PolitiFact Dataset

PolitiFact is a political fact-checking website in which editors rank the credibility of claims that were created by elected officials, candidates, leaders of political parties and political activists in US politic. The PolitiFact's editors validate the claims by pundits, talk show hosts, columnists and widely distributed chain e-mails. The credibility verdict of the claims can be one of six possible ratings: true, mostly true, half true, mostly false, false and pants-on-fire. Along with the verdict, recorded interviews and a list of sources that support their verdict are provided with each claim. This is important for PolitiFact transparency, and it helps readers to judge and convince with the verdict .

PolitiFact dataset was created by Popat et al. [240] in which the claims that were published before December 2017 were extracted. For collecting related evidence, each claim was used to query the Google search engine, and several relevant articles to a claim were retrieved. Similar to Snopes dataset (Snopes-A dataset), relevant articles refer to the articles that support the claim based on several factors mention in Section 5.4. Importantly, any search results coming from the PoltiFact domain are discarded to avoid any kind of bias. At the end, the dataset includes the claims, labels, claims' sources, relevant articles and articles' sources. Each claim has more than one supporting articles that may be retrieved from different web sources. The dataset is publicly available to download [6].

5.6.3 Snopes-B Dataset

Snopes-B dataset is used by [146]. It is originally part of the misinformation dataset for micro-blogging that was developed by Ma et al. [190]. They extracted 992 events and for each event a set of tweets was retrieved from Twitter. Also, 778 of these events were generated

[6] https://www.mpi-inf.mpg.de/dl-cred-analysis/.

from Snopes website during March-December 2015 of which 64% were fake claims. For making the two classes balanced, they further added 214 non-rumour events taken from [53, 163]. The resulting dataset contained 498 rumors and 494 non-rumours. The dataset has various domains including politics, local news, and fun facts. Each claim (or event) is labelled as factually true or false.

Ma et al. [190] dataset did not release the claims or events as a part of their dataset; however, it includes only the set of tweets for each claim. Karadzhov et al. [146] managed to find the original claims for only 761 from snopes.com-based clusters. Their study focused on open-domain setting, thus all the tweets were ignored, and they only used the claims. For extracting evidence articles, their method differed from [240] in which they generated a short query to the Google search API and instead of querying with the full claim, verbs, nouns and adjectives were considered to get high quality search results. Then, the web documents were retrieved, ignoring any results that come from unreliable sources that exist in their database. The web documents were split into three groups and the most similar group to the claim was taken by calculating the cosine similarity. At the end, only one supporting article was related to a target claim (for more details see Chapter 6, Section 6.2.1).

Karadzhov et al. [146] made the dataset (Snopes-B dataset) available with their code[7]. The number of the claims in the dataset exceeds 761 claims (4,856 claims). This is different from what they state in their paper; however, it turns out that they further added more claims from other Snopes dataset in [240]. The dataset includes claim id, label and the claim itself without including any external supporting articles. Therefore, we used the same technique as they proposed to retrieve the article for each claim.

5.7 Conclusion

This chapter introduced misinformation or fake news problem. It presented different definitions of misinformation, which have been widely used in previous studies and the effect of misinformation on society. We highlighted the importance of automatic assessment approach over the manual fact-checking approach for claims assessment. We focused on tackling the problem as a text classification problem, i.e. attempting to automatically detect whether a

[7] github.com/gkaradzhov/FactcheckingRANLP

particular claim is fake/not credible or not. By 'fake/not credible', we meant unverified or untrue claims, or attempts to disseminate information that is inaccurate. Also, the chapter overviewed language-based approach and discussed it corresponding limitations for assessing the credibility of information. Then, the significance of utilising external evidence to enrich the textual data with its context was highlighted and providing it to the users to support the automatic verdict of the system. Finally, we introduced three datasets: Snopes-A, PolitiFact and Snopes-B as resources for text classification efforts that provide external articles with each claim. In next chapter, we present our approach for credibility assessment with a comprehensive analysis for this problem.

Chapter 6

Automatic Assessment Based On Evidence-Aware for Claims Credibility

6.1 Motivation

Fake news or misinformation is intentionally written to mislead readers, which makes it nontrivial to detect them simply based on the text of the claim statement. Also, the structure or origin of the claim is relatively short and contains very limited context. Therefore, the linguistic approach only may not capture the credibility of the claims with either short or no text content. Several researchers have suggested different features to distinguish claims [133,239,254]. Such features can be extracted from the claim text itself [254], author′s history [269] or even from the web considering it as a knowledge source [146, 241]. Despite the availability of various reporting articles related to the claim in the web, few studies have considered it as knowledge source. It considers external evidence in the form of other articles (retrieved from the Web) that confirm or refute a claim. Moreover, it can provide complementary information and gives an interpretative explanation for the user. This agrees with the process of manual fact-checking that entails various evidence from different sources to help in understanding the context of the statements and to support their verdict [45].

The experiments in [146, 241] have utilised a search engine to retrieve external evidence/articles and used them to assess the credibility of the textual claim. However, these articles are long hence carrying the semantics along all-time steps of recurrent models is hard and not necessary. Also, these approaches suffer due to some reasons in [241], the approach is very complex due to the nature in which article sources and claim sources are embedded

and could have influenced the results, as many sources of fake news have a common origin [15]; in [146] four types of similarities are computed and fed through LSTMs. These models might perform poorly in long sentences because of discarded internal interaction between the words of the article [60]. However, Chen et al. [59] incorporated a soft-attention mechanism into the RNNs to pool out distinct temporal-linguistic features with a particular focus. Yet, this approach relies on domain-specific and community-specific features and ignores external evidence, which provides limited context for credibility analysis.

To overcome the limitations of the prior works, this chapter proposes an automated credibility assessment approach. It aims to decrease the burden by supporting humans in validating information based on joint interactions between a claim and its several supporting articles. The idea is motivated by the fact that textual claims are relatively short and could not be reliably used for classification. In contrast, evidence articles can be used to represent the central content of the claim more authentically. For the article input, LSTM neural network was applied and since the article input is long by its nature, the output of the last step of LSTM may not represent the entirely of the article's semantics. Moreover, concatenating all vector representations of multiple words in the article may induce a large vector dimension. Thus, the internal interaction between the words of the article may not be considered. For these reasons, we adapted self-attention mechanism that applies on the top of the LSTM model which can extract different aspects of the article into multiple vector representations. The system then aggregates all the information about the web articles to their target claim by applying a majority vote to asses the claim. Finally, the retrieved article can be visualized to the user as an interpretable explanation. This chapter presents the published work in ACM conference [1].

The research questions and subsequent novel contributions of this work are the following:

RQ-1: Does enriching a textual claim with its supporting articles improve the credibility classification of this claim? (see Section 6.4.5)

RQ-2: How does the system perform compared to the state-of-the-art models? (see Section 6.4.5)

RQ-3: What is the impact of article length on the system performance? (see Section 6.4.6)

RQ-4: Are the results considered satisfactory regarding possible fake news applications? (see

[1] Aloshban, N., 2020, July. ACT: Automatic Fake News Classification Through Self-Attention. In 12th ACM Conference on web Science (pp. 115-124).

Section 6.4.7)

RQ-5: Is the system performance affected by number of evidence articles for a particular claim? (see Section 6.4.7)

RQ-6: How can the system help to interpret the classification results? (see Section 6.4.8)

This chapter first overviews the previous works related to the problem. Then, the proposed approach is described. Finally, we report the experiments and discuss the results.

6.2 Related Work

The research areas most closely related to our work are fake news detection and attention mechanism. This section overviews these two topics.

6.2.1 Fake New Detection

Previous works have typically relied on textual, network and temporal features to characterise and detect fake news. Wang et al. [336] highlighted the importance of social context along with textual features on social media domain. Thus, additional metadata information was used to improve the performance which are subject, speaker profile, party affiliation, and different media sources. Ma et al. [191] and Ruchansky et al. [269] have also adopted RNNs to represent sequential posts and user engagement. Zhang et al. [359] also used RNNs to detect fake news by exploring news creators, articles, subjects ,and their relationships. Despite the performance improvement gained in fake news detection on the social media domain, these models may not effectively execute in detecting fake news in its early stage. Such metadata features are not always available as a prior for fake news detection and user responses to the event comes after this event has been intensely propagated.

Apart from the social media domain, several researchers have investigated fake news detection on the web. Different linguistic features, either content-based features or semantic features, are used for the classification [244, 254]. It aims to find specific writing styles and sensational titles that commonly occur in fake news contents. The study in [240] proposed a pipeline of supervised classifiers that consider the articles' stance, the language style using subjectivity lexicons, the trustworthiness of the sources, and the credibility of the claim. Due to the dependency of the linguistic features on specific topics and the tight coupling to

domain knowledge, hand-crafted linguistic patterns are not yet well suitable for fake news classification, hence, not scalable [269]. To overcome this limitation, several deep learning models have recently been applied to extract the latent embedding features and the experimental results show the potential of these models [146, 241, 269].

Evidence-aware approach shows its importance in providing cues about the controversial nature of the claim and helps in understanding its credibility. Yet, very limited studies are conducted on verifying the credibility of the claims based on the web evidence. For example, Hassan et al. [130] used repositories, which act as a text-based evidence for the claims that are earlier fact-checked. New claims are matched against the claims in the repository, which was built from different fact-checking tools. However, this leads to limit in using this tool to detect the claims that are similar to what already exist in the repository. To overcome this restriction, the authors in [146] used the web as a knowledge source to confirm or reject a claim by generating a query to search the web. Four types of similarities were computed and fed through four LSTMs networks. This approach is brute-forced as there is no explanation for using four different sources of similarity and the complexity of the vectors within different LSTM components could induce information overload, which can affect the performance of the model. The method proposed in [241] also considered external evidence as manually retrieved articles and evaluated them against the language stylistic features, reliability of the web sources and the stance of articles toward the claim. Claim text, claim source, retrieved articles and article sources were fed to Bi-LSTM. However, as argued by [15], a number of sources are well-known to engineer false documents. Using these sources as evidence may boost the performance of the model, as such sources provide good separability (Figure 2 (b) of [241]).

6.2.2 Attention Mechanism

The attention mechanism has been used widely in recent literature. Dos Santos et al. [93] used attention mechanism on the top of CNN and LSTM models to guide the extraction of sentence embedding. The study in [60] proposed an intra-sentence level attention mechanism. This is a fine-grained process for detecting lexical correlations between nearby words. They have applied this model for sentiment analysis and have shown improvement in the performance. In [176],proposed a self-attention mechanism for question encoding. Sim-

ilarly, [178] proposed interpretative sentence embedding through self-attention and used a matrix representation for sentence encoding, which we also follow in our work.

CNN is widely applied in fake news detection given their successes in various text classification tasks. For example, [336] merged the max-pooled text embedding with the metadata representation from Bi-LSTM. However, representing the input embedding by using max-pooling could lose many valuable word meanings. Additionally, Popat et al. [241] used weighted average of the hidden states to represent the sentence encoding into a vector. They calculated the attention weights of the retrieved articles according to their relevance to the claim. Based on these attention weights, the retrieved article can be visualised to the user as an interpretable explanation. However, our approach uses a self-attention mechanism to calculate the attention weights according to the article itself. Through this, we extract the important features of the article into multiple vector representations. Our proposed model can not only automatically learn multiple feature representations, but also generates user comprehensible explanations of the learned representations.

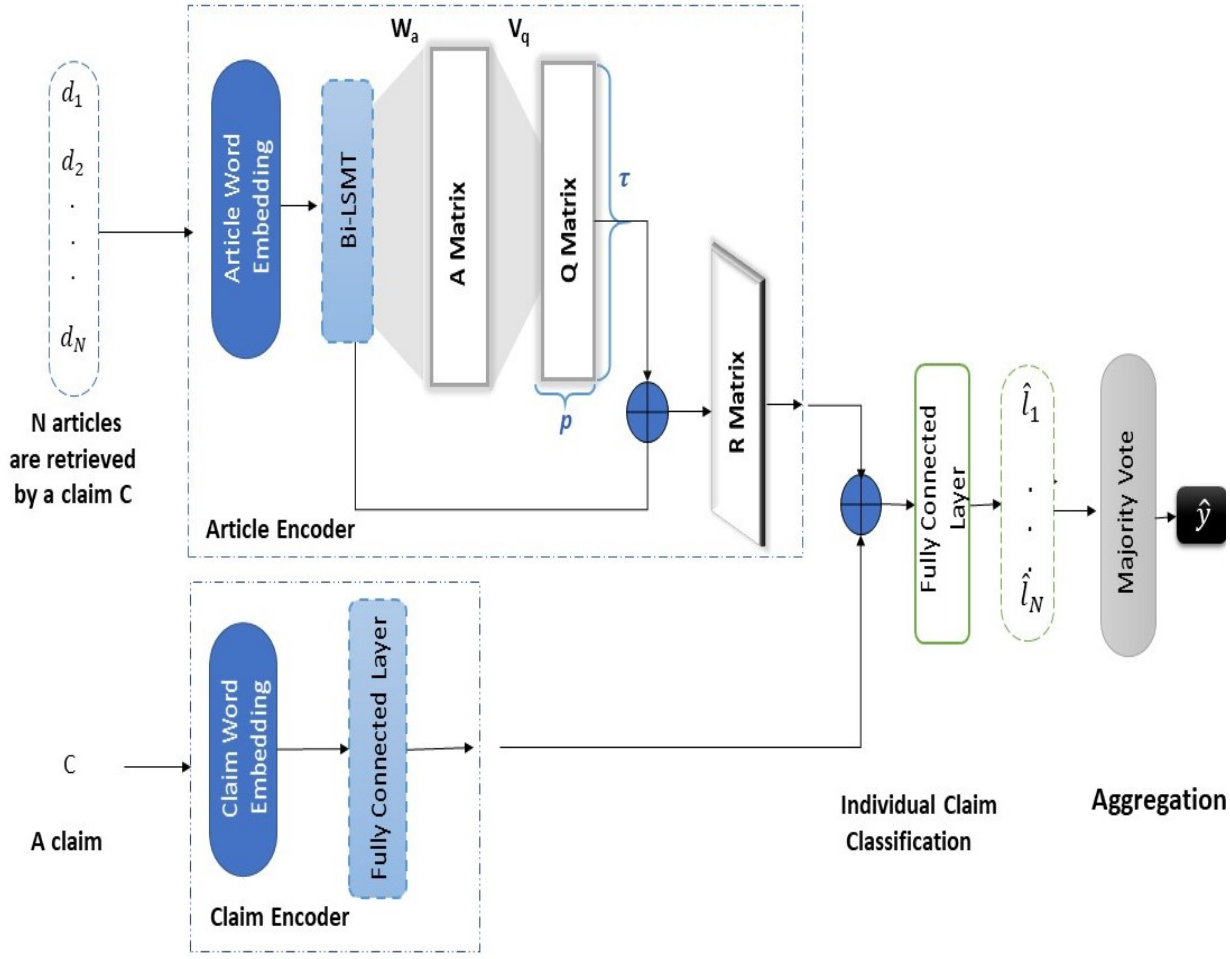

Figure 6.1: The Figure shows the proposed approach which takes as inputs the claim and N articles reported the claim, classes each of them as credible or not. This is performed by concatenating two sequences of feature vectors of the claim and the article and feeding it to a fully connected layer with a softmax activation. Then, we apply majority vote to decide whether the claim belongs to one class or the other.

6.3 The Approach

Individual claims are generally short and containing very limited context. For example, the false claim 'President Obama waived work requirement welfare' lacks of linguistic features needed for a good classifier and hence such short claims may not give a good prediction to the model. Therefore, we are not considering the claim text alone but also the various evidence articles related to it . For example, the reporting article,

'Romney campaign began airing its new TV ad, the ad praises the bipartisan cooperation of President Bill Clinton and a congress to overhaul welfare it then turns partisan and attacks President Obama. Romney TV ad right choice President Obama quietly announced a plan to gut welfare reform by dropping work requirements. Under Obama's plan you would not have to work and would not have to train for a job they just send you your welfare check and goes back to being plain old welfare. Mitt Romney will restore the work requirement because it.....' provides more context and hence we can extract inter-term relations.

Figure 6.1 shows the three major steps of the approach, namely *Encoders*, *Individual Claim classification*, and *Aggregation*. Initially, we treat the claim as the underlying information needed to be credibility checked. Assuming a claim C is reported by a set of N articles $D = \{d_1, d_2, ..., d_N\}$. The tuple of a claim and reporting articles $\{C, d_1\}, \{C, d_2\}, ...\{C, d_N\}$, forms an individual classification outcome $\hat{l} = \{\hat{l_1}, \hat{l_2}, ..., \hat{l_N}\}$. By concatenating them we hope to learn a new mapping that provides better estimation for a claim. Finally, the aggregation step applies a majority vote over $\hat{l}$ to obtain a classification outcome $\hat{y}$ for the claim C. The rest of this section explains each step in detail.

6.3.1 Encoder

In the encoder process, a claim and an article are converted first into sequences of vectors through word embedding step. Then, the claim encoder and article encoder are performed on these vectors.

Word Embedding

The claim and the article are converted into sequences of vectors through word embedding, an approach that has been shown to capture linguistic and semantic characteristics of words,

meaning that words similar along such dimensions tend to be mapped into similar vectors (see Chapter 3, Section 3.5.1 for more information about static word embedding). Particularly, each article is mapped into a sequence $X = (x_1, x_2, ..., x_\tau)$, where x_i is a D-dimensional vector corresponding to the i^{th} word and τ is the maximum number of vectors allowed in X. Moreover, the claim itself is also mapped into a sequence $B = (b_1, b_2, ..., b_k)$, where b_i is a Q-dimensional vector corresponding to the i^{th} word and k is the maximum number of vectors allowed in B. The values of τ and k have been set through cross-validation during the experiments. Consequently, X is represented as a two-dimensional matrix $A \in R^{\tau \times D}$ and B is represented as a two-dimensional matrix $B \in R^{k \times Q}$.

Claim Encoder

After representing a claim input as vectors, the vector representations of the claim $B = (b_1, b_2, ..., b_k)$ are passed through a dense layer to obtain the latent features of the textual claim.

$$S = W_s B + b_s, \tag{6.1}$$

where W_s and b_s are the corresponding weight matrix and bias term.

Article Encoder

- **Bi-LSTM**

  After representing an article input as vectors, the vector representations of the article $X = (x_1, x_2, ..., x_\tau)$ are fed to Bi-LSTM. The main motivation is that the article data is long and it is essential to encode the possible relationships between words in the article. Especially, it understands the context of the article in both directions of a word (see Appendix A, Section 17 for more details). Such a model produces the creation of the network's internal hidden state h_t with the input data x_t and the hidden state of the previous time steps h_{t-1} as follows:

$$\begin{aligned} \overrightarrow{\mathrm{h_t}} &= \overrightarrow{\mathrm{LSTM}}(x_t, \overrightarrow{\mathrm{h}}_{t-1}) \\ \overleftarrow{\mathrm{h_t}} &= \overleftarrow{\mathrm{LSTM}}(x_t, \overleftarrow{\mathrm{h}}_{t-1}) \end{aligned} \tag{6.2}$$

Where $\overrightarrow{h_t}$ denotes to the hidden state for time step t of the forward LSTM and $\overleftarrow{h_t}$ is the hidden state for time step t of the backward LSTM. The final output is produced by concatenated $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ to obtain $h_t = [\overrightarrow{h_t} \oplus \overleftarrow{h_t}]$. After concatenating the output vectors of each time step for τ vectors, a matrix H is generated with the shape of $[\tau, 2s]$, where s denotes to the hidden dimension of the LSTM network.

$$H = (h_1, h_2, ..., h_\tau) \tag{6.3}$$

- **Self-Attention Based on the Article Encoder**

  Attention mechanisms have shown successes in various fields, for example, from question answering and machine translations to image captioning [25, 60, 63, 134, 176, 178, 349]. Motivated by these, we conjecture that these attention techniques can improve the fake news/misinformation detection, especially given the presence of distinctive linguistic features for fake claims.

  Self-attention mechanisms are motivated to mimic the behaviours of human readers, who process text sequentially, from left to right, fixating nearly every word while they read and creating partial representations of sentence prefixes [157]. This can be achieved by utilising a self-attentive mechanism to overall hidden states H of a Bi-LSTM encoder. First, we need to obtain the attention weights q for τ values which is computed as follows :

$$A = \tanh(W_a H^T) \tag{6.4}$$

$$q = softmax(v_q A), \tag{6.5}$$

  where W_a is a trained weight matrix with a shape of $[e, 2s]$ and v_q, a trained parameter vector with size e, where e is a hyperparameter. This results in size $[e, \tau]$ for matrix A and the weighted vector q has a τ dimensional vector over which we apply a softmax to obtain a probability distribution over τ values.

  To obtain attended context vectors r, the weighted sum of each q_i with each value in the vector h_i is calculated.

$$r = q H, \tag{6.6}$$

where r has $2s$ dimensional embedding.

Typically, when reading a long article, we can pay different attention to it and find different parts of the article that express the semantic of the whole article. Therefore, instead of focusing on a specific part of the article by using only a single vector representation r, we need multiple vector representations that focus on p various parts of the article. This will capture the semantics of the article more broadly. To achieve our goal, the vector parameter v_q is replaced by a weight matrix V_q with a shape of $[p,e]$. In details, a linear transformation is performed to convert $2s$ dimensional space, which is produced after applying Bi-LSTM layer (Eq.6.3) to e dimension (Eq.6.4). Then, another linear transformation is performed to convert e dimension to p dimension, which results in p attention weight vectors of size τ.

$$Q = softmax(V_q\, A) \tag{6.7}$$

The final article representation R is computed by multiplying the matrix Q with the hidden states H, which result in p different weighed vectors of size $2s$.

$$R = Q\, H \tag{6.8}$$

- **Attention Diversity**

  Using p different attention weights can enhance the representation of the article with different semantics. Yet, the attention techniques may always produce similar summation weights Q for the article, thus, there will be no difference between multiple rows of attention weights. This yields the article representation R to become infected with redundancy problems. For preventing several attention vectors from being similar or redundant, a penalisation term is introduced to get diverse summation weight vectors in different p. Therefore, different summation embeddings in R can capture several aspects of the article. This can be achieved by ensuring that Q has orthonormal rows. Following Lin et al. [178] in regulating the redundancy, the dot product of Q with its transpose, subtracted by an identity matrix I is applied.

$$J = ||(Q\, Q^T I)||_F^2 \tag{6.9}$$

where $||\cdot||_F$ stands for the Frobenius norm of a matrix, and J will be large when the row values are similar and vice versa. This means that when p rows have apparent variances, J becomes smaller. Therefore, the objective is to minimise both penalty term J and the loss of the model together (see the following section).

