C02029: Doctor of Philosophy
Subject Code: 32903
September 2022

CRICOS Code: 009469A

Computational Understanding of Figurative Language on Social Media

Rhys Biddle

School of Computer Science
Faculty of Eng. & IT
University of Technology Sydney
NSW - 2007, Australia

Computational Understanding of Figurative Language on Social Media

A thesis submitted in fulfilment of the requirements
for the degree of

Doctor of Philosophy
in
Analytics

by

Rhys Biddle

to

Advanced Analytics Institute (AAi)
School of Computer Science
University of Technology Sydney
NSW - 2007, Australia

September 2022

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Rhys Biddle* declare that this thesis, submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

SIGNATURE: Production Note: Signature removed prior to publication.

[Rhys Biddle]

DATE: 7th September, 2022

PLACE: Melbourne, Australia

Abstract

Figurative language in online user-generated text poses challenges to Natural Language Processing (NLP) systems designed to automate the understanding of natural language. This thesis introduces empirical studies that quantify the presence and describes the nature of figurative language in Social Media posts (i.e. Twitter). It also quantifies the impact of figurative language on particular NLP applications and introduces new resources (i.e. datasets and methodologies) for the computational processing of figurative language.

This thesis contains a focused case-study on general figurative language in the context of Public Health Surveillance (PHS) applications that monitor Twitter for health events. Findings indicate that some symptom and disease topics are mentioned in a figurative context more than in a health context, which results in a biased signal. To address this bias, a new annotated dataset and text classifier is proposed that reduces bias by targeting figurative expressions of health-related concepts on Twitter.

There is limited research on the expression of hyperbole on Twitter compared to other types of figurative language (e.g., metaphor). To address this gap, a dataset of tweets annotated for the presence of hyperbole is collected and explored. Findings show that hyperbole is relatively common on Twitter and the expression of hyperbole varies from simple and repetitive to complex and novel. A common theme of hyperbole expression on Twitter is the strong affective-laden intentions of the authors, heightening the importance of hyperbole understanding for affective computing applications. Several text classifiers are proposed that leverage pre-trained language models, affective signals, and privileged information for the detection of hyperbole. Experiments show improvements in the detection of hyperbole and importantly highlight annotation biases inherent in the current annotation scheme for hyperbole detection, which is likely to be a roadblock to further improvements.

This thesis quantifies the occurrence of figurative language on Twitter and demonstrates a considerable and consistent presence. Additionally, figurative language is often mishandled by various NLP resources and is scantly addressed by existing datasets and methodologies. Experiment results show that through direct targeting and careful handling of figurative language, improvements to the detection of figurative language are achievable. However, it is concluded that the complexity and novelty of figurative language requires further algorithmic and data inventions for continued progress.

Acknowledgments

This thesis could not have been completed without the wisdom and guidance of many individuals from whom I sought advice throughout the journey. I acknowledge those who had a strong and direct influence on my thinking process and this thesis; Prof. Guandong Xu, Dr. Shaowu Liu, Dr. Aditya Joshi, Dr. Maciek Rybinksi and Dr. Cecile Paris. I also acknowledge my colleagues that helped me better formulate and articulate my ideas in our many conversations over this journey. To the anonymous reviewers that also provided insightful feedback to my submissions throughout my candidature I also extend my acknowledgements to you.

List of Publications

Related to the Thesis :

- **Biddle, R.**, Joshi, A., Liu, S., Paris, C. and Xu, G., 2020, April. *Leveraging sentiment distributions to distinguish figurative from literal health reports on Twitter*. In Proceedings of The Web Conference 2020 (pp. 1217-1227).
- **Biddle, R.**, Rybinski, M., Li, Q., Paris, C. and Xu, G., 2021, December. *Harnessing Privileged Information for Hyperbole Detection*. In Proceedings of Australasian Language Technology Association (ALTA).

Others :

- Islam, M.R., Liu, S., **Biddle, R.**, Razzak, I., Wang, X., Tilocca, P. and Xu, G., 2021. *Discovering dynamic adverse behavior of policyholders in the life insurance industry*. Technological Forecasting and Social Change, 163, p.120486.