6.3.2 Individual Claim Classification

To perform credibility classification for a claim, which is based on linguistic features of the claim and its evidence articles, the final representations, S and R are concatenated. A fully connected layer is then processed with a linear activation to map output of LSTM layer to a desired output size that produces a label prediction $\hat{l}$ for the claim.

$$\hat{l} = g(W_l(S \oplus R) + b_l) \tag{6.10}$$

Where $\oplus$ denotes the concatenate operation, W_l is a weight matrix, b_l is a bias term and g is a sigmoid that gives as output a vector $\hat{l}$ where $\hat{l}_i$ is the probability of the claim to belong to class i.

We measure the probability error in our model by the following loss function.

$$\mathscr{L}(\mathscr{X}) = -\frac{1}{M}\sum_{m=1}^{M}[l_m \log \sigma(\hat{l_m}) + (1-l_m)\ \log(l-\sigma(\hat{l_m}))] + \lambda J, \tag{6.11}$$

where $\mathscr{X}$ is the training set, M is the total number of samples in $\mathscr{X}$, l_m is the groundtruth of training sample m, λ is a regularization hyperparameter and $\hat{l}_m$ is the classification outcome for the same sample. By this constraint, the parameters in our model are trained by back-propagation. We also use the gradient clipping technique to alleviate the exploding gradient problem with a threshold of 0.5.

6.3.3 Aggregation

Section 6.3 shows that every claim has multiple supporting articles. The claim classification step (section 6.3.2) processes each of them independently so that, for a claim that has N supporting articles, there are N independent classifications $\hat{l_1},..,\hat{l_N}$. These are aggregated into a single classification outcome $\hat{y}$ through a majority voting, meaning that the claim is assigned to the class most frequently represented among the $\hat{l_k}$ values.

Table 6.1: The Table shows statistics of the Snopes-A, PolitiFact and Snopes-B datasets. It provides total number of claims and articles and an accurate information about the classes distribution for claims and articles.

	Snopes-A Dataset	**PolitiFact Dataset**	**Snopes-B Dataset**
Total Claims	4341	3,568	4,856
True Claims	1164 (26.8%)	1867 (52.3%)	1,277 (26.29%)
False Claims	3177 (73.1%)	1701 (47.6%)	3,579 (73.70%)
Total Articles	29,242	29,556	4,856
True Articles	7507 (25.7%)	15,019 (50.8%)	1,277 (26.29%)
False Articles	21,735 (74.3%)	14,537 (49.18%)	3,579 (73.70%)

6.4 Experiment and Result

The main goal of the experiments is to build an objective approach that can help journalists to distinguish between cases. We analyse deeply the performance of the system by conducting several experiments to answer the research questions proposed in Section 6.1. In this section, the experimental dataset is presented. Further, the setting of the hyperparameters of the experiments is explained. Finally, the deep analysis of the results is investigated.

6.4.1 Experimental Datasets

We assess the effectiveness of our model by conducting a set of experiments on three real-world datasets. We use three publicly available datasets for fake news classification: Snopes-A dataset [239], PolitiFact dataset [239] and Snopes-B dataset [146]. This allows us to fairy compare the state of the art models in the literature. The datasets are explained in depth in Chapter 5, Section 5.6.

In Snopes-A and PolitiFact datasets, the articles were retrieved based on different factors. They were assessed based on identifying the language of the reporting articles (i.e. bias and subjectivity), the articles' stance towards the claim (i.e. whether it supports or refutes the claim) and the reliability of the web sources generating the articles. More details about the process is described in Chapter 5 Section 5.4. There are more than two classes: mostly true, half-true, mostly false and half false. As we are considering only binary credibility labels, we map mostly true and half-true into class label true; and mostly false and half false into class label false. Claim sources and evidence sources were ignored in this study. We used

the evidence articles as they exist in these two datasets, which allow us to fairy compare the results reported in [241].

However, Snopes-B dataset comes without any evidence articles and to compare our model in a fair manner with the study in [146], we used their approach to retrieve external articles. A short query was generated to the Google search API and instead of querying with the full claim, verbs, nouns and adjectives were considered to get high quality search results. Then, the web documents were retrieved, ignoring any results that come from unreliable sources that exist in their database.The web documents were split into three groups and the most similar group to the claim was taken by calculating the cosine similarity. At the end, only one supporting article was related to a target claim. Regardless of whether their approach is effective or not (which is behind the scope of this thesis), we used the same technique for a fair comparison. Table 6.1 shows statistics of the three datasets.

6.4.2 Hyperparameter Settings

The datasets have been split into 5 disjoint subsets through a random process such that all supporting articles belonging to a given claim belong to the same subset. In this way, it is possible to apply a k-fold approach $(k = 5)$ and perform claim-independent experiments, meaning that the supporting articles belonging to the given claim never appear in both training and test set. In such a way, it is possible to ensure that the approach actually detects misinformation and does not simply recognise the identity of the claims. Every time a fold has been used as a test set, the union of the remaining four has been split into training set (90% of the material) and validation set (10% of the material). This latter has been used to select the value of the hyperparameters through hyperparameter optimisation.

The space of the hyperparameters (initial learning rate α_0, number of training epochs t, batch size b , number of hidden neurons u_0 for LSTM, self-attention hidden units u_1, L2 regularisation coefficient λ_0 and attention diversity regularisation coefficient λ_1) was searched through a Bayesian Optimisation [225]. At the end of this phase, the hyperparameter values leading to the highest validation accuracy are α= 0.001, t= 30, b= 32, u_0= 32, u_1= 250, λ_1= 0.004 and λ_0= 1.0 for Snopes-A and λ_0= 0.0 for PolitiFact and Snopes-B. According to a practice common in the literature, the initial learning rate has been progressively reduced over the successive training epochs using the expression $\alpha = \alpha_0 \beta^{\phi/\delta}$ where β=0.96 is the

decay rate, ϕ is the global step and δ =500 is the number of decay steps.

Additionally, the number of words at which an article is truncated, has been set to 80 for Snopes-A and 100 for PolitiFact and Snopes-B (section 6.4.6 for more details). The number of words at which a claim is truncated is 9 for all datasets. The embedding size for both claim and article inputs is 100. We set 10 different rows (p) for the matrix embedding (see section 6.3.1). All models and training methodologies have been implemented with Tensorflow. The models are trained through back-propagation and the loss function is the categorical cross-entropy [70]. The cross-entropy loss function was weighted to balance classes for Snopes-A. The models are trained using Adam optimiser [152].

6.4.3 Evaluation

To evaluate and compare the performance of our model with other state-of-the-art methods, we considered various evaluation metrics, which include: AUC, Accuracy, Macro F1 and Micro F1. The prevalent problem in many of the public fake news classification dataset is the number of ground truth fake news higher than true news. This imbalanced distribution between classes gives an unreliable result when using accuracy measures. This induces a situation where a classifier always predicts that the majority class will achieve high accuracy. Area-Under-Curve (AUC) for the ROC (Receiver Operating Characteristic) curve is proven to be statistically consistent and more discriminating than accuracy in an imbalanced classification problem [179]. It shows how well the probabilities from the positive classes are separated from the negative classes. However, we also calculate the accuracy to be comparable with the results reported in the state-of-art models. In addition to the accuracy and AUC matrices, we measure the performance with precision, recall and F1 matrices. F1 score can be interpreted as a weighted average of the precision and recall, this weighted average is either macro average or micro average. Macro average calculates metrics globally between labels, and micro average calculates metrics for each label. (See Appendix A, Section .3, for more details about hyperparameter and model selection).

6.4.4 Baselines

We compared the performance of our approach with the following competitive baseline approaches that have studied external evidence in the fake news detection:

Table 6.2: The Table shows the performance at the level of the claim, i.e. after that a majority vote was applied to all articles reported by a given claim. The performance was computed by true claims accuracy, false claim accuracy, Macro-F1, Micro-F1 and AUC. Since all experiments were repeated 10 times, the values were accompanied by their respective standard errors.

.

Dataset	Models	True Accuracy	False Accuracy	Macro-F1	Micro-F1	AUC
Snopes-A	DeClarE	41.0±0.003	59.0±0.068	49.0±0.026	0.53±0.031	0.50±0.023
	Claims Only	-	-	-	-	-
	Our Approach (Plain)	35.25±0.003	79.6±0.003	55.0±0.002	66.0±0.002	55.0±0.003
	Our Approach (Plain+Attention)	42.37±0.003	80.06±0.003	63.0±0.003	70.0±0.002	63.0±0.002
	Our Approach (Full)	**50.0**±0.016	**81.0**±0.005	**64.2**±0.009	**73.3**±0.007	**64.0**±0.009
PolitiFact	DeClarE	52.0±0.003	52.0±0.003	52.0±0.031	52.5±0.0031	52.0±0.052
	Claims Only	58.01±0.009	47.49±0.001	55.85±0.012	55.82±0.001	55.82±0.002
	Our Approach (Plain)	60.0±0.007	58.1±0.021	57.0±0.025	57.0±0.033	56.8±0.031
	Our Approach (Plain+Attention)	61.0±0.031	60.1±0.041	57.4±0.036	57.4±0.051	57.0±0.023
	Our Approach (Full)	**63.4±0.003**	**60.1±0.006**	**58.0±0.004**	**58.0±0.005**	**59.0±0.005**
Snopes-B	LSTM++ approach	**86.62±0.021**	76.43±0.005	**0.84±0.045**	**0.84±0.007**	62.0±0.024
	Claims Only	-	-	-	-	-
	Our Approach(Plain)	56.21±0.049	90.71±0.054	80.0±0.094	82.0±0.034	83.0±0.025
	Our Approach(Plain+Attention)	61.18±0.032	94.11±0.002	80.0±0.022	83.02±0.002	84.0±0.023
	Our Approach(Full)	76.09±0.058	**95.51±0.047**	80.0±0.010	83.0±0.035	**86.0±0.023**

- DeClarE, which is a recent approach based on LSTM with attention mechanism [241].
- LSTM++ approach, which uses the claim text, snippet, web document and different similarity vectors against the claim and they were fed to LSTM networks [146]

To further investigate the impact of each component and the self-attention mechanism in our approach, we designed several baselines for comparison. They are simplified variations of our approach by removing certain components as follow:

- Our approach (Claim Only), it is considered the claim inputs only, which is a dense layer for the claim encoder, without the evidence articles .
- Our approach (Plain), it is our approach with a dense layer for claim encoder and Bi-LSTM for article encoder without using self-attention.
- Our approach (Plain+Attention), which is our approach with a dense layer for claim encoder and self-attention on the top of Bi-LSTM for article encoder with using simple vector representation instead of two-dimensional representation matrix.
- Our approach (Full), which is an end-to-end algorithm with a dense layer for claim encoder and self-attention on the top of Bi-LSTM, considering a two-dimensional representation matrix for input article.

6.4.5 Experimental Result

Training the models requires a random initialisation step that changes the performance of the approach. Hence, every experiment of this work was replicated 10 times, and the value of every performance metric was the average of the values observed in the 10 repetitions. The low variance over the 10 repetitions implied that the models were fairly resilient to changes in initialisation and, therefore, the averages can be considered realistic estimates of the performance. Table 6.2 shows the baseline and the proposed approaches at the level of the claim, i.e. after that a majority vote was applied to all articles reported by a given claim. Like the state-of-the-art models, we reported our results based on the performance at each particular class (e.g. accuracy for true claims and accuracy for false claims), macro F1, micro F1 and AUC. While the authors of LSTM++ model has distributed the source code2, DeClarE

model does not; thus, we reproduced their code. In both cases, we ran the models to get the results by taking the average of the values of 10 repetitions. We could observe that the overall performance of the proposed model (i.e. Full Approach) outperforms that of baselines to a statistically significant extent in all cases ($p < 10^3$).

Regarding claims only approach, the claim inputs only without enriching them with their relevant articles are insufficient. This is because they do not perform better than claims supplemented with relevant articles to a statistically significant extent. Overall, this means that our approaches benefit to a greater extent from adding complementary information. For example, on PolitiFact dataset, specifically, our approachesclearly outperform the claims only approach. However, Snopes-A and Snopes-B datasets do not learn at all from considering only claim inputs (symbolising in the Table 6.2). In addition, all our model versions with several configurations mostly outperform the baseline models. This means that each component of our model has an important contribution to reach the best misinformation detection performance. If we removed one of the components, the performance would drop by a certain degree, as shown in Table 6.2. Moreover, the system accuracy for detecting fake claims is mostly higher than the true claims. This is important as we aim to notify readers to the suspicious statements as early as possible before widely spread. These observations prove the advantages of our model and validate the effectiveness of the model in detecting fake claim on the web.

To compare our approaches with the state-of-the-art, our model outperforms DeClarE by a large margin on both Snopes-A and PolitiFact datasets. Importantly, the result performance of DeClarE model that we reported in Table 6.2 differs from what was reported in their paper. This is because we reproduced their source code and reported what we observed. One possible explanation of the result differences is that they may apply $k-$fold approach without considering claim independent, meaning that the evidence articles related to a given claim may appear in both training and testing, thereby inducing simple recognition of the claim identity. However, our model outperforms LSTM++ approach in Snopes-B dataset by a large margin of 24% in AUC and 19% in false claims accuracy. The latter is much important measurement because it has the most negative consequences and, therefore, should be as limited as possible. However, the F1 measurements and the true accuracy for LSTM++ are better than our model, probably because each claim in Snopes-B has one best sentence taking from

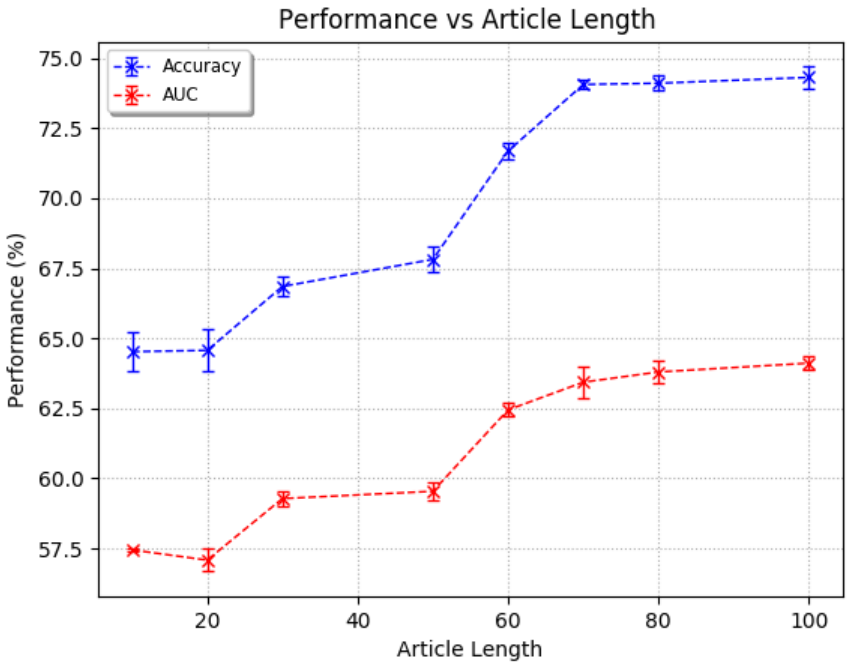

Figure 6.2: The plot shows the performance based on accuracy and AUC as a function of number of words in the article considered, 10 to 80 words counting by tens are analysed. On average, increasing the article length induces better performance.

the supporting article, this is how LSTM++ approach works, while our approach addresses the problem of article length and the effective of retrieving multiple evidence articles. This is important from an operational viewpoint since it mimics real-life checking process.

In the following sections, we conducted different experiments to analyse the performance of the model more deeply. Observably, our model outperforms the other models in all the datasets; thus, what we observed is not just the effect of the data. At that point, choosing one dataset to comprehensively analyse the results is sufficient. In particular, the analysis focuses on the Snopes-A, the most widely benchmark that has been used in the literature.

6.4.6 Analysis of Article Length

As we discussed earlier in this chapter, sequential models are typically difficult to capture long-term dependencies, especially when the input sequences are too long [35, 136]. Therefore, the last hidden state may not accurately express the semantics of the article thus self-attention is proposed. This experiment was conducted to analyse the effect of model's performance regarding article length. Figure 6.2 displays a plot of the relation between the

performance of the model about accuracy and AUC and the number of words in the article considered. In this study, 10 to 80 words counting by tens were analysed and 100 words in the articles were considered (see Section 6.4.2). The model consistently performs better as the length of articles increases, validating that incorporating article semantics by self-attention can capture the article's long-term dependencies. We can conclude that increasing the length of the article can capture all the key factors that contributes to identify the claim identity. This observation follows the actual process of manual fact-checking which entails reading the entire article to give a final decision towards a claim [45].

6.4.7 Analysis of Model Confidence

The application of the majority vote over the articles of a given claim allows us to achieve an accuracy close to 75% at the level of the claims. We compared our findings with several benchmarks and concluded that our results can be considered satisfactory, as it outperforms the state-of-the-art results (see Section 6.4.5). Also, one of the real benchmark for comparison is human fact-checkers, a traditional process to verify information where journalists manually assess the claims. Several studies in social psychology and communications found that human can identify fake contents slightly better than chance. The accuracy ranged in 55%-58%, with a mean accuracy of 54% over 1,000 participants in over 100 studies [315].According to these results, all the proposed approaches in this study seem to outperform human fact-checkers. This indicates that one possible approach for applying misinformation detection technologies is to classify the cases in which the outcome of a system can be trusted while leaving the rest to journalists attention. This conforms with recent developments, indicating that AI driven technologies should work alongside rather than replacing their users [83].

This can be done by considering the confidence score described in Chapter 4, Section 4.3. The score c can be calculated by Equation 4.5, where N is the total number of articles reporting a target claim, and n_D is the number of articles that, for a given claim, is assigned to fake/not credible class. The rationale behind this score is that the more close n_D is to N, the more the approach appears to be confident about its decision. This measure allows one to rank the claims according to the value of c (from largest to smallest) and to consider the accuracy at position k. If higher values of c induce right decisions, the accuracy should be high if only the top positions of the ranking are considered. Figure 6.3 shows the average accuracy (i.e.

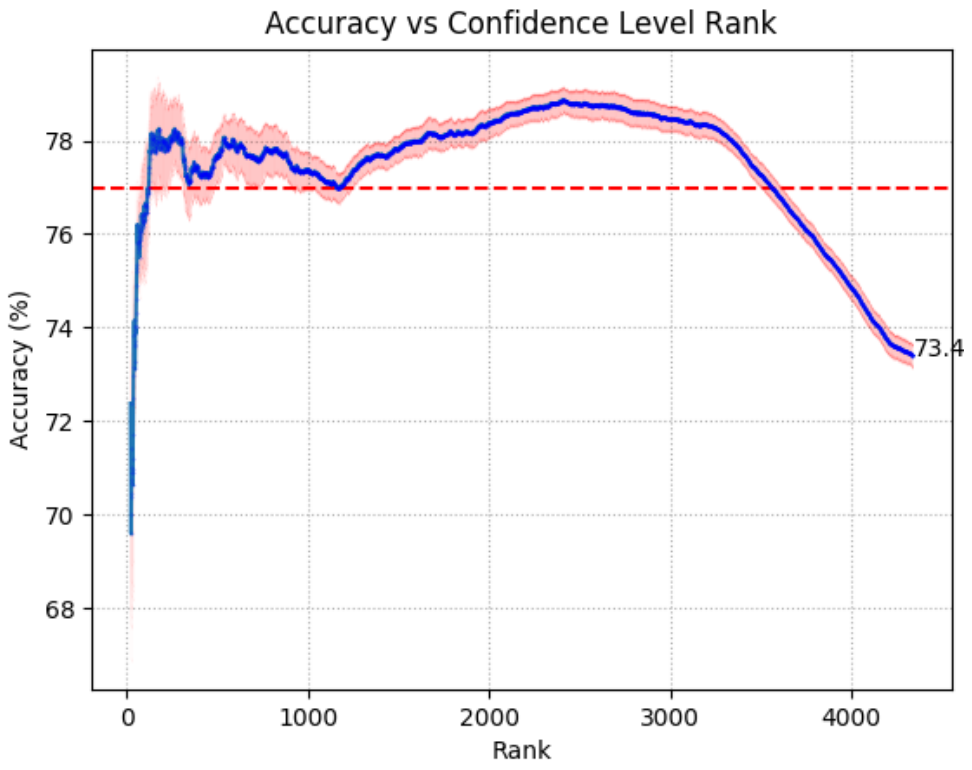

Figure 6.3: The plot shows the accuracy over the k claims for which the approach shows the highest confidence scores for all possible values of k.

average accuracy taken over 10 repetitions with their respective standard errors) at position k of the confidence score ranking, i.e. the accuracy observed when considering only the claims that correspond to the k highest confidence scores. There are very high confidence values for wrong decisions at the beginning, meaning that the approach does not appear to confirm such an expectation. Consequently, higher confidence does not necessarily mean that the classification is correct. One possible explanation for this result is that it is whether the contents are written very carefully to look like the real ones, fooling the reader who does not check for reliability of the sources or the arguments in its content, or the contents are originally true but it looks like the fake ones. This observation highlights the challenges of the problem.

However, it is possible to measure to what extent higher confidence scores tend to be associated to correct classifications and, correspondingly, to measure how much the confidence score can actually be trusted. The plot shows that there are many intervals where the approach has an accuracy exceeding 77% (e.g. when considering the ranking claims between 256 to 348, 755 to 1162 and 2413 to 3561). Also, the approach has an accuracy above 73.4%

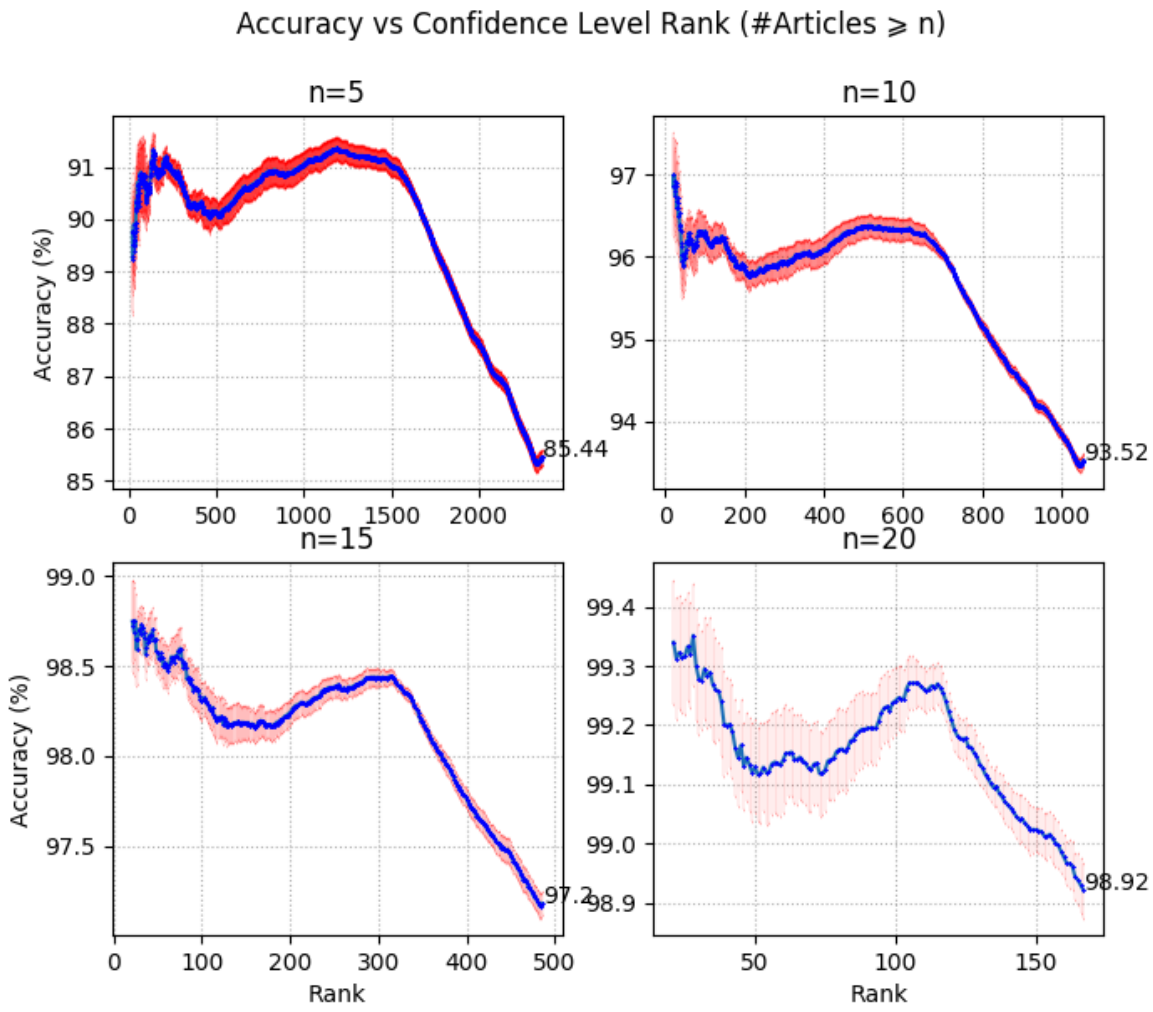

Figure 6.4: The plots show the accuracy when considering only the k claims with the highest confidence values, considering the claims that have more than n evidence articles

when considering 2413 to 4341 ranking claims (corresponding to four-ninths of the data). Here, this observation suggests that the confidence score can be used as a criterion to accept or discard a decision made by the system, while, concurrently, ensuring that the accuracy for the accepted cases is above an acceptable threshold. This is important from an operational viewpoint because it can allow the trained journalists to identify the subset of the claims for which the system performs to an acceptable level, allowing the system to potentially reduce by four-ninths the workload of the trained journalists while still keeping the accuracy above 73.4%. This is critical because it can increase the efficiency of detecting services and, correspondingly, reduce the costs associated to detecting false claim.

The results above tackle all claims regardless of how many articles are available for a claim. Note that the number of supporting articles for each claim in the dataset differs.

Therefore, analysing whether the number of articles reporting a claim impacts the system confidence is imperative. Hence, we considered only the claims that have above n supporting articles and ignored the claims lacking enough materials (i.e. $n < 5$). Figure 6.4 displays plots of the relation between the accuracy when considering only the k claims with the highest confidence values and the number of evidence articles for each claim n. Interestingly, it shows that increasing the number of articles regarding a claim increases the confidence of the system with the correctness of its decisions, especially at the beginning. This highlights that multiple evidence articles for a claim constitute an important source of information for improving the correctness of confidence score of the system. This follows manual fact-checking process since the journalists scan the web to investigate the claim identity, and the more reliable articles the journalists read, the more confident the result is.

6.4.8 Analysis of Attention Weights

When reading an article, a human fact-checker usually focuses on several words with different attention weights. Similarly, in our model, the attention weights purpose is to highlight how much each word influences an article during the learning process. This visualisation gives insight into the internal process. Therefore, we can know which words the embedding considers and which are ignored by the embedding. This also interprets learning representations to the end-users. Higher weights are coloured by darker shades to show the most significant words in the learning process that play a major role in credibility decision. Figure 6.4.8 exemplifies true and false claims, along with their supportive articles and each article has different (p) weights(see section 6.3.1).Different article weights help to extract the latent information and provide higher level semantics of the article, which the model exploits. For example, in Figure 6.4.8(a) the words in the article *'allow','turn on your camera'* and *'could read'* support the model to recognize the credibility of a claim *'Photos taken with your smart-phone can provide others with the locations of the people pictured and allow hackers to clone your phone'*. However, Figure 6.4.8(b) exemplifies false claim *'Football player Michael Vick broke legs in auto- mobile accident'*. The model highlights different words that help to understand the context such as: *'misleading', 'did not' and 'satire'*.

Android app can turn on your phone camera to allow hacker to snoop around your homePersonal data and private moment can be glean from image By Eddie Wrenn for Mail Online Published A new app can virtually steal from your home by turn on your phone camera and beaming image back to thief The software can even build up a 3D model of your home from which the hacker can inspect your room potentially glean information about valuable in your home calendar entry a well a spy on you The app be create by US military expert at Naval Surface Warfare Center in Crane Indiana to show how cybercriminals could operate in the future New method Hackers can build up a virtual image of a room leave and then inspect detail at close- up - pick up detail such a cheque number bottom right The Plaice Raider creator even demonstrate how they could read the number of a cheque book

Android app can turn on your phone camera to allow hacker to snoop around your homePersonal data and private moment can be glean from image By Eddie Wrenn for Mail Online Published A new app can virtually steal from your home by turn on your phone camera and beaming image back to thief The software can even build up a 3D model of your home from which the hacker can inspect your room potentially glean information about valuable in your home calendar entry a well a spy on you The app be create by US military expert at Naval Surface Warfare Center in Crane Indiana to show how cybercriminals could operate in the future New method Hackers can build up a virtual image of a room leave and then inspect detail at close- up - pick up detail such a cheque number bottom right The Plaice Raider creator even demonstrate how they could read the number of a cheque book

Android app can turn on your phone camera to allow hacker to snoop around your homePersonal data and private moment can be glean from image By Eddie Wrenn for Mail Online Published A new app can virtually steal from your home by turn on your phone camera and beaming image back to thief The software can even build up a 3D model of your home from which the hacker can inspect your room potentially glean information about valuable in your home calendar entry a well a spy on you The app be create by US military expert at Naval Surface Warfare Center in Crane Indiana to show how cybercriminals could operate in the future New method Hackers can build up a virtual image of a room leave and then inspect detail at close- up - pick up detail such a cheque number bottom right The Plaice Raider creator even demonstrate how they could read the number of a cheque book

Android app can turn on your phone camera to allow hacker to snoop around your homePersonal data and private moment can be glean from image By Eddie Wrenn for Mail Online Published A new app can virtually steal from your home by turn on your phone camera and beaming image back to thief The software can even build up a 3D model of your home from which the hacker can inspect your room potentially glean information about valuable in your home calendar entry a well a spy on you The app be create by US military expert at Naval Surface Warfare Center in Crane Indiana to show how cybercriminals could operate in the future New method Hackers can build up a virtual image of a room leave and then inspect detail at close- up - pick up detail such a cheque number bottom right The Plaice Raider creator even demonstrate how they could read the number of a cheque book

(a) True Claim: "Photos taken with your smart-phone can provide others with the locations of the people pictured and allow hackers to clone your phone."

unsubstantiated misleading messages Facebook related satire oddities security computer security malware threats email security spam control extras scam victim stories nutshell special features Michael Vick did not break his legs in a car accident outline message circulating vigorously via social media claims that oft reviled football player Michael Vick has broken both legs in a car accident brief analysis the claims in the message are untrue Michael Vick did not break his legs in a car accident the information is derived from a fake news item published via the satirical entertainment website detailed analysis and references below example scroll down

unsubstantiated misleading messages Facebook related satire oddities security computer security malware threats email security spam control extras scam victim stories nutshell special features Michael Vick did not break his legs in a car accident outline message circulating vigorously via social media claims that oft reviled football player Michael Vick has broken both legs in a car accident brief analysis the claims in the message are untrue Michael Vick did not break his legs in a car accident the information is derived from a fake news item published via the satirical entertainment website detailed analysis and references below example scroll down

unsubstantiated misleading messages Facebook related satire oddities security computer security malware threats email security spam control extras scam victim stories nutshell special features Michael Vick did not break his legs in a car accident outline message circulating vigorously via social media claims that oft reviled football player Michael Vick has broken both legs in a car accident brief analysis the claims in the message are untrue Michael Vick did not break his legs in a car accident the information is derived from a fake news item published via the satirical entertainment website detailed analysis and references below example scroll down

unsubstantiated misleading messages Facebook related satire oddities security computer security malware threats email security spam control extras scam victim stories nutshell special features Michael Vick did not break his legs in a car accident outline message circulating vigorously via social media claims that oft reviled football player Michael Vick has broken both legs in a car accident brief analysis the claims in the message are untrue Michael Vick did not break his legs in a car accident the information is derived from a fake news item published via the satirical entertainment website detailed analysis and references below example scroll down

unsubstantiated misleading messages Facebook related satire oddities security computer security malware threats email security spam control extras scam victim stories nutshell special features Michael Vick did not break his legs in a car accident outline message circulating vigorously via social media claims that oft reviled football player Michael Vick has broken both legs in a car accident brief analysis the claims in the message are untrue Michael Vick did not break his legs in a car accident the information is derived from a fake news item published via the satirical entertainment website detailed analysis and references below example scroll down

unsubstantiated misleading messages Facebook related satire oddities security computer security malware threats email security spam control extras scam victim stories nutshell special features Michael Vick did not break his legs in a car accident outline message circulating vigorously via social media claims that oft reviled football player Michael Vick has broken both legs in a car accident brief analysis the claims in the message are untrue Michael Vick did not break his legs in a car accident the information is derived from a fake news item published via the satirical entertainment website detailed analysis and references below example scroll down

unsubstantiated misleading messages Facebook related satire oddities security computer security malware threats email security spam control extras scam victim stories nutshell special features Michael Vick did not break his legs in a car accident outline message circulating vigorously via social media claims that oft reviled football player Michael Vick has broken both legs in a car accident brief analysis the claims in the message are untrue Michael Vick did not break his legs in a car accident the information is derived from a fake news item published via the satirical entertainment website detailed analysis and references below example scroll down

unsubstantiated misleading messages Facebook related satire oddities security computer security malware threats email security spam control extras scam victim stories nutshell special features Michael Vick did not break his legs in a car accident outline message circulating vigorously via social media claims that oft reviled football player Michael Vick has broken both legs in a car accident brief analysis the claims in the message are untrue Michael Vick did not break his legs in a car accident the information is derived from a fake news item published via the satirical entertainment website detailed analysis and references below example scroll down

unsubstantiated misleading messages Facebook related satire oddities security computer security malware threats email security spam control extras scam victim stories nutshell special features Michael Vick did not break his legs in a car accident outline message circulating vigorously via social media claims that oft reviled football player Michael Vick has broken both legs in a car accident brief analysis the claims in the message are untrue Michael Vick did not break his legs in a car accident the information is derived from a fake news item published via the satirical entertainment website detailed analysis and references below example scroll down

unsubstantiated misleading messages Facebook related satire oddities security computer security malware threats email security spam control extras scam victim stories nutshell special features Michael Vick did not break his legs in a car accident outline message circulating vigorously via social media claims that oft reviled football player Michael Vick has broken both legs in a car accident brief analysis the claims in the message are untrue Michael Vick did not break his legs in a car accident the information is derived from a fake news item published via the satirical entertainment website detailed analysis and references below example scroll down

(b) False Claim: "Football player Michael Vick broke legs automobile accident.".

Figure 6.5: The Figure shows user interpretation via attention weights.

6.5 Conclusions

The chapter described a novel fake news detection algorithm based on the signal from the textual claim and set of evidence articles. To capture higher level semantics for the articles and to mimic human reading process, we applied self-attention mechanism, in which we represented each article using a matrix-based representation. Then, a majority vote over several external articles of a given claim was applied to assess the claim's credibility.

Extensive experiments on large-scale real-world data demonstrated the effectiveness of the system over state-of-the-art baselines. The experiments also showed the impact of the article's length on the performance in which the system consistently performs better as the length of the article increases. Also, the experiments illustrated several application scenarios where the proposed approaches used confidence measures that identify the likeliest cases to be correctly classified. In this way, the systems could reduce by four-ninths the workload of the trained journalists while still keeping the accuracy above 73.4%. The analysis of the performance also showed that the more supporting articles for a claim, the higher the performance of the system. This agrees with manual fact-checking process in which the journalists scan the web to investigate the claim identity, and the more reliable articles the journalists read, the more confident the result is. Finally, the system provided interpretability of results, which potentially help a reader in understanding the classification decision.

Chapter 7

Conclusions

7.1 Introduction

Binary text classification problems have been widely studied and addressed in many real applications, such as depression detection and misinformation identification. In particular, with the latest NLP and text mining breakthroughs, several studies have developed applications that exploit text classification methods. However, the classification of texts that include an insufficient or limited number of words is particularly challenging. In this thesis, we can define the insufficiency of texts as any text that does not adequately represent the critical features of a problem. Alternatively, the text features alone do not function as well as when they are paired with additional information sources. Enriching texts using domain-specific knowledge is one of the successful solutions for addressing the problems described. In this thesis, we extensively researched the textual data in the depression and misinformation domains and made a list of contributions for each domain.

The rest of this chapter is organised as follows. Section 7.2 summarises the contributions of this thesis for each domain, and Section 7.3 discusses limitations of the research and prospective future research.

7.2 Contribution Summary

Regarding the depression problem, we found that most previous studies have addressed depression by the inference from behaviour of scores resulting from the administration of self-assessment questionnaires, such as BDI-II or different versions of PHQ (see Chapters 3 and

4). In this thesis, however, the disparity between depressed and non-depressed participants has been made by psychiatrists rather than by administering self-assessment questionnaires. This is a privilege since it increases the probability of data representing the true difference between depressed and non-depressed participants. Alternatively, it ensures that the problem addressed in the work is depression detection and not the inference of self-assessment scores. This is remarkable due to multiple biases of self-assessment questionnaires and inconsistency that may happen in filling out the questionnaires, especially by people affected by depression (see Chapters 3 and 4).

All the experiments in Chapters 3 and 4 were conducted over a corpus of 59 interviews. To effectively tackle this limited amount of data, the interviews were segmented into clauses, i.e. to manually extracted linguistic units that include a noun, a verb and a complement. Given that the average number of clauses per participant is 114, this allows one to perform, for every person, a large number of clause level decisions. In particular, Chapter 3 examined whether applying a high-quality contextualised embedding (i.e. BERT) for the interview transcriptions can contribute to the result. It showed that using of BERT rather than traditional word embedding does not induce performance improvements, possibly because there is a mismatch between the dictionary used during the currently available versions of BERT for Italian texts and the dictionary of the data used in this work, which may be due to the small vocabulary size of the multilingual BERT-base edition compared to the English one. Also, it is probably because the clauses are short (the average length is 3.9 words); thus, the context might carry insufficient contextual information.

No single indicator on its own can sufficiently identify depression signs due to intrinsic variations in the speech system [79]. This suggests that linguistic cues alone might not be adequate to explain a person's mental characteristics and states, necessitating the inclusion of knowledge from other sources. Therefore, Chapter 3 introduced the proposed model that integrates speech signals and their transcriptions via varying multimodal methods that consider both what people say and how they say it. The key explanation for focusing on linguistic and acoustic features of speech is that depression interferes with the neural processes underlying language and communication, therefore leaving detectable traces in both what people say and how they say it. For textual data performance, the experiments show that the efficiency of textual transcriptions does not produce better performance compared to when it is paired with

its signals (74.1% vs 85.5% on the best multimodal method). This highlights the significance of utilising another source of information other than relying only on textual transcriptions.

A more comprehensive analysis can be found in Chapter 4. The experiments showed that the key variations between the methodologies of unimodal and multimodal are that the multimodal methodologies appear to have more uniform clause-level accuracy across the participants. This is significant since it contributes to greater accuracy at the person level, the metric that really matters from an application viewpoint. This result originates from the tendency of certain participants, particularly depressed ones, to manifest their condition either by what they say or by how they say it, but not by both. Also, the analysis of GMU weights showed that the role of language is likely to be more important for control participants than the depressed. These are important findings that are unprecedented in the literature. One of the most important consequences of the observations is that the two modalities appear to be a source of *diversity*—the tendency of different classifiers to make different mistakes [252]. Such a property was shown to increase the chances of classifier ensembles [153] to outperform their best members [162]. Therefore, in the experiments of this work, diversity across modalities might be at the origin of the significant performance difference between the multimodal approach and the best unimodal recogniser (83.5% vs 74.1% based on accuracy). In this respect, the main question seems to be not whether there is a modality that is better than the others (as the state-of-the-art in Chapter 3 seems to suggest), but whether it is possible to find multiple modalities that can correct each other when one or some of them do not carry reliable information. Furthermore, the experiments suggested that the modality carrying the most reliable information could be different for people belonging to different classes. This further confirms that the best strategy is not necessarily looking for the best modality but for a set of modalities that cover all groups of people appearing in the data.

Chapter 4 also illustrates several application scenarios in which the proposed approaches use confidence measures to identify the likeliest cases to be correctly classified. The experiments showed that the doctors' workload can be reduced by up to two-thirds, although still maintaining above 90% accuracy. This is significant due to the increasing efficiency of screening services, thereby reducing the costs related to the diagnosis of depression. In this respect, our approaches can assist psychiatric and counselling services by potentially enabling psychiatrists to focus on challenging situations while leaving machines to handle the

more obvious ones. This supports the previous findings, indicating that the most effective way to utilise AI is by supporting people and not replacing them.

Fast depression detection addresses several issues in clinical practice. First, depressed people are inclined to avoid social relations to speak less than non-depressed people (see Chapter 3). The potential of identifying depression with minimal material can help to tackle such a pattern and produce good results for individuals who have difficulties in retaining an interview. Second, detecting depressed individuals out of many people who contact counselling services due to momentary distress, but are not having depression, is critical. Therefore, in Chapter 4 deeply analysed the performance as a function of the time, based on the number of clauses, to see how many materials the system need to detect depression. The experiments show that it is possible to perform depression detection with less than 10 seconds without significant performance losses, especially for recall. The results do not depend on the protocol applied at the beginning of the interviews, but on the amount of data. This observation may explain that the state of depression patients tends to be manifested consistently, thus there is a high likelihood of accurately classifying any clause they utter.

Furthermore, the problem of misinformation was studied in Chapters 5 and 6. We highlighted the importance of utilising external evidence, such as web evidence, for credibility analysis of claims. Its importance lies in providing additional information that represents the central content of the claim more authentically and gives the user an interpretative explanation (see Chapter 5 for more details). In particular, Chapter 7introduced an approach designed to reduce the burden by assisting humans in verifying the veracity of textual claims based on joint interactions between a claim and its evidence articles. The experiments showed that the systems can reduce the workload of the trained journalists by four-ninths while still keeping the accuracy above 73.4%. This is important because it can improve the efficiency of detection services and, consequently, lower the costs of detecting false claims.

Chapter 6 showed that utilising complementary information for a claim is critical based on the fact that textual claims are quite short and may not be used accurately for classification. Confirming that, the experiments of this study notice that using claim inputs alone without supplementing them with their relevant articles is inadequate. This is because the performance of claims regarding the articles performs better, to a statistically significant extent, than the claims approach performance does alone. In addition, Chapter 6 also analysed

the system confidence, and it showed that when the system has high confidence values, this does not always induce its correct classification decisions. One potential reason for such a finding is that the contents are written very deliberately similar to the true ones or the contents are true but seem to be false. Such insight emphasises the problem's difficulties.

In addition, the effect of evidence articles on the number of articles required for a claim and article length were studied in Chapter 6. Regarding the number of articles, we observed that increasing the number of evidence articles enabled a better assessment of the credibility of claims. This finding agrees with the manual fact-checking process since the journalists scan the web to investigate the claim identity, and the more factual articles the journalists read, the more confident the result is. Regarding article length, the experiments showed that article length is critical in the credibility assessment of claims. It showed that increasing the length of the article can capture all the key factors that contribute to identifying the claim identity. This finding follows the actual process of manual fact-checking, which involves reviewing the whole article before reaching a final judgement on a claim.

Several machine learning models, including regression, Naive Bayes, decision tree and random forest are naturally interpretable. In regression, for example, coefficient weights show the importance of the features. Likewise, traditional feature selection methods also provide clarification of model decisions by providing the contribution of each feature. This highlights the importance of providing user-interpretable explanations as an additional production of the system that can justify the credibility assessment (see Chapter 5. In Chapter 6, we conducted an experiment using interpretable evidence that aimed to specify the informative words of evidence articles. This visualisation shows insight into the internal process; thus, we can know which words the embedding considers and which are ignored by the embedding.

7.3 Limitations and Future Work

This section discusses the limitations that arise from the thesis as a whole and provides possible directions for future research. Based on the thesis organisation, in Chapter 1, Section 1.4, the main works were introduced in Parts I and II; thus, this section is based on these parts only.