NOMENCLATURE

API Application Programming Interface

BERT Bidirectional Encoder Representations from Transformers

BiLSTM Bi-Directional Long Short Term Memory Network

BNC British National Corpus

CNN Convolutional Neural Network

ECF Extreme Case Formulation

ELMo Embeddings from Language Model

GloVe Global Vectors for Word Representation

GRNN Gated Recurrent Neural Network

HMC Health Mention Classification

ICD International Statistical Classification of Diseases and Related Health Problems

ICF International Classification of Functioning, Disability and Health

IR Information Retrieval

KNN K-Nearest Neighbour

LDA Latent Dirichlet Allocation

LIWC Linguistic Inquirer Word Count

LSTM Long Short Term Memory Network

MELC Metaphor in end-of-life Care

MIP Metaphor Identification Procedure

NLG Natural Language Generation

NLP Natural Language Processing

NLU Natural Language Understanding

PHS Public Health Surveillance

RNN Recurrent Neural Network

RTT Round Trip Translation

SVM Support Vector Machine

TER Translation Edit Ratio

ULMFit Universal Language Model Fine Tuning for Text Classification

VAD Valence Arousal Dominance

w2v Word2Vec

WHO World Health Organization

Table of Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1 Thesis Statement

This thesis is an investigation into the expression of figurative language in social media, particularly Twitter, as well as an exploration in to the use of machine learning methods to detect and interpret these expressions. The following thesis statement captures the central question that is addressed by the work presented in this thesis;

Accurate computational understanding of figurative language on social media is a complex task that requires modification of existing and creation of new datasets and methodologies.

This question is addressed via the introduction of new datasets on the phenomena of figurative language, qualitative and quantitative analysis of these datasets and the introduction of novel machine learning algorithms evaluated on these datasets.

1.2 Background

This section provides a brief overview of three topics that are important to the work presented in this thesis, including figurative language, social media and Natural Language Processing (NLP). Definitions and relationship between these three topics and how they are related in the context of this thesis are provided.

A critical review of the existing literature relating to these topics will be covered in chapters 2, 3, 4 and 5 rather than in a single literature review at the beginning of the

thesis.

1.2.1 Figurative Language

Figurative Language is a specific type of natural language that has been extensively studied and debated for centuries, resulting in a vast body of theories and definitions [198]. In addition to the longevity of research into figurative language, scholars from disciplines such as linguistics, philosophy, psychology, cognitive science and literary criticism have studied the phenomena [187, 198]. Identifying a single concise definition that adequately captures all aspects of figurative language is a challenging task given such a large body of diverse literature. For the purpose of this thesis, it is necessary to define figurative language by drawing from several modern definitions.

Figurative language is often defined in terms of a contrast with the phenomenon of literal language. The Oxford Dictionary defines figurative language as *"Departing from a literal use of words; metaphor"* [46]. A more detailed definition states that the intended meaning of a figurative utterance does not coincide with the literal meaning of the words and sentences contained within that particular utterance [63]. It is claimed that the recognition of figurative language is only possible because of a contrast with more literal language [45].

The provided definitions indicate that figurative language is a phenomenon that is understood in terms of literal language. This suggests that an understanding of literal language precedes that of figurative language. The Oxford Dictionary provides the following definition of literal language; *"Taking words in their usual or most basic sense without metaphor or exaggeration"* [46]. Slightly problematically, figurative language and literal language are defined in terms of each other by the Oxford Dictionary. A literal meaning can be more comprehensively defined as being whatever the dominant linguistic theory determines the meaning to be, based off a rule-by-rule interpretation of the utterance components [63, 235]. Following on from this definition of literal language, figurative language can be defined as a form language in which the rule-by-rule interpretation of the utterance components differs from what the author intends to convey [35], see Figure 1.1.

This deviation in meaning from the literal to figurative sense can be achieved in a myriad of ways, including through the use of metaphor, hyperbole, irony, personification, sarcasm, metonymy, idioms, analogy and many other literary devices [171, 187, 235]. The precise definition of these devices, boundaries between them and their importance

Figure 1.1: **Figurative Language Utterances**

Utterance, intended meaning and example sentence

have also been a source of debate between scholars from various disciplines for centuries [63].

This thesis begins with a focus on general figurative language, Part I, followed by an in-depth focus on hyperbole, Part II.

1.2.2 Online User-Generated Content and Social Media

The *social web* and *web 2.0* are terms that are used to denote the fundamental shift in the way users interact and contribute to the internet, from consumers of relatively static information to active producers and consumers of dynamic information [68, 154, 249]. Much like figurative language, there has been debate over the usage and meaning of these particular terms [154]. The semantics of these terms is outside the scope of this thesis. However, this fundamental shift in user interaction with the internet is a crucial event that has lead to the development of web-based platforms that facilitate social interaction and networking. These platforms are broadly referred to as social media [74, 249].

An important phenomenon to emerge from social media is 'searchable talk'. This refers to how discourse around a specific topic, geographical location, point in time or any combination of these may be found by searching the vast stores of data collected by social media platforms [249]. This data provides a snapshot of discourse that has no parallel at any period in history.

The low signal-to-noise ratio in discourse on social media is a problem given the sheer amount of content generated by users as well as user anonymity and lack of moderation in some spaces [50]. In order to '*cut though the noise*' and achieve virility users resort to

various strategies for generating content that catches the attention of users [15, 207]. One such strategy for user-generated text content is the use of novel figurative language to amuse users and gain attention [167]. Through the use of figurative language, authors can adopt personas, impart a sense of vividness to their content or pique the interest of the audience by crafting complex scenarios that require effort on behalf of the audience to interpret [73].

The effectiveness of figurative language in capturing the interest of users has led to the New York Times proclaiming a '*death by internet hyperbole*' [14], the Independent arguing '*How Hyperbole 'won the internet"* [77] and the Guardian claiming that '*Exaggeration is the official language of the internet*' [21].

A core focus of this thesis is the use of figurative language on social media platforms (i.e., Twitter), looking at the expression, intentions and diversity of figuration.

1.2.3 Natural Language Processing

Natural Language Processing (NLP) is an interdisciplinary field with the goal of computationally performing tasks involving natural human languages [95, 131]. The field of NLP has been an area of fervent academic research and commercial applications for several decades. As a result, NLP has grown into a broad field with sub-fields that cover domains and applications which may be categorised differently (see Table 1.1[40] for one such categorisation of NLP applications).

Most relevant to this thesis is the *Text analytics* and *Natural Language Generation* application domains. Particularly the problem of automated classification of text into pre-defined categories within the text analytics domain, commonly referred to as text classification or text categorisation [44]. With respect to natural language generation, this thesis looks at the generation of natural language as a means of interpreting figurative language.

The granularity of text classification tasks range from the classification of an entire document, a paragraph, a sentence or sub-sentence fragments to one or more pre-defined categories. In addition to this variation in granularity, there is also considerable variation in the domains that have seen successful application of text classification models. All the possible variations and applications of text classification methods has resulted in considerable breadth of datasets and methodologies, a single review that covers all these resources and related works is an enormous task and outside the scope of this thesis (see [4, 6, 105, 253] for details).

Application	Description
Machine translation	The automated translation of natural language content from a source language to a target language whilst preserving meaning of source text.
Speech technologies	The conversion of an audio signal of a linguistic utterance into textual form, referred to as Speech Recognition, or the conversion of the textual form of an utterance into an audio signal (Speech Synthesis or Text-to-Speech).
Dialog interfaces	Interactive interfaces that allow a user to communicate with an automated system using natural language and receive responses in natural language
Text analytics	The identification of particular content in text data, typically via the classification of text sources or via the extraction of certain content from text sources.
Natural language generation	The automated generation of linguistic content.
Writing assistance	Automated systems that embellish, provide suggestions or corrections to human generated text.

Table 1.1: **The landscape of NLP**

[40]

This thesis focuses on the text classification problem in the context of figurative language. The analysis of figurative language by NLP systems is a considerable challenge and has been described as one of the most arduous topics confronting researchers in the field [180, 187]. Despite significant advancements in text classification and NLP in recent years, the ability to adequately detect and comprehend figurative language is lacking [3, 180, 187]. A key focus of this thesis is the challenges in processing figurative language in online user-generated content using NLP techniques.

In addition to a focus on the text analytics domain, the thesis also deals with the application domain of natural language generation. The task of interpreting figurative language has been formulated as a mono-lingual machine translation task (i.e., paraphrase generation) [18, 164, 203, 206]. However, the work in this area is scant and will be expanded upon in this thesis.

Style Transfer (ST) is another mono-lingual machine translation task related to some of the generation methods described in this section. Many different transfer tasks, associated datasets and models have been proposed. Such as the transfer of texts from informal to formal English [183], the transfer of texts in to Shakespearean style [244] and the transfer of product reviews from positive to negative amongst several others [117]. The transfer of text from figurative to literal is an application of ST that is explored

in this thesis.

A core focus of this thesis is the evaluation of existing as well as the proposal, implementation and evaluation of new NLP models for various tasks related to understanding figurative language.