Part I

- In this work, we conducted several experiments on a new dataset that included a relatively small number of interviews. Having a larger number of subjects would allow validation of the investigated approaches and allow for the use of more advanced techniques. Future work should consider collecting more data and releasing the dataset to the public to allow for comparisons with state-of-the-art approaches.

- Gender, age, comorbidity and cultural variances are factors that can highly affect the diagnostic appropriateness. According to the National Comorbidity Survey, depression was found to affect around twice as many females (21.3%) than males (12.7%) [150]. Moreover, depression was found in 50%, 50–75% and 25% of patients suffering from Parkinson's disease, eating disorder and cancer, respectively. Future work should analyse the results of depression deeply regarding the participants' age, education level and gender to better understand how these factors may affect system performance.

- This work on depression detection involves Italian speakers; thus, it examined depressed individuals from broadly similar cultures. Future work should improve cross-cultural investigations and enhance the awareness of the effect of cultures on depression behaviour by gathering multiple cultural backgrounds in the data. Therefore, it has been stated that different cultures have different displays of depression and different cultural acceptance of depression. These differences could or could not influence an objective depression detection system. It would be critical to have the same paradigm to allow for the comparison of findings.

- The label in the dataset represents a snapshot of an individual's mental state. That is the current score of the individual's depressive symptoms does not reflect continuous data regarding a person's condition. Therefore, it is difficult to gauge how a system would fare over time. Future work should involve a longitudinal analysis that measures the efficiency of depression detection systems over time.

- This work investigated the diagnosis of clinical depression from linguistic and acoustic aspects of speech. However, other cues could be investigated in future work. The visual

modality, including body movement, head pose and eye activity, could be a rich source of cues to detect depression.

- Every interview was manually transcribed and segmented into clauses. However, speech segmentation and transcriptions can be performed automatically using advanced speaker techniques. Future work should consider a fully automated system to segment and transcribe the interviews, which might be feasible for the task of detecting depression.

Part II

- Our models for automated credibility assessment assign a classification label indicating how likely the information is to be true or false. This is done by enriching the claim with its supporting articles retrieved from the web. To rely on these retrieved articles, evaluating the article's trustworthiness (i.e. such as reliability and stance determination of the article) is critical. However, in this thesis, the evaluation of the article's trustworthiness is not considered. Future work should investigate and evaluate the evidence and automate the process of generating evidence.

- The analysis of the system confidence showed that a high confidence value of the system does not always induce correct decisions in classification. Future work should further investigate which cases the system makes wrong decisions while it has high confidence in them.

- This work on credibility assessment focuses on textual contents. With the rise of technology, however, misinformation has also affected digital content. Future work should consider developing models that assess the credibility of multimedia content. Overall, a significant number of important research questions remain open in the area of depression and misinformation domains, and we believe they should be tackled in future studies, possibly relying on our work in this thesis as a foundation.

Appendix A: Methodology

Most of the existing methods of the machine learning perform properly because of the features of inputs and human-designed representations. When machine learning is applied to the features of the data, it is simply about weight optimisation to get the best final prediction. However, deep learning can be considered as the formation of representation learning and machine learning. It seeks to jointly learn proper features over multiple levels of growing complexity and abstraction, and the final prediction.

In this Appendix, the reasons for deep neural network based models are reviewed and the general structure of a neural network is defined. Moreover, more advanced architectures of neural networks are examined including MLP, RNN, LSTM and Bi-LSTM. This chapter also explains how words and signals are represented in a way the machine can understand. Finally, the chapter shows different techniques for combining multiple representations of models.

.1 Why Deep Learning Is Important?

In our research, we utilised deep learning models for several reasons. First, hand-crafted features are time-consuming, frequently over-specified, and uncompleted. Considering having more than one modality or task, the process of extracting features is repeated for each of them. Therefore, deep learning algorithms are important as they can learn features automatically; the whole learning process can be automated more easily and many tasks could be addressed.

Second, many models in Natural Language Processing (NLP) are based on count-based models which suffers from generalisation when particular words during testing do not appeared within the training set. The so-called 'curse of dimensionality' is another characteristic of this problem. Because an index vector over a large vocabulary is very sparse, models can simply overfit to the training set. The traditional solution to such a problem is to either

use aforementioned hand-crafted features or rather simple target functions, such as in linear models. Instead of discrete word counts, deep learning models of language often utilise distributed word representation, which captures similarities between words and make models more robust. Distributed word representation is described in more details in Section .4.

Finally, deep learning models learn multiple levels of representation which are similar to the human brains. In NLP, for example, humans can handle sentences as compositions of words and phrases, and deep learning algorithms can use recursive architectures to process and compose meaningful representations through compositionality (see Section .4).

.2 Neural Networks: Definitions and Basics

Artificial Neural Networks (ANNs) are information processing models that are developed to simulate the network of neurons in human brain [198,266,271]. The underling structure of an ANN is a network of nodes connected to each other by weighted connections, with different weights. This is inspired by the biological model where the nodes refer to neurons and the weighted connections refer to power of the synapses between the neurons. The degree of importance of the given connection in ANN is indicated by the *weigh coefficient*. Producing an input to number of neurons activates the network, and this activation then extends to all the network beside the weighted connections. The electrical motion of the nerve cells has a sequence of sharp points. The purpose of activating the ANN node is to simulate the average firing rate of these sharp points.

Over the years, various ANNs have arisen with generally different properties. The major difference between them is the form of the connections. The first type is ANNs with noncyclic, which means that the computation can be performed sequentially. This type of networks are called *Feed-Forward Neural Networks* (FNNs). The wide application of FNNs is the *Multilayer Perceptron (MLP)*, networks and it is simply called *Neural Networks* (NNs) (see Section .2.2). The other type is ANNs with cycles, which means the processing can feed into itself. This type of network are called *Recursive or Recurrent Neural Networks* (RNNs) and they are far more complex, and we will describe them later in section .2.3. [39,271,338].

In the rest of this section, the structure of simple neuron model is presented. In addition, MLP, RNN, LSTM and Bidirectional LSTM models (Bi-LSTM) are described. The basic

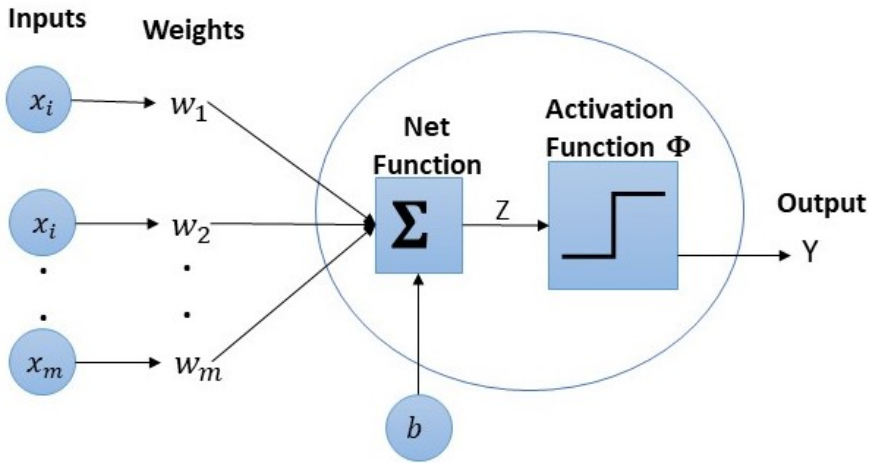

Figure 1: McCulloch and Pitts' neuron maps m inputs to one y output. The inputs x_i are multiplied by the weights w_i, and the neurons sum their values. The neuron activities when this sum is greater than specific threshold; otherwise it does not. The Figure is adapted from [196].

training process of any ANNs is also highlighted.

.2.1 Neuron Model

The most widely used neuron model is based on McCulloch and Pitts' neuron [198]. Figure 1 shows the graphical representation of the neuron (it is also called perceptron). The neuron receives input as a vector of length m $(x_1, x_2, \ldots x_m)$, where the neuron has a set of free parameters which learns during the training process.

$$\theta = (w, b), \qquad \in \mathbb{R} \times \mathbb{R}^m \tag{1}$$

where b is referred to bias and $w = (w_1, w_2, \ldots w_m)$ is termed the synaptic weight vector. Therefore, $m+1$ is the number of parameters of this neuron model.

A neuron comprises two components which are the *net function* and the *activation function* (it is also called transfer function). These two functions are described below:

Net Function

The net function defines how the input signals are combined inside the neuron. The weighted sum of inputs is computed as follows.

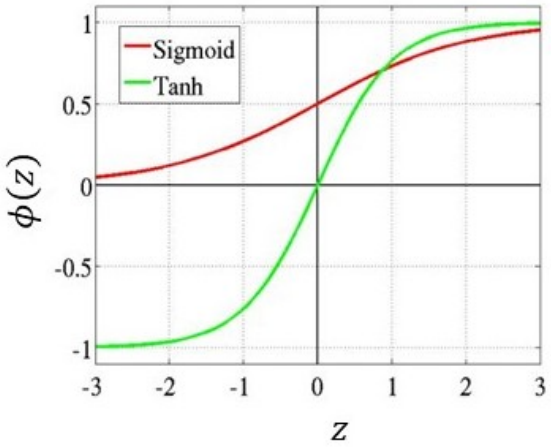

Figure 2: The graph of the two activation functions showsing their distinct output values. The Tanh function outputs ranges from -1 to 1, whereas Sigmid function outputs ranges from 0 to 1.

$$z = b + \sum_{i=1}^{m} x_i w_i \tag{2}$$

Then, the linear combination z is transformed by activation function Φ to specify the output signal y from the neuron regarding its net input signal z (see the following paragraph).

Activation Function

The activation function is extremely crucial which is a mathematical function attached to each neuron in a neural network. It is biologically inspired by activities in human brains where certain neurons are either firing (or are activated) or they are not firing (or are not activated) by different stimuli. The activation function does the non-linear transformation to the input, making it capable to learn and perform more complex tasks. It regulates whether the signal will progress further through the network to affect the result. It maps the resulting values into the desired range which varies from a function to another based on the problem itself. It also aims to fine-tune the weights of the inputs until the margin of error of neural network is minimal. The activation function Φ can be formulated as follows.

$$y = \Phi(z) \tag{3}$$

The most commonly activation functions are *logistic sigmoid*, *softmax* and *hyperbolic tangent (tanh)*. Figure 2 illustrates the two distinct activation functions.

The logistic sigmoid function transforms the values between the range 0 and 1. It is widely used for binary classification problems. The sigmoid function can be defined as:

$$\Phi(z) = \sigma(z) = \frac{1}{1 - e^{-z}} \tag{4}$$

However, the softmax function is often described as a combination of multiple sigmoid functions. It determines a probability value for each class by representing a categorical distribution, i.e. a probability distribution over 'N' different possible outcomes. It is suitable for multi-class problems since the output for a neuron (z) depends on the output of the other neurons in the same layer. At the end, the total values add up to 1, and all class probabilities are between 0 and 1. The mathematical expression of softmax function is:

$$\Phi(z) = softmax(z) = \frac{e^{z}}{\sum_{i=1}^{N} e^{z_i}} \tag{5}$$

Finally, the tanh function is very similar to the sigmoid function. The difference is in the symmetry around the origin. The range of values is from -1 to 1. The mathematical expression of tanh function is

$$\Phi(z) = tanh(z) = \frac{e^{2z} - 1}{e^{2z} + 1} \tag{6}$$

.2.2 MultiLayer Perceptron Model

A multilayer perceptron network (MLP) or neural network (NNs) is a feed forward artificial neural network. The underlying structure of MLP is a layered structure comprising w a set of neurons. Each layer contains some number of identical neurons, where each neuron is connected to all neurons in the following layer. There are three types of layers which are input layer, one or more hidden layers of neurons and output layer. A three-layer neural network is called *Shallow Neural Network*, whereas a network with more than three layers is called *Deep Neural Network*. The MLP with a single hidden layer is shown in Figure 3. The input patterns are provided to the input layer. The output signals, in which input patterns may map, are provided to the output layer. The layers between these two layers are hidden layers where the summed input weights are passed through a node's activation function. It is hypothesised that the hidden layers infer latent features in the input data that have predictability towards the outputs, which describes the *feature extraction process*.

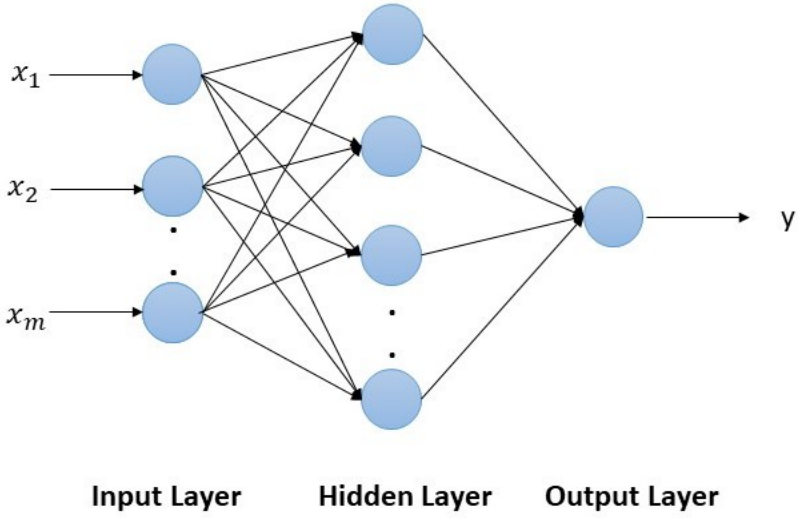

Figure 3: The figure illustrates the structure of simple multilayer perceptron network (MLP), comprising multiple layers of connected neurons, which are the input layer, one hidden layer and the output layer.

In practice, each neuron in a specific layer is linked to all neurons in the subsequent layer, this interconnection between the i^{th} and j^{th} neurons are characterised by the weight coefficient w_{ij}. Consider an MLP with a set of neurons in a layer A, a set of neurons in the following layer B and x_j presenting the output of neuron j, each neuron is computed as in equation 2. More generally the equation can be written as follows:

$$a_i = b_i + \sum_{j \in A} w_{ij} x_j, \qquad \forall i : i \in B \tag{7}$$

where b_i represents the bias of perceptron i. To get the output of hidden perceptron i, the linear combination a_i is transformed by activation function by equation 3, more generally:

$$x_i = \Phi(a_i) \tag{8}$$

The output of the neuron flows to all the neurons in the following layer till the final outputs of the network would be produced. However, the output of MLP depends only on the current input and not on the past or future inputs; thus, it is not suitable for sequential data.

.2.3 Recurrent Neural Networks

Traditional feedforward neural networks are confined to looking at individual instances rather than analysing sequential inputs. Sequential data is common various domains which have

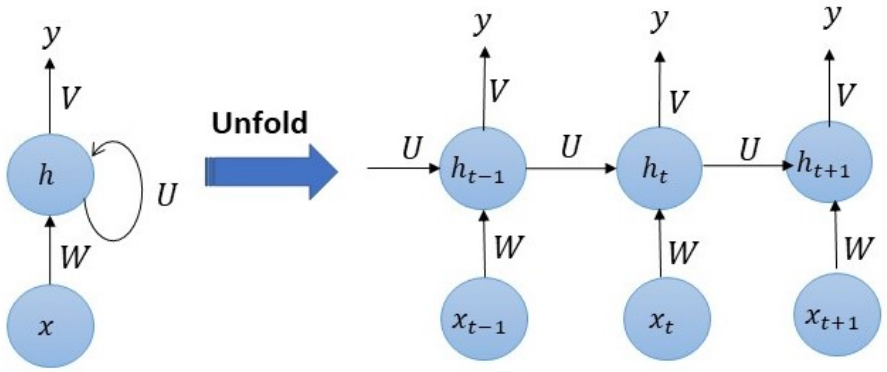

Figure 4: The Figure shows unrolling an RNN over sequential data over time which shows weight sharing across time steps. RNN has three types of layers: the input layer x, the hidden layer h, and the output layer y. If we unfold this loop, the standard RNN can be considered as copying the same structure multiple times, and the state h of each copy is taken as an input to its successor.

time-dependent individual instances, such as NLP, speech recognition and computational biology. NNs address each instance independently thus the advantage that can be taken by exploiting this sequential information is lost. One solution to attribute sequential dependency is a window-based method. It concatenates a fixed number of successive data instances together and process them as one data point, similar to moving a fixed size sliding window over stream of data. This method was applied in [113] for time sequence prediction and in [206] for acoustic modelling. However, it depends on the factor of choosing the optimal window size, where a small window size may not capture the longer dependencies, while a larger window size may add unnecessary noise. Particularly, a window-based approach may not scale when there are long term dependencies in data ranging over hundreds of time steps [113].

Moreover, considering these two sentences, *'I went to Roma in 2019'* and *'In 2019, I went to Roma'*, they have the same meaning but the details are in different positions of the sequence. Feeding these two sentences into a neural network for a prediction task, the model will assign different weights to 'to Roma' at each moment in time. This is because, for each input feature, the neural network has different parameters, thus the network learn all the rules of the language independently. Yet, Recurrent Neural Network (RNN) shares the same weights across multiple time steps and, thus the same weight will be assigned to 'to Roma'. Parameter sharing allows a model to be extended to examples of different lengths and to be generalized across examples. It is especially important when the the same part of information appears at multi-position within an input sequence [123].

RNNs are developed [183] which are well-known to work well for learning tasks where the input data is sequential. It processes the input sequence one instance at a time and preserves a hidden state vector which acts as a memory for past information. Alternatively, the state of a hidden neuron at time step t is a function of all inputs from previous time steps. Therefore, the recurrent connection from the end of the hidden layer to the beginning can be viewed as creating a kind of 'memory'. This allows them to utilize both current input and past information while making future predictions. The concept of memory is useful with sequential data, giving an example from NLP, *'I had cleaned my room'* has different meaning from *'I had my room cleaned'*. Therefore, it is important to understand the context of each word by looking to the words before or after it. Critically, the RNN framework does not require a limited fixed length on the prior context; the context encoded in the previous hidden layer contains information extending back to the beginning of the sequence.

Figure 6 illustrates the structure of the basic RNN model where on the right side of the figure, it shows the unfolded version of the RNN, which can be seen as a deep feed-forward neural network with the number of layers equivalent to the number of time steps in the input sequence and with shared weight matrices W,U and V between layers. Given a sequence of inputs $(x_1,\ldots,x_T)$, the model sees at each t time step a current sequence element x_t and the hidden state vector from previous time step h_{t-1}. Therefore, the hidden state is updated to h_t as follow:

$$h_t = tanh(Wx_t + Uh_{t-1} + b_h), \tag{9}$$

where W accounts for a weight matrix between input and the hidden layer, and U is a weight matrix connecting the hidden layer to itself at the previous time step t. In this way the current output h_t depends on all the previous inputs $x_t{}'$ (for $'t \leqslant t$).

The output y_t is computed as a function of hidden state as follows:

$$y_t = \Phi(Vh_t + b_y) \tag{10}$$

where Φ is an activation function, V is the output weight matrix and b_y is the bias.

Although this approach manages to achieve relatively good accuracy on many problems involving temporal data, it suffers from *vanishing and exploding gradient* problems. These problems occur when the input has long-range dependencies. During training of a deep net-

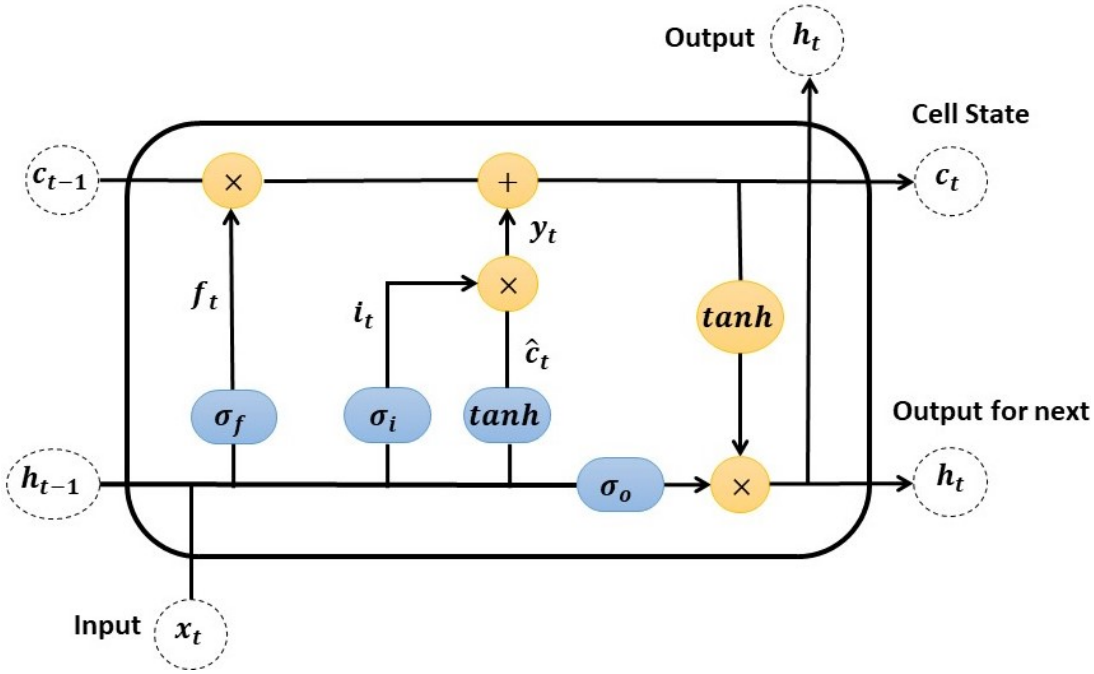

Figure 5: The figure shows the architecture of a recurrent cell in a Long Short-Term Memory Network (LSTM). + and x circles depict linear operations, while σ_f, σ_u and σ_o are the sigmoids used in the forget, update and output gates respectively [38].

work, the gradients are being propagated back in time all the way to the initial layer and the gradients are calculated in the deeper layers by continuous matrix multiplications. When the values are so small, they reduce proportionally till they vanish and, thus, a gradient not able to have a significant impact on the parameters that need to be adjusted. Alternatively, if the values are very large, they eventually explode and, in turn, an unstable network [135]. To overcome this problem, LSTM networks have been introduced and they are proven to be very useful in learning long-term dependencies compared to standard RNNs. LSTM networks have become the most popular variant of RNN, the next section describes in details its structure.