1.3 Research Questions

In order to address the core statement underpinning this thesis, *Accurate computational understanding of figurative language on social media is a complex task that requires modification of existing and creation of new datasets and methodologies.*, three research questions have been devised:

i. How does figurative language occur on social media and how does this differ in comparison to the occurrence of figurative language in traditional forms of communication?

ii. How adequate are current resources (i.e., datasets, models) for the accurate detection and interpretation of figurative utterances found on social media?

iii. How can the computational detection and interpretation of figurative utterances be improved?

1.4 Research Objectives

Several artifacts will be produced as a result of answering the three research questions outline in Section 1.3. The production of these artifacts are considered objectives of this thesis and are as follows:

i. Create annotated datasets that enable the study of figurative language on social media.

ii. Quantify the phenomenon of figurative language on social media and how this impacts the predictive performance of existing NLP text classification models.

iii. Develop and evaluate machine learning algorithms for the task of figurative language understanding on social media

1.5 Thesis Contributions

This section will outline contributions made towards answering the research questions and addressing both research objectives and gaps in the literature. These contributions will be presented based on the chapter of the thesis in which the contribution is described in detail.

1.5.1 Figurative Language and Public Health

The design, collection, annotation and analysis of a dataset consisting of Twitter posts mentioning health related content is introduced in Chapter 2. This content provides several contributions:

- This is the first data resource to provide both data and annotations to analyse the relationship between health related concepts and figurative language usage on Twitter.
- Analysis of this dataset provided quantitative evidence of figurative language in the context of symptom and disease words on Twitter
- Findings showed that figurative language occurs frequently on social media in the context of symptom and disease words, importantly it was observed that some symptom and disease words were more likely to be used in a figurative sense than in a literal sense.
- Experiments were designed to quantify the impact of figurative language on public health applications that rely on mentions of disease and symptom words.
- Experiment findings showed that the majority of false positives were a result of figurative expressions of disease and symptom words. This result provided evidence for the need to address this bias caused by figurative expressions of disease and symptom words.
- It was observed that a particularly challenging feature of the figurative language usage on Twitter was hyperbolic expressions.

1.5.2 Hyperbole

The design, collection, annotation and analysis of a dataset consisting of Twitter posts as well as the design, generation and annotation of a probing suite is introduced in Chapter

3. A number of contributions emerged from the content introduced in this part of the thesis:

- This is the first data resource containing data and annotations of hyperbolic expressions on Twitter.
- Analysis of this data provided quantitative evidence of hyperbolic expressions on Twitter.
- A significantly greater prevalence of hyperbole was observed on Twitter compared to that found in corpus studies on hyperbole in different communicative forms (i.e., conversational English).
- With respect to the intentions of figurative language, Hyperbole was commonly used on Twitter to convey strong sentiment, which highlights the importance of understanding hyperbole for computational tasks concerned with identifying affective content in text (i.e., sentiment analysis).
- A detailed look at the diversity of hyperbole on Twitter showed that some hyperbolic expression were simply parroted by different Twitter users but also a number of novel, elaborate and specific hyperbole were constructed by Twitter users. This finding indicated that adapting to novel hyperbole expressions is a key challenge in computational analysis and understanding of hyperbole.

1.5.3 Towards Hyperbole Detection

The evaluation of existing methods for hyperbole detection, and the introduction of several novel methodologies, in Chapter 4 indicated that:

- Hyperbolic language on social media is a challenging phenomena. It was observed that hyperbole on Twitter was harder to accurately detect compared to idiomatic hyperbole suggesting that current NLP methodologies were inadequate for the task, especially detecting hyperbole on Twitter.
- The use of affective signals was combined in various ways in hyperbole detection models showing some improvement on the task of hyperbole detection.
- A novel hyperbole detection methodology optimised via contrastive-loss and pre-trained language modelling showed improved performance of the task of hyperbole detection.

- Detailed error analysis provides a foundation for future research on the challenging task of hyperbole detection in social media text.

1.5.4 Towards Hyperbole Interpretation

The evaluation of various natural language generation models for hyperbole interpretation in Chapter 5 showed that:

- The first dataset of parallel hyperbolic Tweets and manually composed de-hyperbolised Tweets was used to pose the hyperbole interpretation problem as mono-linguistic machine translation problem.

- The automatic generation of literal interpretations is a challenging task. It was observed that generic paraphrases do not adequately interpret the hyperbolic content present in an expression.