.2.4 Long Short-Term Memory

Long Short-Term Memory Networks (LSTMs) are special kind of RNNs and were proposed in 1997 by Sepp Hochrieiter and Jürgen Schmidhuber. They can learn long-term dependencies. They avoid vanishing and exploding gradient problems by introducing an adaptive gating mechanism and an explicitly defined memory cell (also called cell state), which preserves information across multiple time steps. Each neuron has a memory cell and three gates: in-

put, output and forget. These gates regulate the information flow into and out of the memory cell. The gate mechanism is based on sigmoidal activation functions, which output values in the range of 0 and 1, which refers to whether the corresponding entry can go through (1) or not (0).

LSTMs come in many forms, but all of them have some form of input, forget and output gates. The basic architecture of the LSTM unit, which is used in this work, is depicted in Figure 5.

Forget Gate

The forget gate f_t decides what the existing information is forgotten from the cell state. Alternatively, how much information on the previous memory cell c_{t-1} should be remembered. For each entry in the cell state c_{t-1}, h_{t-1} and x_t are considered then a value between 0 and 1 is assigned by pushing the output of the forget gate through the sigmoid function, which refers to whether the corresponding entry is kept (1) or removed (0).

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f), \tag{11}$$

where W_f is the weight and b_f is the bias.

Input gate

The input gate i_t is to decide what the new information stores into the new memory cell c_t. This occurs in two steps using two parallel neural network layers. The first layer determines which values should be updated in the state. Previous internal state h_{t-1} and the current input of x_t are used to calculate the input gate i_t as follows:

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i). \tag{12}$$

This layer is passed through sigmoid function where the output will be close to 0 for the values that the gate decides to leave unchanged, and the output will be close to 1 for the values that the gate decides to change. The second step produces new candidate values $\widetilde{c_t}$ that might be updated in the state with the same two input nodes.

$$\widetilde{c_t} = \tanh(W_c.[h_{t-1}, x_t] + b_c) \tag{13}$$

Finally, the old memory cell c_{t-1} requires to be updated into the new cell state c_t. This happens based on the multiplication of i_t and $\widetilde{c_t}$, which is the new candidate values scaled by how much it is decided to update each state value, added to the filtered previous internal state c_{t-1} by the forget gate.

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c_t} \tag{14}$$

Here c_t is proceed as input for the next time step.

Output Gate

The output gate o_t regulates which part of the memory cell c_t should flow into the hidden state h_t using sigmoid function to the previous hidden state and current input as follows.

$$o_t = \tanh\left(W_o.[h_{t-1}, x_t] + b_o\right) \tag{15}$$

The next hidden state h_t is then calculated by passing the updated cell state (c_t) through an elementwise tanh. This is then multiplied by the output gate o_t with values in the range of 0 to 1 to decide which element should be considered in the hidden state h_t.

$$h_t = o_t * \tanh(c_t) \tag{16}$$

Generally, the last hidden state h_T of a sequence of length T represents the entire sequence.

.2.5 Bidirectional Long Short-Term Memory Network

Conventional LSTMs only consider the previous context of data for training as it processes sequences in temporal order. Since simply looking at previous context may not be sufficient to understand the context, future context is also important to explore. Therefore, Bidirectional RNNs were introduced in 1997 by Schusterand Paliwa [287]. Recently, Bidirectional LSTMs (Bi-LSTMs) are applied more commonly in the literature, the standard LSTM networks are extended by adding another layer where the hidden connections are streamed in reverse temporal order.

The basic structure of Bi-LSTM is shown in Figure 17, where there is two LSTMs for each forward and backward sequence, and both are connected to the output layer. Therefore,

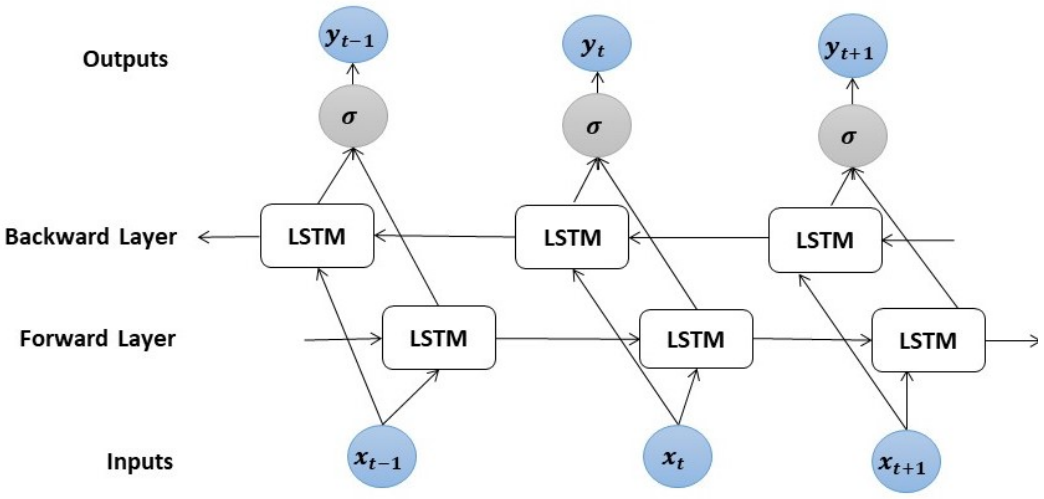

Figure 6: The Figure shows the basic structure of the Bi-LSTM network. The LSTM nets at the bottom indicate the forward feature. The above nets are used for backward. Both networks are concatenated and connected to a common activation layer σ to produce outputs..

the model can capture both the previous and the future time steps of the input sequence.

For each input sequence $(x_1, x_2, ..., x_T)$, the hidden states sequence $(\overrightarrow{h_1}, \overrightarrow{h_2}, ..., \overrightarrow{h_T})$ of the output of the forward LSTM and the hidden states sequence $(\overleftarrow{h_1}, \overleftarrow{h_2}, ..., \overleftarrow{h_T})$ of the backward LSTM are concatenated at each time step as follows.

$$h_t = [\overrightarrow{h_t} \oplus \overleftarrow{h_t}], \qquad \in \mathbb{R}^{2u} \tag{17}$$

Where u is the number of hidden unit in each unidirectional LSTM. The entire hidden states sequence are formed as follows:

$$(h_1, h_2, ..., h_T), \qquad \in \mathbb{R}^{T \times 2u} \tag{18}$$

Generally, the last hidden state h_T of the sequence of length T represents the entire sequence.

.2.6 Network Training

Network training is the fundamental part of machine learning procedure. Before training ANNs, the *initialisation phase* occurs. Several approaches are available to initialise any type of ANNs, the most common one is to set the weights W to a small negative or positive random values. Besides, the bias $b(l)$ for each layer l is set to a number, which is not 0. The

performance of this initialisation process apparently will not result in a good prediction, thus the training process is essential. After this initialisation step, the actual training starts. The remaining of this section presents the basic training process of any form of ANNs.

Loss Function

A training step comprises the *forward pass*, where the training samples are passed forward through the network (the processes that was described in details in Section .2). The network predicts the probability for an independent label $\hat{y}$ for each instance where $\hat{y} = f(x, \theta)$. The goal of the training process is to learn the network's parameters for each layer as such making each training example closest to the true value (y). The distance between the network's output ($\hat{y}$) and the true output (y) is measured, this is called *cost function* or *loss function*. The Mean Squared Error (MSE) and the Cross-Entropy Loss functions are the most common cost functions used in literature. In classification tasks, the loss function we want to minimise is usually cross-entropy. This work uses the binary cross-entropy cost function which is calculated with:

$$J(\hat{y}, y) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\ \log(l - (\hat{y}_i)] \tag{19}$$

where y is the ground-truth label. Getting low loss means that the network's predictions are very close to the true values.

Optimisation Algorithm

The goal of training any ANNs is to minimise the cost function J. This is achieved by iteratively updating the network's parameters θ. Therefore, learning is an optimisation problem with the following criterion:

$$\min_{\theta} J(f(x, \theta), y) \tag{20}$$

To minimise the difference between the network's output and the target output, the optimal weights for the neurons are discovered by performing *a backward pass*, moving back from the network's prediction to the neurons that generated that prediction. Each time the network processes the whole set of data (both a forward pass and a backward pass), it is called an *epoch*. This process is an iterative process, which continues to reduce the error by every epoch until an acceptable level of errors is obtained. This entire process is called *back-*

propagation process or *backward pass*. It identifies how the network's performance changes for each parameter in the network (θ). Alternatively, it tracks the derivatives of the activation functions in each successive neuron, to find weights that bring the loss function to a minimum, which will generate the best prediction, with a given step (this step is called *learning rate*). The mathematical process for this approach is called *gradient descent*.

The gradient descent method is an iterative *optimisation algorithm*, which updates the weights of the model in the opposite direction of the gradient to minimise the loss function. In the simplest version, it has the following form:

$$\theta := \theta - \alpha \frac{\partial J(\theta)}{\partial \theta} \tag{21}$$

where α represents the learning rate.

ANNs often have a large number of parameters to optimise. The gradient is therefore not efficient regarding the computation time and update speed. This means that it must compute millions of forward propagations, before it can compute one backpropagation step for updating the weights. For one update, it processes the entire training set, which can have millions of examples. Therefore, one possible solution is to split the training set into *mini-batches* (its size can be set by batch size hyperparameter) and the gradient is calculated on each mini-batch, this Gradient Descent is called *Stochastic Gradient Descent (SGD)*.

In the recent years, several new optimizers have been proposed to tackle complex training scenarios where gradient descent methods behave poorly. One of the most widely used and practical optimizers for training deep learning models is Adaptive Moment Estimation (Adam) [152]. It is a stochastic optimisation algorithm that calculates adaptive learning rate for different parameters from estimates of the first and second moments of the gradients. Thus, Adam adapts its learning rate α during training and optimizes it for every parameter. Practically, Adam optimiser outperforms SGD in many complex tasks [152].

The selection of the learning rate α is important because choosing a too small step value may result into long time training of the algorithm, and it could stuck in a local minima. Alternatively, if α is too big, it may jump the valley with the best solution. Therefore, learning rate is one of the main hyperparameter of a neural network and should be carefully selected. The next section discusses the network hyperparameters, including learning rate and optimiser algorithm.

.3 Hyperparameter and Model Selection

To test the generalisation of the model, the training set is further split in a validation set. Hence, there is three datasets: a training set, a validation set and a test set. The validation set is used to measure the generalisation error and is not used for the training. Since the model is chosen by maximising the performance on validation set, the predicted performance on the validation set has a bias. Therefore, the performance should be measured on the test set, and this is a good approximation of the performance on unseen data.

One of the best technique used to test the effectiveness of the algorithm is *cross validation methods* (CV). It is a re-sampling procedure used especially with limited dataset where a portion of the given data is kept for training the model and another for testing its performance. Several CV techniques are used for splitting the dataset including the *leave-one-out* (LOO) and the *k-fold* methods. In the LOO procedure, all the samples are trained except for one sample and the prediction is performed in the out sample and, thus, the average error is computed and is used to evaluate the overall model. In addition, one subject can have several samples and, hence, the LOO can also be implemented in a leave-one-subject-out manner, where all samples from a specific subject are excluded each time. In the k-fold procedure, k partitions of the dataset are split, where one partition is kept each time for testing and the others used for training the model. This manner is repetitive for k times.

Machine learning algorithms automatically adjust and learn their internal parameters based on data. However, there is a subset of parameters that is not learned and that have to be configured by the scientists. Such parameters are often referred to as *'hyperparameters'*. With these hyperparameters, the algorithms' behaviours can be changed and the capacity of the models can be regulated. The hyperparameters for ANNs are numerous, but the most important ones are: epochs, batch size, number of neurons, number of hidden layers, activation functions, optimisation algorithms, and loss functions. Selecting an optimal set of hyperparameters is crucial because it influences the performance of the model substantially.

Tuning a hyperparameter scheme to find optimum topology of the model is time-consuming and tedious. Thus, *hyperparameter optimisation* methods are termed to find the best hyperparameter combination that gives the best performance on a hold-out validation set. Grid Search, Random Search, and automated hyperparameter optimisation methods have been

commonly employed. Grid Search and Random Search create a grid of hyperparameter values. Specifically, in Grid Search, the value combinations will be exhaustively explored to find the hyperparameter values combination that gives the best accuracy values. This method is a costly approach, assuming having n hyperparameters and each hyperparameter has two values, then the total number of configurations is 2^n, making this method very inefficient. However, Random Search navigates the grid of hyperparameters randomly, which repeatedly selects random combinations from the grid until the certain number of iterations is achieved. Although it manages to give good hyperparameters combination, it is hard to be certain that it is the best combination [36]. In contrast, automatic hyperparameter tuning forms knowledge about the relation between the hyperparameter settings and model performance to make a smarter choice for the next parameter settings. It uses different techniques such as Bayesian optimisation that conducts a guided search for the best hyperparameters. Bayesian optimisation applies a Gaussian process to model the surrogate and typically optimises the Expected Improvement, which is the expected probability that new trials will improve upon the current best observation. Bayesian optimisation can yield better hyperparameter combinations than Grid Search and Random Search algorithms [115, 225].

.4 Natural Language Processing: Text Representation

NLP is a sub-field of Artificial Intelligence (AI) for computational techniques which helps computers to understand, process and manipulate human written language such as sentiment analysis. It allows computers to execute various natural language tasks including part-of-speech tagging (POS) [334], syntactic parsing [58], named entity recognition [168], semantic role labelling [360] and machine translation [367]. NLP techniques depend mostly on machine learning to derive meaning from human languages, which help to understand *what people say*. For machine to understand the text, word representation is essential to represent words as feature vectors as real-valued vectors. This process is called *word representation or feature extraction*. This section illustrates the different types of word representation methods including static representation and dynamic representation.

.4.1 Static Representation

Representing the words in the static form can be divided into two different techniques: local representation and distributional representation. Local representation is to represent the words into sparse and high-dimensional vectors. This type of representation suffers from high dimensionality and data sparsity, especially with a large size of vocabulary. Conversely, distributed representation can address such a problem by representing the words with dense and low-dimensional vectors, which have been trained on large data with the goal of transferring it to other NLP problems. This section describes briefly these different techniques for static word representation.

Local Representations

In the early age of NLP, local representation such as *one-hot word representation* was introduced. Each word is represented to a sparse discrete vector which is all zero values except the index of the specific word which is marked with a one. This representation, however, is meaningless because it lacks the relationship of words with each other. For example, *'apple'* and *'banana'* should be near to each other in the semantic space due to their similarity in contexts. If the word *'apple'* is changed by *'banana'*, regardless whether it has seen the sentence *'This banana is fresh'* by virtue of banana occurring in the same context of apple, the sentence's probability should be estimated. However, one-hot representation is unsuccessful to do this because all the word vectors are orthogonal to each other which means that the cosine similarity of any two distinct word vectors is 0. Moreover, if there is a dictionary of n-words, this requires n-dimensional vector for each word, thereby making the training model on this representation infeasible.

For capturing the syntactic and semantic similarity between words, additional features of word representation are utilised, including morphology and part of speech. The intuition behind it is that the linguistic concept of distributional hypothesis states that *'words occurring in similar contexts seem to have similar meanings'* [27]. Word is represented by a vector whose values are the count of words that appear in context which may semantically capture the similarity between words. With large corpus, for example, we can observe that the contexts of *banana* is closer to the contexts of *apple*. If V_w is the word vocabulary and V_s is the predefined context word vocabulary. A metric W is created to quantify the relation of words

with their contexts.

$$W_{ij} = count(w_i, c_j), W \in \mathbb{R}^{|V_w| \times |V_c|} \tag{22}$$

where $count(w_i, c_j)$ is the the number of times a w_i appears in the context of c_j.

However, the co-occurrence is not the only measurement that captures the correlation of words, since high weights could be assigned to word-context pairs containing common contexts. Therefore, term frequency–inverse data frequency model (TF-IDF) was proposed to solve such a problem by applying weighting factors [138]. Unlike co-occurrence vectors, TF-IDF considers not just the occurrence of a word in a single sentence (context) but in the entire corpus. The weights of word-context pairs are decreased in proportion to their frequency in the corpus. Alternatively, weighs down the frequent words (less significant words) while scaling up the rare ones (more significant words).

Given a document collection D, a word w and a document $d \in D$, TF-IDF is claculated as follow:

$$w_d = f_{w,d} * log(|D|/f_{w,D}) \tag{23}$$

where $f_{w,d}$ calculates the word count of w in d, $|D|$ is the corpus size and $f_{w,D}$ is the number of documents in which w occurs in D [274].

The main problem with this type of representations is the high dimensionality of vectors. If there is a dictionary of n-words, n-dimensional vector for each word is required, thereby making the training model on this representation infeasible. Furthermore, it can suffer from data sparsity caused by having vast vocabularies and a given word would be represented by a large vector comprising mostly zero values.

Distributed Representation

The traditional word representation models mentioned above are easy to develop. However, the semantic of elements of larger granularity, such as phrases and sentences, is difficult to capture. To solve this problem, the expressive power of neural network are employed a neural network based approach using contextualised information [202, 233]. The distributed representations are real-valued vectors to flexibly represent semantics of natural language. Can distributed word representations used to improve language modelling? Instead of hand-crafted word representation, different ways of learning representations are considered directly

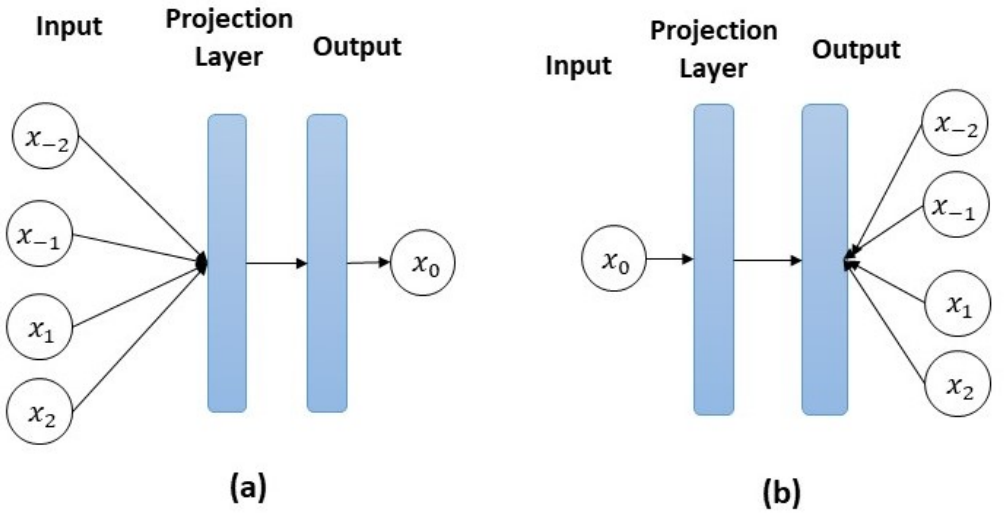

Figure 7: The Figure shows the neural network architecture of two different word2vec models: (a) Bag-of-Words model (CBOW) and (b) Continuous Skip-gram model. archite. In the CBOW architecture, the model predicts a target word given a set of surrounding context words. In contrast, the Skip-gram architecture tries to predict a set of context words given a target word.

from the data to solve some task at hand. Such representations are learned implicitly with the language modelling task in a neural network architecture. This results in a low, dense, real-valued distributed word representation which is called *distributed representation* or *word embedding*.

In details, words can be initialised with random vectors in a lookup table. These initialised representations can be updated via backpropagation which results in a dense, real-valued distributed word representation. While having distributed representations for each word in the vocabulary, a probability of sequence of words is computed as a function of the word representations. This induces another question, what type of functions can be used to assign word representations to probabilities? Word representations are learned by training neural network as parameters of the model. Backpropagation is then used to learn the function from word vectors to probabilities and the word vectors themselves that minimises the difference between next-word prediction features and target values.

The most common learning techniques for unsupervised learning to learn word embeddings are *Continuous Bag-of-Words model* (CBOW) and *Continuous Skip-gram model* [202]. Figure 7 illustrates these two methods. Skip-gram is the conditional probability for generating n surrounding words (context) given a word, where n is the context window size.

However, CBOW maximises the probability of a word given its context (surrounding words). It predicts the central target word based on the context words proceeding and following it in the text sequence. Generally, these two approaches use shallow neural network, an MLP network with one hidden layer, to learn word representation. In practice, Word2Vec is the most popular implementation for both CBOW and skip-gram models.

Recently, Wikipedia2Vec is presented: an optimised tool for learning embeddings of words and entities from Wikipedia [350]. It learns embeddings of words and entities simultaneously and places similar words and entities close to one another in a continuous vector space. This is achieved by extending Word2Vec's skip-gram model, which learns to predict the context word for a given target word, with two sub-modules: the link graph module and the anchor context module. The link graph module learns to estimate neighbouring entities given an entity in the link graph of Wikipedia entities. The anchor context module learns to predict neighbouring words given an entity using a link that points to the entity and its neighbouring words. Wikipedia2Vec has been applied in different important fields including text classification [351] and paraphrase detection [94].

The main drawback of distributed representation introduced above is that it learns embeddings by looking at the occurrences of nearby words which is limited on the local context of a given word. For instance, CBOW and skip-gram incorporate 5 to 10 context words in practice that influence a word embedding. Therefore, they lack to project all global connections of the word. Besides, they suffer from capturing the polysemous of words in different context where each word is represented by a single prototype vector that does not change with its context [237].

.4.2 Dynamic Word Embedding

A more effective way to address the polysemy problem is using *dynamic embeddings* or *contextualised embeddings*. Essentially, the static word embedding models generate the same embedding for the same word in different contexts. Instead of learning a fixed number of contexts per word, dynamic word embedding captures word semantics in different contexts to address the issue of polysemous and the context-dependent word semantics [237].

Bidirectional Encoder Representations from Transformers (BERT) is the most common example of dynamic word embedding which has recently improved the state-of-the-art in

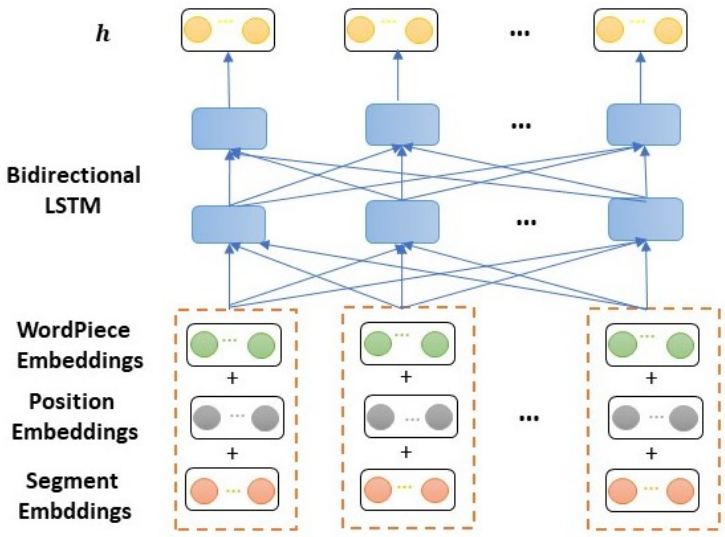

Figure 8: The figure shows neural network architecture of BERT. The input word piece, position and segment embeddings are summed [337].

word embeddings [90]. BERT has a transformer encoder that captures both left and right contexts through reading the entire sequence of words at once. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word). It is pre-trained with two objectives—masked language model (MLM) and next sentence prediction (NSP) pretraining task. The MLM task is for predicting randomly masked words given its context whereas the NSP task is for capturing the relationship between sentences by predicting if a sentence B is followed by sentence A.

BERT utilises WordPiece embeddings [347] for tokenisation instead of word ones. The input is tokenised into word pieces so that each word piece is an element of the dictionary, which in effect, splits token such as 'playing' to 'play' and '##ing'. This is because the tokens may not be contained in the pre-trained vocabulary of BERT as the BERT model has a specific fixed vocabulary. Moreover, this kind of WordPiece tokeniser has a certain way of handling Out-of-Vocab words (OOV). In addition, special tokens are inserted into the start of the sequence ([CLS]), which contains the special classification embedding, and the end of each sequence ([SEP]) for separating segments or denoting the end of the sequence.

The embedding layer is used to get the vector representation for each word in a sequence. It comprises word piece embedding, the segment embedding, and the position embedding. Specifically, word piece embedding is obtained through the corresponding embedding matrices. Position embedding is used to capture the order information of the sequence which

is ignored during the self-attention process. Segment embedding is used to distinguish between two different sequences of the input. The word, segment and position embeddings are summed up to create the final input embeddings for a sequence (Figure 8).

Two model sizes for BERT have been developed, including BERT-Base, which has 12 encoder layers, each having a hidden size of 768 and 12 attention heads (110M parameters), and BERT-Large, which has 24 encoder layers, each having a hidden size of 1024 and 16 attention heads (330M parameters). For each word, the word representation can be extracted from any of the encoder layers. BERT models have been trained on general domain corpora, such as English Wikipedia and Books Corpus. They have released English language and multilingual versions. The latter supports 104 languages in a single model, which has a large shared vocabulary, including Italian, German, Arabic and Japanese. The vocabulary size of multilingual model is 119,547 WordPiece tokens for all of 104 languages compared to 28,996 tokens for English-only model.

In contrast to fixed word embeddings, the text representations are learned based on sequential context than word concurrency. Moreover, it further learns sentence-level information by sentence-level encoders than only extract local semantic information of individual words.

.5 Computational Paralinguistics: Speech Representation

It is a fact that a human speech contains both the basic verbal message along with paralinguistic information. Paralinguistics mean 'alongside linguistic'. It is the study of non-verbal properties of speech that is based on the qualities of your voice and the way you vocalise [283]. In signal processing and machine learning, it is being a mainstream subject and one of the hot topics within Social Signal Processing [325]. It is a non-linguistic function that is embedded in the verbal acoustic message and can be consciously controlled by the speaker such as intentions, attitudes, emphasis and speaking styles.

Conventional NLP approaches concentrate on linguistic content analysis and word representation. With the wider availability of recorded speech, analysing the states and traits of speakers are increased [284]. Compare with computational linguistics, computational paralinguistics analyses *how people say* rather than *what people say*. The ability to analyse

paralinguistic features has induced progress in a multitude of speech processing tasks, such as speaker verification [132,173,323], age identification [201], personality recognition [284], speech emotion recognition [121, 216, 313], conversation analysis [170, 172], medical diagnostics [41, 82, 194, 281] and depression detection [76].

This section discusses the extracting features from paralinguistic information using Mel frequency cepstral Coefficients (MFCCs), a feature widely used in speech processing tasks.

.5.1 Mel Frequency Cepstral Coefficients

Feature extraction aims to transform the speech signal into a parametric representation of reduced dimensionality, providing a good discriminability between classes to be identified. Mel Frequency Cepstral Coefficients (MFCCs) are a unit of representation related to the human auditory system and are very distinguishing for speech processing tasks. Several studies (e.g. phone recognition [85], speaker identification [257] and claim identification [181]) have extracted the coefficients of MFCCs to identify the paralinguistic features,

It is short-term spectral features that are widely applied in the area of audio and speech processing. Human ears have different bandwidths with different frequencies, and the MFCCs are based on the difference of frequencies that the human ear can distinguish. To detect the patterns of speech and audio, filters are placed linearly at low frequencies (below 1000 Hz) and are placed logarithmically at high frequencies (above 1000 Hz). It is useful as the voice depends on the shape of vocal tract, including tongue and teeth. Representation of short-time power spectrum of sound is essentially a representation of the vocal tract. The process of MFCCs comprises several steps, performed in order, which are described below.

Frame Blocking

Speech signal varies over time. To gain stable acoustic characteristics, speech signal should be processed over adequately short period known as *frames*. Speech analysis studies short segments that capture enough information in which the features inside the frames should remain relatively stationary. Therefore, the speech signal is divided into a sequence of N frames, where each frame can be analysed independently and are represented by a single feature vector, with next frames separated by M samples ($M < N$), and the adjacent frames are overlapped by $N - M$ samples. The typical length of the frame size is about $20 - 40ms$

with overlap to the frames by 15*ms* [139]. It is necessary to choose a reasonable frame size in which it can provide good spectral resolution. If the frame size is too large, it will not capture the local spectral properties, whereas if the frame size is too small, there will not be enough samples to get a reliable spectral estimate. The standard frame size for the speech analysis during speech recognition is set to 25*ms*, as within this short period, the speech signal's properties are fairly stationary.

Windowing Function

To smooth the signal and minimise the disruptions at the start and at the end of the frames, windowing function is applied over the frames where the frame and window function is being multiplied. Hanning and Hamming are commonly applied windowing functions to enhance the continuity between each frame and its adjacent frame, after the signal is segmented into frames. Practically, the spectral distortion is reduced using a window that tapers the speech sample to zero at both the beginning and the end of each frame. This windowing process $W[n]$ is applied to the input speech frame $S[n]$ as follows:

$$X[n] = S[n] * W[n], \qquad where \qquad 0 \leq n \leq N-1 \tag{24}$$

where N stands for the quantity of samples within every frame, and X[n] represents the output signal after multiplying the input signal $S[n]$ and the window function $W[n]$.

Spectral Estimation (Discrete Fourier Transform)

For the spectral analysis, discrete Fourier transform (DFT) converts each windowed frame from the time domain to the frequency domain [139]. The fast Fourier Transform (FFT) is a computationally efficient algorithm for implementing the DFT in which each frame has a given set of N samples that are converted into frequency domain as follows:

$$X[k] = \sum_{n=0}^{N-1} X[n] e^{(-j\frac{2\pi}{N}kn)}, \qquad where \qquad k = 0, 1, ..., N-1 \tag{25}$$

Here $X[n]$ is the framed speech signal, and $X(k)$ are spectral coefficients where their values are comprise real and imaginary values. These values result in complex numbers, however, only the absolute values (frequency magnitudes) are considered to perform further process.

By calculating DFT, we can obtain the magnitude spectrum for the k^{th} frequency component in the original signal.

Mel Spectrum

Conventional spectral analysis results in a varying frequencies and the signal is linearly spaced frequencies. This means that the resolution (variance between adjacent frequencies) is the same at all frequencies. However, the human auditory system has unevenly resolution. For instance, human may be able to identify small changes in frequency in low frequency audios, and big changes in frequency in high frequency audios. Thus, there is a need to adapt some new techniques on the spectral output from the preceding step (spectral estimation) to replicate the human auditory response. Hence, Fourier transform signal is passed through triangular band-pass filters known as *Mel-filter bank* to wrap the output frequencies to the Mel-scale [304]. A Mel is a unit of measure that mimics the human perception of sound. A popular formula to convert from frequency scale to mel-scale $m(f)$ is given as follows:

$$mel(f) = 2595 * log_{10}(\frac{1+f}{700}) \tag{26}$$

where $mel(f)$ is the frequency in mels, and f is the normal frequency in Hz.

The set of filters are a set of triangular windows that are distant uniformly with overlapping on the Mel-frequency axis. These filters are roughly a linear frequency spacing below 1kHz and turns logarithmically afterward. The power spectrum of each frame is the input of the mel-filter-bank and each filter output is the sum of its filtered spectral components, known as *mel-spectrum*. This can be described in the given formula.

$$Y[k] = \sum_{k} |X[k]| H_i(Mel(f)) \tag{27}$$

Where X[k] is the DFT at frequency k, H_i is the mel-spaced-filter-bank. An example of the mel-filter-bank is illustrated in Figure 9.

The human response to the audio level is logarithmic. Use of logarithm squeezes vigorous amount of values and sorts frequency estimates less sensitive to slight variations in the input such as power variation because of microphone distortions. Logarithm of the square

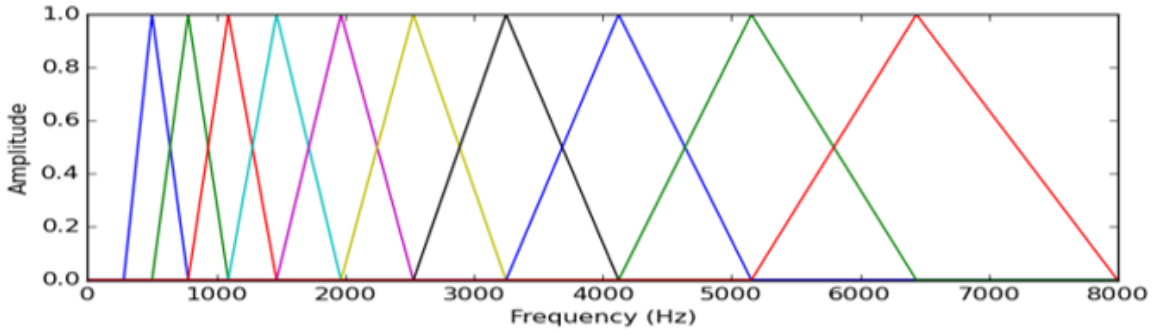

Figure 9: The Figure shows a filter bank of 10 filters used in MFCC

magnitude of the mel-filter bank output $Y[k]$, is defined as follow:

$$E[m] = log|Y[k]|^2 \tag{28}$$

Discrete Cosine Transform

Discrete cosine transform (DCT) is computed to the log-spectral-energy vector $E[m]$ to transform the log Mel spectrum from the frequency domain to the time domain [139]. This results in several Mel-scale cepstral coefficients, which are the standard MFCC. The mathematical formula for calculating the cepstral coefficients is as follows.

$$C[i] = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} E[m] cos(\frac{\pi i}{M}(m - \frac{1}{2})) \tag{29}$$

where M is the total number of cepstral coefficients extracted from each frame. Typically, the first 12 coefficients are considered in the literature, since these 12 values depict the information about the vocal tract only.

Energy and Deltas

The first 12 cepstral coefficients of MFCCs contain the most salient information needed for speech recognition. To achieve higher accuracy, energy from each frame is computed, which is considered as the 13^{th} feature. The energy in each frame can be calculated as the summation

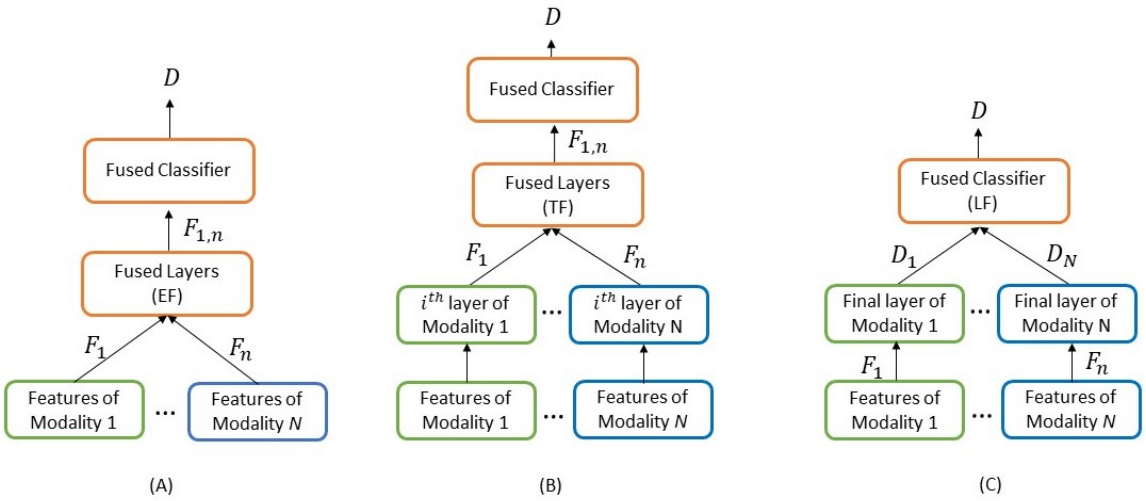

Figure 10: The Figure shows different fusion strategies for multiple modalities: (A) Early Fusion (EF) where all the features from different modalities F_1 to F_n are fused using an EF unit to obtain single feature vector F_1,n which is passed as input of the model to get the final result D, (B) Intermediate Fusion (TF) where the intermediate features for each channel obtained from layer i of NN are fused using a TF unit, and then the combined feature vector is passed to the model for further analysis, and Late Fusion (LF) where the individual decision from each channel D_1 to D_n are fused using an LF unit to obtain a final decision D.

of the power of the samples over time, as follows.

$$Energy = \sum_t x^2[t] \tag{30}$$

In addition, the speech signal changes from frame to frame. To represent dynamic nature of the audio, the change in the cepstral features over time is computed by adding a *delta* (velocity feature) and a *double delta* (acceleration feature) for each of the 13 features. To calculate deltas, the distance between frames is defined by the following formula:

$$d(t) = \frac{C(t+1) - C(t-1)}{2} \tag{31}$$

where $d(t)$ represents the delta values, at time t, for a specific cepstral value $C(t)$. This results in 13 delta features that show the variation between frames in the respective cepstral feature. Furthermore, the 13 double delta features show the variation between frames in the respective delta features. At the end, the entire signal is transformed to a sequence of 39 cepstral vectors.

.6 Multimodal Representation

The integration of multiple sources, modals or modes to perform an analysis task is referred to as *multimodal fusion*. Sources, modals or modes are, essentially, channels of information. The data from multiple modalities are semantically correlated, and sometimes they provide complementary information to each other. Therefore, it reflects patterns that may not be visible when individual modality infer a decision in isolation from other modalities. The key idea of utilising multimodal fusion is its ability to fill missing modality given the observed ones which can increase the accuracy of the overall decision-making process. For example, fusion of speech features along with manual transcription features for depression detection enhances the result compared to what would be obtained using a single medium [210].

One recurrent question with multimodal fusion is where the fusion should be applied. The most widely used strategy is to fuse the information close to the data, which is known as *early fusion*. The other fusion approach is applied at the decision level, which is known as *late fusion*. Another fusion strategy is in between these fusions which is known as *intermediate fusion*. For neural networks, the fusion can be done at any level between the input and the output of the unimodal networks. Figure 10 shows different variants of the early, intermediate and late level fusion strategies. In this section, we will highlight these three fusion strategies.

.6.1 Early Fusion

Early Fusion (EF) is also called fusion in feature space. In early fusion, joint representation of input features from different modalities are formed. For n different modalities, a set of features F_1 to F_n are concatenated into a feature vector $F_{1,n}$ before being fed to the the model. The model,73 in the scope of this research, can be any type of neural network models that are used for the task at hand such as feed-forward network, RNN, LSTM or CNN. Typically, a preprocessing step is necessary on the concatenated features to share the same statistical properties such as normalisation techniques. An illustration of the EF strategy is provided in Figure 10-A. It shows an instance of the early fusion multimodal analysis task in which the extracted features are first fused using an EF unit, and then the combined feature vector is passed to the model for analysis.

This fusion strategy is advantageous in that it can utilise the relation between multiple

features from different modalities at an early stage, which helps in better task achievement. Moreover, it only involves one learning phase on the combined feature vector [299]. However, it is often difficult to combine features of different natures into a common homogeneous representation [348], for instance, two time sequences of different sampling rates or different lengths. One possibility is to flatten each representation to a one-dimensional vector before the concatenating process. This solution often is undesirable because flattening the data changes the structure of it [227].

.6.2 Late Fusion

Late fusion approach (LF) is also called decision level approach. Unimodal approaches produce local decisions from D_1 to D_n which are obtained based on individual unimodal features (F_1 to F_n). These local decisions are then combined using an LF unit to make a fused decision vector that is analysed further to obtain a final decision D about the problem [57]. An illustration of LF approach is shown in Figure 10-C.

The late fusion strategy has many advantages over early fusion. Unlike early fusion, late fusion focuses on the individual predictive strength of each modality. In addition, it can be performed to a broader set of learning problems because it does not suffer from fusing features in representation space. This is because the outputs of multiple unimodal classifiers are in the same form, such as class labels or class confidence measures. However, it fails in utilising the feature level correlation among modalities. In addition, different classifiers have their local decisions and, thus, the training process for them seems to be tedious and time-consuming.

Various techniques are used for late fusion approach. Regarding neural network with softmax layer, it produces a discrete probability distribution over the available classes. A comprehensive and comparative study of various combination rules such as sum, product, max, min, median and majority voting, was studied in [154]. They suggested that the sum rule is less exposed to the error of individual classifiers when estimating posterior class probability. It is the most straightforward approach, probably the most commonly applied technique for combined multiple classifiers. In this study, the sum rule was performed in which the outputs of multiple classifiers can be combined into one multimodal class prediction by taking the

class with the maximum index value [227]:

$$\hat{c} = arg \max_{c \in C} \sum_{i=1}^{n} p(c|S_i) \tag{32}$$

where C is the set of all possible classes, n is the number of modalities, and S is the decision vector extracted from modality i.

.6.3 Intermediate Fusion

Intermediate fusion (TF) is neither early nor late fusion levels, it is in between them. The architecture of intermediate fusion is built based on the deep neural network. Neural networks comprise hierarchy of layers that transform input data into a higher level representation through multiple layers. Therefore, the internal data representation (intermediate features) can be extracted from any layers of processing on each unimodal representation. The extracted learned representations of all modalities, F_1 to F_n, are concatenated and then fed to another model that learns to embed them in a new multimodal space that is better in optimizing their similarity. An illustration of the TF strategy is provided in Figure 10-B. It shows an instance of the intermediate fusion multimodal analysis task in which the intermediate features are fused using an TF unit and then the concatenated feature vector is passed to another model for further analysis [227].

This type of fusion can be beneficial over early fusion, since the unimodal processing layers can vary in a way matching the nature of each modality. Moreover, the concatenated representation can be fed into any number of layers, this provides intense processing procedure of multimodal representation. Therefore, the intermediate fusion can be seen more powerful than late fusion models.

.7 Conclusion

This study explored the importance of deep neural networks. It overviewed the basic structure of neural networks and the more advanced structures, including MLP, RNN, LSTM and Bi-LSTM. It also highlighted the different approaches of representing words and speech. The different techniques for combining different modes were also considered.

Bibliography

[1] Saeed Abdullah and Tanzeem Choudhury. Sensing technologies for monitoring serious mental illnesses. *IEEE MultiMedia*, 25(1):61–75, 2018.

[2] Aseel Addawood and Masooda Bashir. "what is your evidence?" a study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 1–11, 2016.

[3] Amber Afshan, Jinxi Guo, Soo Jin Park, Vijay Ravi, Jonathan Flint, and Abeer Alwan. Effectiveness of voice quality features in detecting depression. In *Interspeech*, pages 1676–1680, 2018.

[4] Rajesh K Aggarwal and Guojun Wu. Stock market manipulations. *The Journal of Business*, 79(4):1915–1953, 2006.

[5] T. Al Hanai, M.M. Ghassemi, and J.R. Glass. Detecting depression with audio/text sequence modeling of interviews. In *Proceedings of Interspeech*, pages 1716–1720, 2018.

[6] Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, volume 2522, pages 1716–1720, 2018.

[7] M. Al Jazaery and G. Guo. Video-based depression level analysis by encoding deep spatiotemporal features. *IEEE Transactions on Affective Computing (to appear)*, 2019.

[8] Mohamad Al Jazaery and Guodong Guo. Video-based depression level analysis by encoding deep spatiotemporal features. *IEEE Transactions on Affective Computing*, 2018.

[9] Mohammed Al-Mosaiwi and Tom Johnstone. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, 6(4):529–542, 2018.

[10] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear. Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Transactions on Affective Computing*, 9(4):478–490, 2018.

[11] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear. Head pose and movement analysis as an indicator of depression. In *Proceedings of the IEEE International Conference on Affective Computing and Intelligent Interaction*, pages 283–288, 2013.

[12] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Michael Breakspear, and Gordon Parker. Detecting depression: a comparison between spontaneous and read speech. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7547–7551. IEEE, 2013.

[13] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Tom Gedeon, Michael Breakspear, and Gordon Parker. A comparative study of different classifiers for detecting depression from spontaneous speech. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8022–8026. IEEE, 2013.

[14] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36, 2017.

[15] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554, 2019.

[16] Murray Alpert, Enrique R Pouget, and Raul R Silva. Reflections of depression in acoustic measures of the patient's speech. *Journal of affective disorders*, 66(1):59–69, 2001.

[17] Tim Althoff, Kevin Clark, and Jure Leskovec. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476, 2016.

[18] M Ananny. Checking in with the facebook fact-checking partnership. *Columbia Journalism Review*, pages 84–117, 2018.

[19] L. Andrade, J.J. Caraveo-Anduaga, P. Berglund, R.V. Bijl, R. De Graaf, W. Vollebergh, E. Dragomirecka, R. Kohn, M. Keller, R.C. Kessler, N. Kawakami, C. Kiliç, D. Offord, T. Bedirhan Ustun, and H.-U. Wittchen. The epidemiology of major depressive episodes: Results from the international consortium of psychiatric epidemiology (ICPE) surveys. *International Journal of Methods in Psychiatric Research*, 12(1):3–21, 2003.

[20] J. Arevalo, T. Solorio, M. Montes-y G/'omez, and F.A. González. Gated multimodal units for information fusion. arxiv:1702.01992, arXiv, 2017.

[21] American Psychiatric Association et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[22] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.

[23] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[24] R Michael Bagby, Andrew G Ryder, Deborah R Schuller, and Margarita B Marshall. The hamilton depression rating scale: has the gold standard become a lead weight? *American Journal of Psychiatry*, 161(12):2163–2177, 2004.

[25] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[26] Meital Balmas. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. *Communication research*, 41(3):430–454, 2014.

[27] CE Bazell. Studies in linguistic analysis. special volume of the philological society, vii, 205 pp., 5 plates. oxford: Basil blackwell, 1957. 70s. *Bulletin of the School of Oriental and African Studies*, 22(1):182–184, 1959.

[28] Aaron T Beck. *Depression: Clinical, experimental, and theoretical aspects*. University of Pennsylvania Press, 1967.

[29] Aaron T Beck. *Cognitive therapy and the emotional disorders*. Penguin, 1979.

[30] Aaron T Beck and Brad A Alford. *Depression: Causes and treatment*. University of Pennsylvania Press, 2009.

[31] Aaron T Beck, C Ward, M Mendelson, J Mock, and J Erbaugh. Beck depression inventory (bdi). *Arch Gen Psychiatry*, 4(6):561–571, 1961.

[32] A.T. Beck and B.A. Alford. *Depression: Causes and Treatment*. University of Pennsylvania Press, 2009.

[33] Patrice Bellot, Antoine Doucet, Shlomo Geva, Sairam Gurajada, Jaap Kamps, Gabriella Kazai, Marijn Koolen, Arunav Mishra, Véronique Moriceau, Josiane Mothe, et al. Report on inex 2013. In *ACM SIGIR Forum*, volume 47, pages 21–32. ACM New York, NY, USA, 2013.

[34] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[35] Yoshua Bengio, Patrice Y. Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Networks*, 5(2):157–166, 1994.

[36] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.

[37] Adam J Berinsky. Rumors and health care reform: Experiments in political misinformation. *British journal of political science*, 47(2):241–262, 2017.

[38] Filippo Maria Bianchi, Enrico Maiorino, Michael C Kampffmeyer, Antonello Rizzi, and Robert Jenssen. An overview and comparative analysis of recurrent neural networks for short term load forecasting. *arXiv preprint arXiv:1705.04378*, 2017.

[39] Christopher M Bishop et al. *Neural networks for pattern recognition*. Oxford university press, 1995.

[40] Gerhard Blanken, Jürgen Dittmann, Hannelore Grimm, John C Marshall, and Claus-W Wallesch. *Linguistic disorders and pathologies: An international handbook*, volume 8. Walter de Gruyter, 2008.

[41] Tobias Bocklet, Elmar Nöth, Georg Stemmer, Hana Ruzickova, and Jan Rusz. Detection of persons with parkinson's disease by acoustic, vocal, and prosodic analysis. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 478–483. IEEE, 2011.

[42] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[43] Charles F Bond Jr and Bella M DePaulo. Accuracy of deception judgments. *Personality and social psychology Review*, 10(3):214–234, 2006.

[44] E.H. Bos, A.L. Bouhuys, E. Geerts, T.W.D.P. Van Os, and J. Ormel. Lack of association between conversation partners' nonverbal behavior predicts recurrence of depression, independently of personality. *Psychiatry Research*, 142(1):79–88, 2006.

[45] Petter Bae Brandtzæg and Asbjørn Følstad. Trust and distrust in online fact-checking services. *Commun. ACM*, 60(9):65–71, 2017.

[46] Paul R Brewer, Dannagal Goldthwaite Young, and Michelle Morreale. The impact of real news about "fake news": Intertextual processes and political satire. *International Journal of Public Opinion Research*, 25(3):323–343, 2013.

[47] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. 1998.

[48] S Brindha, K Prabha, and S Sukumaran. A survey on classification techniques for text mining. In *2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 1–5. IEEE, 2016.

[49] Alexander Budanitsky and Graeme Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational linguistics*, 32(1):13–47, 2006.

[50] H. Cai, X. Zhang, Y. Zhang, Z. Wang, and B. Hu. A case-based reasoning model for depression based on three-electrode eeg data. *IEEE Transactions on Affective Computing (to appear)*, 2019.

[51] Michael P Caligiuri and Joel Ellwanger. Motor and cognitive aspects of motor retardation in depression. *Journal of affective disorders*, 57(1-3):83–93, 2000.

[52] Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. Affective computing and sentiment analysis. In *A practical guide to sentiment analysis*, pages 1–10. Springer, 2017.

[53] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684, 2011.

[54] Pew Research Centre. News Use Across Social Media Platforms 2018 . `https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/`, 2018. [Online; accessed 22-February-2021].

[55] Wallace Chafe. Integration and involvement in speaking, writing, and oral literature. *Spoken and written language: Exploring orality and literacy*, pages 35–54, 1982.

[56] Iti Chaturvedi, Erik Cambria, Roy E Welsch, and Francisco Herrera. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, 44:65–77, 2018.

[57] Vassilios Chatzis, Adrian G Bors, and Ioannis Pitas. Multimodal decision-level fusion for person authentication. *IEEE transactions on systems, man, and cybernetics-part a: systems and humans*, 29(6):674–680, 1999.

[58] Danqi Chen and Christopher D Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, 2014.

[59] Tong Chen, Xue Li, Hongzhi Yin, and Jun Zhang. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 40–52. Springer, 2018.

[60] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 551–561, 2016.

[61] Michael Chmielewski, Lee Anna Clark, R Michael Bagby, and David Watson. Method matters: Understanding diagnostic reliability in dsm-iv and dsm-5. *Journal of abnormal psychology*, 124(3):764, 2015.

[62] Noam Chomsky and Morris Halle. The sound pattern of english. 1968.

[63] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585, 2015.

[64] Gary Christopher and John MacDonald. The impact of clinical depression on working memory. *Cognitive neuropsychiatry*, 10(5):379–399, 2005.

[65] Cindy Chung and James W Pennebaker. The psychological functions of function words. *Social communication*, 1:343–359, 2007.

[66] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193, 2015.

[67] David A Clark, Aaron T Beck, Brad A Alford, Peter J Bieling, and Zindel V Segal. Scientific foundations of cognitive theory and therapy of depression, 2000.

[68] Manny Cohen. Fake news and manipulated data, the new gdpr, and the future of information. *Business Information Review*, 34(2):81–85, 2017.

[69] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. Detecting depression from facial actions and vocal prosody. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7. IEEE, 2009.

[70] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P.P. Kuksa. Natural language processing (almost) from scratch. *CoRR*, abs/1103.0398, 2011.

[71] Nadia K Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

[72] William E Cooper and Jeanne Paccia-Cooper. *Syntax and speech*. Number 3. Harvard University Press, 1980.

[73] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, 2015.

[74] N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski. Analysis of acoustic space variability in speech affected by depression. *Speech Communication*, 75:27–49, 2015.

[75] N. Cummins, V. Sethu, J. Epps, J.R. Williamson, T.F. Quatieri, and J. Krajewski. Generalized two-stage rank regression framework for depression score prediction from speech. *IEEE Transactions on Affective Computing (to appear)*, 2019.

[76] Nicholas Cummins. Automatic assessment of depression from speech: paralinguistic analysis, modelling and machine learning. *School of Electrical Engineering and Telecommunications, PhD Thesis, UNSW Australia, Sydney, Australia*, 2016.

[77] Nicholas Cummins, Alice Baird, and Björn W Schuller. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*, 151:41–54, 2018.

[78] Nicholas Cummins, Julien Epps, Michael Breakspear, and Roland Goecke. An investigation of depressed speech detection: Features and normalization. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[79] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

[80] Nicholas Cummins, Vidhyasaharan Sethu, Julien Epps, James R Williamson, Thomas F Quatieri, and Jarek Krajewski. Generalized two-stage rank regression framework for depression score prediction from speech. *IEEE Transactions on Affective Computing*, 2017.

[81] R.C.N. D'Arcy, J.F. Connolly, E. Service, C.S. Hawco, and M.E. Houlihan. Separating phonological and semantic processing in auditory sentence processing: A high-resolution event-related brain potential study. *Human Brain Mapping*, 22(1):40–51, 2004.

[82] Louis Daudet, Nikhil Yadav, Matthew Perez, Christian Poellabauer, Sandra Schneider, and Alan Huebner. Portable mtbi assessment using temporal and frequency analysis of speech. *IEEE journal of biomedical and health informatics*, 21(2):496–506, 2016.

[83] T.H. Davenport. *The AI advantage: How to put the artificial intelligence revolution to work*. MIT Press, 2018.

[84] Randall Davis and Douglas B Lenat. Knowledge-based systems in artificial intelligence. 1982.

[85] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[86] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*, 2013.

[87] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[88] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE, 2014.

[89] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[90] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[91] Shubham Dham, Anirudh Sharma, and Abhinav Dhall. Depression scale recognition from audio, visual and text analysis. *arXiv preprint arXiv:1709.05865*, 2017.

[92] Heinrich Dinkel, Mengyue Wu, and Kai Yu. Text-based depression detection: What triggers an alert. *arXiv preprint arXiv:1904.05154*, 2019.

[93] Cícero Nogueira dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. Attentive pooling networks. abs/1602.03609, 2016.

[94] Phuc H Duong, Hien T Nguyen, Hieu N Duong, Khoa Ngo, and Dat Ngo. A hybrid approach to paraphrase detection. In *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 366–371. IEEE, 2018.

[95] Emile Durkheim. Suicide: a study in sociology [1897]. *Translated by JA Spaulding and G. Simpson (Glencoe, Illinois: The Free Press, 1951)*, 1951.

[96] Emile Durkheim and A Suicide. *A study in sociology*. Routledge & K. Paul London, 1952.

[97] Ullrich KH Ecker, Joshua L Hogan, and Stephan Lewandowsky. Reminders and repetition of misinformation: Helping or hindering its retraction? *Journal of Applied Research in Memory and Cognition*, 6(2):185–192, 2017.

[98] Paul Ekman. *Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition)*. WW Norton & Company, 2009.

[99] M.S. El Ayadi, M.and Kamel and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[100] Vyvyan Evans. Cognitive linguistics. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(2):129–141, 2012.

[101] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM, 2013.

[102] Florian Eyben, Martin Wöllmer, and Björn Schuller. Openear—introducing the munich open-source emotion and affect recognition toolkit. In *2009 3rd international conference on affective computing and intelligent interaction and workshops*, pages 1–6. IEEE, 2009.

[103] Zhou Faguo, Zhang Fan, Yang Bingru, and Yu Xingang. Research on short text classification algorithm based on statistics and rules. In *2010 Third international symposium on electronic commerce and security*, pages 3–7. IEEE, 2010.

[104] Mehrdad Farajtabar, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha. Fake news mitigation via point process based intervention. *arXiv preprint arXiv:1703.07823*, 2017.

[105] Maurizio Fava and Kenneth S Kendler. Major depressive disorder. *Neuron*, 28(2):335–341, 2000.

[106] Song Feng, Ritwik Banerjee, and Yejin Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual MeetRao Rao ing of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 171–175, 2012.

[107] Vanessa Wei Feng and Graeme Hirst. Detecting deceptive opinions with profile compatibility. In *Proceedings of the sixth international joint conference on natural language processing*, pages 338–346, 2013.

[108] William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1168, 2016.

[109] Andrew J Flanagin and Miriam J Metzger. *Digital media and youth: Unparalleled opportunity and unprecedented responsibility*. MacArthur Foundation Digital Media and Learning Initiative, 2008.

[110] Alistair J Flint, Sandra E Black, Irene Campbell-Taylor, Gillian F Gailey, and Carey Levinton. Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. *Journal of psychiatric research*, 27(3):309–319, 1993.

[111] World Economic Forum. Outlook on the Global Agenda 2014. `http://reports.weforum.org/outlook-14/`, 2017. [Online; accessed 17-February-2021].

[112] Daniel Joseph France, Richard G Shiavi, Stephen Silverman, Marilyn Silverman, and M Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering*, 47(7):829–837, 2000.

[113] Ray J Frank, Neil Davey, and Stephen P Hunt. Time series prediction and neural networks. *Journal of intelligent and robotic systems*, 31(1-3):91–103, 2001.

[114] Kathleen C Fraser, Frank Rudzicz, and Graeme Hirst. Detecting late-life depression in alzheimer's disease through analysis of speech and language. In *Proceedings of the*

Third Workshop on Computational Linguistics and Clinical Psychology, pages 1–11, 2016.

[115] Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[116] Justin P Friesen, Troy H Campbell, and Aaron C Kay. The psychological advantage of unfalsifiability: The appeal of untestable religious and political ideologies. *Journal of personality and social psychology*, 108(3):515, 2015.

[117] J. Garber, C.M. Gallerani, and S. A. Frankel. Depression in children. In I.H. Gotlib and C.L. Hammen, editors, *Depression in children*, pages 405–443. The Guilford Press, 2009.

[118] E. Geerts, N. Bouhuys, and R.H. Van den Hoofdakker. Nonverbal attunement between depressed patients and an interviewer predicts subsequent improvement. *Journal of Affective Disorders*, 40(1-2):15–21, 1996.

[119] S. Gilbody, D. Richards, S. Brealey, and C. Hewitt. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): A diagnostic meta-analysis. *Journal of General Internal Medicine*, 22(11):1596–1602, 2007.

[120] J.M. Girard, J.F. Cohn, M.H. Mahoor, S. Mavadati, and D.P. Rosenwald. Social risk and depression: Evidence from manual and automatic facial expression analysis. In *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 1–8, 2013.

[121] Yuan Gong and Christian Poellabauer. Continuous assessment of children's emotional states using acoustic analysis. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 171–178. IEEE, 2017.

[122] Yuan Gong and Christian Poellabauer. Topic modeling based multi-modal depression detection. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 69–76, 2017.

[123] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[124] Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128. Citeseer, 2014.

[125] A. Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer Verlag, 2012.

[126] John J Gumperz, Hannah Kaltman, and MARY CATHERINE O'Connor. Cohesion in spoken and written discourse: Ethnic style and the transition to literacy. *Coherence in spoken and written discourse*, 12:3–19, 1984.

[127] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[128] Max Hamilton. The hamilton rating scale for depression. In *Assessment of depression*, pages 143–152. Springer, 1986.

[129] Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. The quest to automate fact-checking. *world*, 2015.

[130] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. Claimbuster: The first-ever end-to-end fact-checking system. *PVLDB*, 10(12):1945–1948, 2017.

[131] Lang He and Cui Cao. Automated depression analysis using convolutional neural networks from speech. *Journal of biomedical informatics*, 83:103–111, 2018.

[132] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. End-to-end text-dependent speaker verification. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5115–5119. IEEE, 2016.

[133] Stefan Helmstetter and Heiko Paulheim. Weakly supervised learning for fake news detection on twitter. In *IEEE/ACM 2018 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018, Barcelona, Spain, August 28-31, 2018*, pages 274–277, 2018.

[134] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701, 2015.

[135] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.

[136] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[137] Florian Hönig, Anton Batliner, Elmar Nöth, Sebastian Schnieder, and Jarek Krajewski. Automatic modelling of depressed speech: relevant features and relevance of gender. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[138] Fei Huang and Alexander Yates. Distributional representations for handling sparsity in supervised sequence-labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 495–503, 2009.

[139] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Reddy. *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, 2001.

[140] Z. Huang, J. Epps, D. Joachim, and M. Chen. Depression detection from short utterances via diverse smartphones in natural environmental conditions. In *Proceedings of Interspeech*, pages 3393–3397, 2018.

[141] Satoshi Imai, Takao Kobayashi, Keiichi Tokuda, T Masuko, K Koishida, S Sako, and H Zen. Speech signal processing toolkit (sptk), 2009.

[142] C. Irons. *Depression*. Palgrave, 2014.

[143] Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. Claimrank: Detecting check-worthy claims in arabic and english. *arXiv preprint arXiv:1804.07587*, 2018.

[144] Mario Jarmasz. Roget's thesaurus as a lexical resource for natural language processing. *arXiv preprint arXiv:1204.0140*, 2012.

[145] J. Joshi, R. Goecke, G. Parker, and M. Breakspear. Can body expressions contribute to automatic depression analysis? In *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–7, 2013.

[146] Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. Fully automated fact checking using external sources. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 344–353, 2017.

[147] Ray D Kent and Y-J Kim. Toward an acoustic typology of motor speech disorders. *Clinical linguistics & phonetics*, 17(6):427–445, 2003.

[148] R.C. Kessler, P. Berglund, O. Demler, R. Jin, D. Koretz, K.R. Merikangas, A.J. Rush, E.E. Walters, and P.S. Wang. The epidemiology of major depressive disorder: Results from the National Comorbidity Survey Replication (NCS-R). *Journal of the American Medical Association*, 289(23):3095–3105, 2003.

[149] R.C. Kessler and E.E. Walters. Epidemiology of DSM-III-R major depression and minor depression among adolescents and young adults in the national comorbidity survey. *Depression and Anxiety*, 7(1):3–14, 1998.

[150] Ronald C Kessler, Katherine A McGonagle, Marvin Swartz, Dan G Blazer, and Christopher B Nelson. Sex and depression in the national comorbidity survey i: Lifetime prevalence, chronicity and recurrence. *Journal of affective disorders*, 29(2-3):85–96, 1993.

[151] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, and Khairullah Khan. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1):4–20, 2010.

[152] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[153] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[154] Josef Kittler, Mohamad Hatef, Robert PW Duin, and Jiri Matas. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239, 1998.

[155] David Klein and Joshua Wueller. Fake news: A legal perspective. *Journal of Internet Law (Apr. 2017)*, 2017.

[156] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[157] Lars Konieczny. Locality and parsing complexity. *Journal of psycholinguistic research*, 29:627–645, 2016.

[158] Helena Chmura Kraemer, David J Kupfer, Diana E Clarke, William E Narrow, and Darrel A Regier. Dsm-5: how reliable is reliable enough? *American Journal of Psychiatry*, 169(1):13–15, 2012.

[159] Jarek Krajewski, Sebastian Schnieder, David Sommer, Anton Batliner, and Björn Schuller. Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech. *Neurocomputing*, 84:65–75, 2012.

[160] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613, 2001.

[161] Srijan Kumar and Neil Shah. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*, 2018.

[162] L.I. Kuncheva and C.J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

[163] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining*, pages 1103–1108. IEEE, 2013.

[164] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1675–1684, 2016.

[165] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*, 2017.

[166] Genevieve Lam, Huang Dongyan, and Weisi Lin. Context-aware deep learning for multi-modal depression detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3946–3950. IEEE, 2019.

[167] Charles Lama, Brian Leungb, Cora Yipb, and Jason Yungb. A linguistic approach to misinformation in chinese. *Proceedings http://ceur-ws. org ISSN*, 1613:0073, 2020.

[168] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

[169] Katie Langin. Fake news spreads faster than true news on twitter—thanks to people, not bots. *Science Magazine*, 2018.

[170] Kornel Laskowski, Mari Ostendorf, and Tanja Schultz. Modeling vocal interaction for text-independent participant characterization in multi-party conversation. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 148–155, 2008.

[171] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[172] Chi-Chun Lee, Athanasios Katsamanis, Matthew P Black, Brian R Baucom, Panayiotis G Georgiou, and Shrikanth Narayanan. An analysis of pca-based vocal entrainment measures in married couples' affective spoken interactions. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[173] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1695–1699. IEEE, 2014.

[174] Douglas B Lenat and Ramanathan V Guha. *Building large knowledge-based systems; representation and inference in the Cyc project*. Addison-Wesley Longman Publishing Co., Inc., 1989.

[175] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, et al. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, 9(3):1350–1371, 2015.

[176] Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. *CoRR*, abs/1607.06275, 2016. Withdrawn.

[177] Chloe Lim. Checking how fact-checkers check. *Research & Politics*, 5(3):2053168018786848, 2018.

[178] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130, 2017.

[179] Charles X Ling, Jin Huang, Harry Zhang, et al. Auc: a statistically consistent and more discriminating measure than accuracy. In *IJCAI*, volume 3, pages 519–524, 2003.

[180] Tal Linzen, Grzegorz Chrupała, and Afra Alishahi. Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018.

[181] Marco Lippi and Paolo Torroni. Argument mining from speech: Detecting claims in political debates. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[182] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[183] Zachary C Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

[184] Yang Liu and Yi-Fang Wu. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[185] Elizabeth F Loftus. Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & memory*, 12(4):361–366, 2005.

[186] David E Losada and Fabio Crestani. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer, 2016.

[187] L. A. Low, N. C. Maddage, M. Lech, L.B. Sheeber, and N.B. Allen. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–586, 2011.

[188] Lu-Shih Alex Low, Namunu C Maddage, Margaret Lech, Lisa B Sheeber, and Nicholas B Allen. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–586, 2010.

[189] Hans Peter Luhn. Key word-in-context index for technical literature (kwic index). *American documentation*, 11(4):288–295, 1960.

[190] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. 2016.

[191] Jing Ma, Wei Gao, and Kam-Fai Wong. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the*

Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pages 708–717, 2017.

[192] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 35–42, 2016.

[193] Mohammad Mahyoob, Jeehaan Al-Garaady, and Musaad Alrahaili. Linguistic-based detection of fake news in social media. *Forthcoming, International Journal of English Linguistics*, 11(1), 2020.

[194] Nicolas Malyska, Thomas F Quatieri, and Douglas Sturim. Automatic dysphonia recognition using biologically-inspired amplitude-modulation features. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I–873. IEEE, 2005.

[195] Morgan Marietta, David C Barker, and Todd Bowser. Fact-checking polarized politics: Does the fact-check industry provide consistent guidance on disputed realities? In *The Forum*, volume 13, pages 577–596. De Gruyter, 2015.

[196] Stephen Marsland. *Machine learning: an algorithmic perspective*. CRC press, 2015.

[197] Benoit Mathieu, Slim Essid, Thomas Fillon, Jacques Prado, and Gaël Richard. Yaafe, an easy to use and efficient audio feature extraction software. In *ISMIR*, pages 441–446, 2010.

[198] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[199] F.A. McDougall, F.E. Matthews, K. Kvaal, M.E. Dewey, and C. Brayne. Prevalence and symptomatology of depression in older people living in institutions in england and wales. *Age and Ageing*, 36(5):562–568, 2007.

[200] Hongying Meng, Di Huang, Heng Wang, Hongyu Yang, Mohammed Ai-Shuraifi, and Yunhong Wang. Depression recognition based on dynamic facial and vocal expres-

sion features using partial least square regression. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 21–30, 2013.

[201] Florian Metze, Jitendra Ajmera, Roman Englert, Udo Bub, Felix Burkhardt, Joachim Stegmann, Christian Muller, Richard Huber, Bernt Andrassy, Josef G Bauer, et al. Comparison of four approaches to age and gender recognition for telephone applications. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–1089. IEEE, 2007.

[202] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[203] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[204] A.J. Mitchell, A. Vaze, and S. Rao. Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet*, 374(9690):609 – 619, 2009.

[205] Alex J Mitchell, Amol Vaze, and Sanjay Rao. Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet*, 374(9690):609–619, 2009.

[206] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton. Acoustic modeling using deep belief networks. *IEEE transactions on audio, speech, and language processing*, 20(1):14–22, 2011.

[207] Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. Automatic stance detection using end-to-end memory networks. *arXiv preprint arXiv:1804.07581*, 2018.

[208] Michelle Morales, Stefan Scherer, and Rivka Levitan. A cross-modal review of indicators for depression detection systems. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 1–12, 2017.

[209] Michelle Morales, Stefan Scherer, and Rivka Levitan. A linguistically-informed fusion approach for multimodal depression detection. In *Proceedings of the Fifth Workshop*

on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pages 13–24, 2018.

[210] Michelle Renee Morales and Rivka Levitan. Speech vs. text: A comparative analysis of features for depression detection systems. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 136–143. IEEE, 2016.

[211] M.R. Morales and R. Levitan. Speech vs. text: A comparative analysis of features for depression detection systems. In *proceedings of the IEEE Spoken Language Technology Workshop*, pages 136–143, 2016.

[212] Charles A Morgan III, Steven Southwick, George Steffian, Gary A Hazlett, and Elizabeth F Loftus. Misinformation can influence memory for recently experienced, highly stressful events. *International journal of law and psychiatry*, 36(1):11–17, 2013.

[213] Eni Mustafaraj and Panagiotis Takis Metaxas. The fake news spreading plague: was it preventable? In *Proceedings of the 2017 ACM on web science conference*, pages 235–239, 2017.

[214] John Naughton. The evolution of the internet: from military experiment to general purpose technology. *Journal of Cyber Policy*, 1(1):5–28, 2016.

[215] Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, 5(3):217–226, 2014.

[216] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623, 2003.

[217] Brendan Nyhan and Jason Reifler. The effect of fact-checking on elites: A field experiment on us state legislators. *American Journal of Political Science*, 59(3):628–640, 2015.

[218] Brendan Nyhan and Jason Reifler. Estimating fact-checking's effects. *Arlington, VA: American Press Institute*, 2015.

[219] Mental Health Commission of Canada Strategic Plan 2017-2022. Mental health commission of canada. Technical report, 2017.

[220] Kuan Ee Brian Ooi, Lu-Shih Alex Low, Margaret Lech, and Nicholas Allen. Prediction of clinical depression in adolescents using facial image analaysis. 2011.

[221] World Health Organization. Depression let's talk" says WHO, as depression tops list of causes of ill health. https://www.who.int/news/item/30-03-2017--depression-let-s-talk-says-who-as-depression-tops-list-o 2017. [Online; accessed 17-February-2021].

[222] World Health Organization. Depression can lead to suicide. https://www.who.int/news-room/fact-sheets/detail/depression, 2020. [Online; accessed 17-February-2021].

[223] World Health Organization et al. Depression and other common mental disorders: global health estimates. Technical report, World Health Organization, 2017.

[224] World Health Organization, World Health Organization, et al. May 2017. *URL http://www. who. int/mental_ health/management/depression/en*, 1(10), 4.

[225] Michael A Osborne. *Bayesian Gaussian processes for sequential prediction, optimisation and quadrature*. PhD thesis, Oxford University, UK, 2010.

[226] Douglas O'shaughnessy. *Speech Communications: Human And Machine (IEEE)*. Universities press, 1987.

[227] Sharon Oviatt, Björn Schuller, Philip Cohen, Daniel Sonntag, Gerasimos Potamianos, and Antonio Krüger. *The handbook of multimodal-multisensor interfaces, Volume 2: Signal processing, architectures, and detection of emotion and cognition*. Morgan & Claypool, 2018.

[228] Asli Ozdas, Richard G Shiavi, Stephen E Silverman, Marilyn K Silverman, and D Mitchell Wilkes. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Transactions on Biomedical Engineering*, 51(9):1530–1540, 2004.

[229] Yasin Özkanca, Cenk Demiroglu, Asli Besirli, and Selime Celik. Multi-lingual depression-level assessment from conversational speech using acoustic and text features. In *Interspeech*, pages 3398–3402, 2018.

[230] Dimitri Palaz, Ronan Collobert, and Mathew Magimai Doss. Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. *arXiv preprint arXiv:1304.1018*, 2013.

[231] R. Pascanu, T. Mikolov, and Y. Bengio. Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, 2:417, 2012.

[232] D.L. Paulhus and S. Vazire. The self-report method. In R.W. Robins, R.C. Fraley, and R.F. Krueger, editors, *Handbook of Research Methods in Personality Psychology*, pages 224–239. Gilford, 2007.

[233] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[234] Gordon Pennycook, Tyrone D Cannon, and David G Rand. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general*, 147(12):1865, 2018.

[235] A. Pentland. *Honest Signals*. MIT press, 2007.

[236] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*, 2017.

[237] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[238] P Pichot. Self-report inventories in the study of depression. In *New Results in Depression Research*, pages 53–58. Springer, 1986.

[239] Kashyap Popat. Assessing the credibility of claims on the web. In *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, pages 735–739, 2017.

[240] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1003–1012, 2017.

[241] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 22–32, 2018.

[242] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[243] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*, 2017.

[244] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. abs/1702.05638, 2017.

[245] Chris Poulin, Brian Shiner, Paul Thompson, Linas Vepstas, Yinong Young-Xu, Benjamin Goertzel, Bradley Watts, Laura Flashman, and Thomas McAllister. Predicting the risk of suicide by analyzing the text of clinical notes. *PloS one*, 9(1):e85733, 2014.

[246] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011.

[247] Tom Pyszczynski and Jeff Greenberg. Self-regulatory perseveration and the depressive self-focusing style: a self-awareness theory of reactive depression. *Psychological bulletin*, 102(1):122, 1987.

[248] Thomas F Quatieri and Nicolas Malyska. Vocal-source biomarkers for depression: A link to psychomotor activity. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[249] Syed Arbaaz Qureshi, Mohammed Hasanuzzaman, Sriparna Saha, and Gaël Dias. The verbal and non verbal signals of depression–combining acoustics, text and visuals for estimating depression level. *arXiv preprint arXiv:1904.07656*, 2019.

[250] Lawrence Rabiner. On the use of autocorrelation analysis for pitch detection. *IEEE transactions on acoustics, speech, and signal processing*, 25(1):24–33, 1977.

[251] Dr Rajni and Nripendra Narayan Das. Emotion recognition from audio signal. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 5(6), 2016.

[252] R. Ranawana and V. Palade. Multi-classifier systems: Review and a roadmap for developers. *International Journal of Hybrid Intelligent Systems*, 3(1):35–61, 2006.

[253] Josyula R Rao, Pankaj Rohatgi, et al. Can pseudonymity really guarantee privacy? In *USENIX Security Symposium*, pages 85–96, 2000.

[254] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, 2017.

[255] Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg. Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 81–88, 2019.

[256] Emna Rejaibi, Ali Komaty, Fabrice Meriaudeau, Said Agrebi, and Alice Othmani. Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *arXiv preprint arXiv:1909.07208*, 2019.

[257] Jimmy Ren, Yongtao Hu, Yu-Wing Tai, Chuan Wang, Li Xu, Wenxiu Sun, and Qiong Yan. Look, listen and learn—a multimodal lstm for speaker identification. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[258] Yanping Ren, Hui Yang, Colette Browning, Shane Thomas, and Meiyan Liu. Performance of screening tools in detecting major depressive disorder among patients with coronary heart disease: a systematic review. *Medical science monitor: international medical journal of experimental and clinical research*, 21:646, 2015.

[259] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

[260] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic. AVEC 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the International Workshop on Audio/Visual Emotion Challenge*, pages 3–9, 2017.

[261] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9, 2017.

[262] Ruty Rinott, Lena Dankin, Carlos Alzate, Mitesh M Khapra, Ehud Aharoni, and Noam Slonim. Show me your evidence-an automatic method for context dependent evidence detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 440–450, 2015.

[263] Marko Robnik-Šikonja and Igor Kononenko. Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600, 2008.

[264] M. Rohanian, J. Hough, and M. Purver. Detecting depression with word-level multimodal fusion. In *Proceedings of Interspeech*, pages 1443–1447, 2019.

[265] Morteza Rohanian, Julian Hough, Matthew Purver, et al. Detecting depression with word-level multimodal fusion. *Proc. Interspeech 2019*, pages 1443–1447, 2019.

[266] Frank Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY, 1961.

[267] Victoria L Rubin. Deception detection and rumor debunking for social media. *The SAGE Handbook of Social Media Research Methods*, pages 342–364, 2017.

[268] Victoria L Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17, 2016.

[269] Natali Ruchansky, Sungyong Seo, and Yan Liu. CSI: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 797–806, 2017.

[270] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133, 2004.

[271] DE Rumellhart. Learning internal representations by error propagation. *Parallel distributed processing*, 1:318–362, 1986.

[272] Saurabh Sahu and Carol Y Espy-Wilson. Speech features for depression detection. In *INTERSPEECH*, pages 1928–1932, 2016.

[273] Tara N Sainath, Ron J Weiss, Andrew Senior, Kevin W Wilson, and Oriol Vinyals. Learning the speech front-end with raw waveform cldnns. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[274] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[275] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[276] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

[277] Klaus R Scherer. Vocal affect expression: A review and a model for future research. *Psychological bulletin*, 99(2):143, 1986.

[278] Stefan Scherer, Gale M Lucas, Jonathan Gratch, Albert Skip Rizzo, and Louis-Philippe Morency. Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews. *IEEE Transactions on Affective Computing*, 7(1):59–73, 2015.

[279] Stefan Scherer, Giota Stratou, Gale Lucas, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, Louis-Philippe Morency, et al. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, 32(10):648–658, 2014.

[280] Stefan Scherer, Giota Stratou, and Louis-Philippe Morency. Audiovisual behavior descriptors for depression assessment. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 135–140, 2013.

[281] Jean Schoentgen. Vocal cues of disordered voices: an overview. *Acta Acustica united with Acustica*, 92(5):667–680, 2006.

[282] B. Schuller and A. Batliner. *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[283] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27:4–39, 2013.

[284] Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, et al. The interspeech 2012 speaker trait challenge. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[285] Isabelle Schumann, Antonius Schneider, Claudia Kantert, Bernd Löwe, and Klaus Linde. Physicians' attitudes, diagnostic process and barriers regarding depression diagnosis in primary care: a systematic review of qualitative studies. *Family practice*, 29(3):255–263, 2011.

[286] Isabelle Schumann, Antonius Schneider, Claudia Kantert, Bernd Löwe, and Klaus Linde. Physicians' attitudes, diagnostic process and barriers regarding depression diagnosis in primary care: a systematic review of qualitative studies. *Family practice*, 29(3):255–263, 2012.

[287] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[288] H Andrew Schwartz, Johannes Eichstaedt, Margaret L Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125, 2014.

[289] H Andrew Schwartz and Lyle H Ungar. Data-driven content analysis of social media: a systematic overview of automated methods. *The ANNALS of the American Academy of Political and Social Science*, 659(1):78–94, 2015.

[290] M.L. Seghier, F. Lazeyras, A.J. Pegna, J.-M. Annoni, I. Zimine, E. Mayer, C.M. Michel, and A. Khateb. Variability of fMRI activation during a phonological and

semantic language task in healthy subjects. *Human Brain Mapping*, 23(3):140–155, 2004.

[291] M Sharifa, Roland Goecke, Michael Wagner, Julien Epps, Michael Breakspear, Gordon Parker, et al. From joyous to clinically depressed: Mood detection using spontaneous speech. In *Twenty-Fifth International FLAIRS Conference*, 2012.

[292] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 395–405, 2019.

[293] Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 626–637, 2020.

[294] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[295] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1), September 2017.

[296] Kai Shu, Suhang Wang, and Huan Liu. Understanding user profiles on social media for fake news detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 430–435. IEEE, 2018.

[297] Olympia Simantiraki, Paulos Charonyktakis, Anastasia Pampouchidou, Manolis Tsiknakis, and Martin Cooke. Glottal source features for automatic speech-based depression assessment. In *INTERSPEECH*, pages 2700–2704, 2017.

[298] D Smirnova, E Sloeva, N Kuvshinova, A Krasnov, D Romanov, and G Nosachev. 1419–language changes as an important psychopathological phenomenon of mild depression. *European Psychiatry*, 28:1, 2013.

[299] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, 2005.

[300] Christina Sobin and Harold A Sackeim. Psychomotor symptoms of depression. *American Journal of Psychiatry*, 154(1):4–17, 1997.

[301] Brian Stasak, Julien Epps, Nicholas Cummins, and Roland Goecke. An investigation of emotional speech in depression classification. In *Interspeech*, pages 485–489, 2016.

[302] HH Stassen et al. Speaking behavior and voice sound characteristics in depressive patients during recovery. *Journal of psychiatric research*, 27(3):289–307, 1993.

[303] HH Stassen, S Kuny, and D Hell. The speech analysis approach to determining onset of improvement under antidepressants. *European Neuropsychopharmacology*, 8(4):303–310, 1998.

[304] Stanley Smith Stevens, John Volkmann, and Edwin B Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

[305] Shannon Wiltsey Stirman and James W Pennebaker. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine*, 63(4):517–522, 2001.

[306] E. Stockings, L. Degenhardt, Y.Y. Lee, C. Mihalopoulos, A. Liu, M. Hobbs, and G. Patton. Symptom screening scales for detecting major depressive disorder in children and adolescents: a systematic review and meta-analysis of reliability, validity and diagnostic utility. *Journal of Affective Disorders*, 174:447–463, 2015.

[307] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706, 2007.

[308] Briony Swire, Ullrich KH Ecker, and Stephan Lewandowsky. The role of familiarity in correcting inaccurate information. *Journal of experimental psychology: learning, memory, and cognition*, 43(12):1948, 2017.

[309] Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. Defining "fake news" a typology of scholarly definitions. *Digital journalism*, 6(2):137–153, 2018.

[310] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[311] James Thorne and Andreas Vlachos. Automated fact checking: Task formulations, methods and future directions. *arXiv preprint arXiv:1806.07687*, 2018.

[312] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

[313] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204. IEEE, 2016.

[314] Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[315] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

[316] Ana Valdivia, M Victoria Luzón, Erik Cambria, and Francisco Herrera. Consensus vote models for detecting and filtering neutrality in sentiment analysis. *Information Fusion*, 44:126–135, 2018.

[317] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the International Workshop on Audio/Visual Emotion Challenge*, pages 3–10, 2016.

[318] M. Valstar, B. Schuller, J. Krajewski, J. Cohn, R. Cowie, and M. Pantic. AVEC 2014 – The three dimensional affect and depression challenge. In *Proceedings of the ACM International Workshop on Audio/Visual Emotion Challenge*, pages 1–9, 2014.

[319] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. AVEC 2013: The continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the ACM International Workshop on Audio/visual Emotion Challenge*, pages 3–10, 2013.

[320] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10, 2016.

[321] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*, pages 3–10, 2014.

[322] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM, 2013.

[323] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4052–4056. IEEE, 2014.

[324] A. Vinciarelli and G. Mohammadi. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014.

[325] Alessandro Vinciarelli, Maja Pantic, Hervé Bourlard, and Alex Pentland. Social signals, their function, and automatic analysis: a survey. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 61–68, 2008.

[326] Nguyen Vo and Kyumin Lee. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 275–284, 2018.

[327] Nguyen Vo and Kyumin Lee. Learning from fact-checkers: analysis and generation of fact-checking language. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–344, 2019.

[328] Nguyen Vo and Kyumin Lee. Standing on the shoulders of guardians: Novel methodologies to combat fake news. In *Disinformation, Misinformation, and Fake News in Social Media*, pages 183–210. Springer, 2020.

[329] Theo Vos, Amanuel Alemu Abajobir, Kalkidan Hassen Abate, Cristiana Abbafati, Kaja M Abbas, Foad Abd-Allah, Rizwan Suliankatchi Abdulkader, Abdishakur M Abdulle, Teshome Abuka Abebo, Semaw Ferede Abera, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet*, 390(10100):1211–1259, 2017.

[330] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[331] Michael Wagner. Prosody as a diagonalization of syntax. evidence from complex predicates. In *PROCEEDINGS-NELS*, volume 34, pages 587–602, 2004.

[332] Jerome C Wakefield and Steeves Demazeux. *Sadness Or Depression?: International Perspectives on the Depression Epidemic and Its Meaning*, volume 15. Springer, 2015.

[333] Fulton Wang and Cynthia Rudin. Falling rule lists. In *Artificial Intelligence and Statistics*, pages 1013–1022. PMLR, 2015.

[334] Peilu Wang, Yao Qian, Frank K Soong, Lei He, and Hai Zhao. Part-of-speech tagging with bidirectional long short-term memory recurrent neural network. *arXiv preprint arXiv:1510.06168*, 2015.

[335] Philip S Wang, Matthias Angermeyer, Guilherme Borges, Ronny Bruffaerts, Wai Tat Chiu, Giovanni De Girolamo, John Fayyad, Oye Gureje, Josep Maria Haro, Yueqin Huang, et al. Delay and failure in treatment seeking after first onset of mental disorders in the world health organization's world mental health survey initiative. *World psychiatry*, 6(3):177, 2007.

[336] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 422–426, 2017.

[337] Yuxuan Wang, Yutai Hou, Wanxiang Che, and Ting Liu. From static to dynamic word representations: a survey. *International Journal of Machine Learning and Cybernetics*, pages 1–20, 2020.

[338] Paul J Werbos. Generalization of backpropagation with application to a recurrent gas market model. *Neural networks*, 1(4):339–356, 1988.

[339] WHO Document Production Services. Depression and other common mental disorders. Technical report, World Health Organization, 2017.

[340] James R Williamson, Elizabeth Godoy, Miriam Cha, Adrianne Schwarzentruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F Quatieri. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 11–18. ACM, 2016.

[341] James R Williamson, Thomas F Quatieri, Brian S Helfer, Gregory Ciccarelli, and Daryush D Mehta. Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 65–72. ACM, 2014.

[342] James R Williamson, Thomas F Quatieri, Brian S Helfer, Rachelle Horwitz, Bea Yu, and Daryush D Mehta. Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 41–48. ACM, 2013.

[343] J.R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T.F. Quatieri. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the International Workshop on Audio/Visual Emotion Challenge*, pages 11–18, 2016.

[344] Lars Willnat and David Hugh Weaver. *American Journalist in the digital age: Key findings*. School of Journalism, Indiana University, 2014.

[345] Andrew Gordon Wilson, Jason Yosinski, Patrice Simard, Rich Caruana, and William Herlands. Proceedings of nips 2017 symposium on interpretable machine learning. *arXiv e-prints*, pages arXiv–1711, 2017.

[346] Liang Wu and Huan Liu. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the eleventh ACM international conference on Web Search and Data Mining*, pages 637–645, 2018.

[347] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[348] Zhiyong Wu, Lianhong Cai, and Helen Meng. Multi-level fusion of audio and visual features for speaker identification. In *International Conference on Biometrics*, pages 493–499. Springer, 2006.

[349] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[350] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. Wikipedia2vec: An optimized implementation for learning embeddings from wikipedia. *arXiv preprint arXiv:1812.06280*, 2018.

[351] Ikuya Yamada and Hiroyuki Shindo. Neural attentive bag-of-entities model for text classification. *arXiv preprint arXiv:1909.01259*, 2019.

[352] L. Yang, D. Jiang, and H. Sahli. Integrating deep and shallow models for multi-modal depression analysis—hybrid architectures. *IEEE Transactions on Affective Computing (to appear)*, 2019.

[353] Le Yang, Dongmei Jiang, and Hichem Sahli. Integrating deep and shallow models for multi-modal depression analysis—hybrid architectures. *IEEE Transactions on Affective Computing*, 2018.

[354] Le Yang, Dongmei Jiang, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 53–59. ACM, 2017.

[355] Y. Yang, C. Fairbairn, and J.F. Cohn. Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing*, 4(2):142–150, 2012.

[356] Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *Icml*, volume 97, page 35. Nashville, TN, USA, 1997.

[357] Ying Yang, Catherine Fairbairn, and Jeffrey F Cohn. Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing*, 4(2):142–150, 2012.

[358] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social media mining: an introduction*. Cambridge University Press, 2014.

[359] Jiawei Zhang, Limeng Cui, Yanjie Fu, and Fisher B Gouza. Fake news detection with deep diffusive network model. *arXiv preprint arXiv:1805.08751*, 2018.

[360] Jie Zhou and Wei Xu. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association*

for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1127–1137, 2015.

[361] X. Zhou, K. Jin, Y. Shang, and G. Guo. Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing (to appear)*, 2019.

[362] Xinyi Zhou and Reza Zafarani. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*, 2018.

[363] Xiuzhuang Zhou, Kai Jin, Yuanyuan Shang, and Guodong Guo. Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing*, 2018.

[364] Y. Zhu, Y. Shang, Z. Shao, and G. Guo. Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. *IEEE Transactions on Affective Computing*, 9(4):578–584, 2017.

[365] Yu Zhu, Yuanyuan Shang, Zhuhong Shao, and Guodong Guo. Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. *IEEE Transactions on Affective Computing*, 9(4):578–584, 2017.

[366] Jörg Zinken, Katarzyna Zinken, J Clare Wilson, Lisa Butler, and Timothy Skinner. Analysis of syntax and word use to predict successful participation in guided self-help for anxiety and depression. *Psychiatry research*, 179(2):181–186, 2010.

[367] Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, 2013.