- Models trained for neutralizing of subjective bias, a similar task to hyperbole interpretation, do not adequately remove hyperbolic content.

- Detailed error analysis provides a foundation for future research on the challenging task of automated hyperbole interpretation.

1.6 Thesis Structure

The main content of this thesis is presented in three parts.

- **Part I**, is on the figurative expression of health related concepts on Twitter and how this can impact public health applications that rely on NLP methods that use Twitter as a data source.

- **Part II**, focuses on general expressions of a particular form of figurative language (i.e., hyperbole) on Twitter and general idiomatic usage. The introduction of empirical studies quantify how these expressions are realised and probes NLP models on their detection and interpretation of hyperbole.

- **Part III**, synthesises the results from the first two parts of the thesis, discusses overall findings and lays the platform for future research directions.

The literature review and background specific to the topics covered in a particular chapter are presented at the beginning of that chapter. This is instead of a large literature review at the beginning of the thesis that covers the diverse range of topics covered throughout the thesis.

Part I

Figurative Language and Public Health

CHAPTER 2

FIGURATIVE LANGUAGE AND HEALTH MENTION CLASSIFICATION

2.1 Introduction

This chapter details a study on the figurative expression of health concepts (i.e., symptom and disease words) on Twitter and how these expressions impact public health applications that rely on Twitter for input data.

The content in this chapter addresses the research questions and objectives outlined in Section 1.3 in the following ways:

i. Tweets related to health concepts are collected and annotated for the presence of figurative language. This provides data to understand *how figurative language occurs on social media*, in the context of public health on the Twitter platform. (Research Question i, Research Objective i)

ii. Experiments are conducted on text classifiers trained for the detection of health related content with a *focus on the impact of figurative language on the accuracy of these NLP classifiers*. (Research Question ii, Research Objective ii)

iii. A text classification model is proposed and evaluated for detecting health events that target figurative expressions. This is to understand *how improvements can*

be made to increase the accuracy of figurative language detection in the context of classifying health events on Twitter. (Research Question iii, Research Objective iii)

The content in this chapter is structured as follows;

- Section 2.2 details the connection between figurative language, health and public health applications as well as motivations for such an empirical study.
- Section 2.3 details the collection and annotation of **HMC2019**, an English language Twitter dataset that contains annotated tweets relating to various symptoms and diseases
- Section 2.4 describes the design of experiments that aim to quantify the impact of figurative language on classifiers trained to detect health mentions.
- Section 2.5 details experiments on different techniques for computing word representations and the impacts on the detection of figurative usage of health concepts.
- Section 2.6 proposes a methodology, and accompanying experiments, that focus on capturing figurative health mentions to correct the bias revealed in experiments from Section 2.4
- Section 2.7 discusses the results of experiments described in Section 2.6.
- Section 2.8 details a manual error analysis that focuses on classification errors
- Section 2.9 concludes the chapter

2.2 Figurative Language and Public Health

Figurative language is ubiquitous across various registers of communication [140] and is considered by some scholars to be fundamental to our conceptualisation of the world around us [108]. Several researchers have documented the prevalence of figurative language in various registers of communication. Ironic figures of speech were found in 8% of all conversational turns in an analysis of spoken conversations between friends [60]. Corpus studies revealed that metaphorical expressions occurred in every third sentence of general-domain text on average [204]. Turning to online user-generated text, a study of online debate forums found sarcasm in 12% of utterances [232].

The use of figurative language in healthcare communication has also been well documented and explored [23, 42, 61, 70, 76, 146, 220]. Patients use figurative language

to help describe the at-times abstract emotions and sensations they experience during illness. Practitioners and researchers use figurative language to establish a stronger connection with individual patients and the broader community.

The prolific use of various metaphors by cancer patients has been observed in various settings in different languages [70, 76, 199, 200]. Two common metaphors expressed by cancer patients involve relating their experience with the disease via the *Battle* (e.g., 'fighting a *battle* against cancer') and *Journey* ('cancer is another obstacle along the journey') metaphors. The Metaphor in end-of-life Care (MELC) project provides analysis of online discussion forums of patients sharing their experiences living with late stage cancer in the United Kingdom [201]. The authors identified strong usage of these metaphors to describe their experiences and comes to terms with their illness. An analysis of blog posts by Swedish patients with advanced cancer also identified strong usage of these metaphors to express their emotions as they progressed through their illness [76]. The authors also identified what they refer to as the *Imprisonment* and *Burden* (e.g., carry, lift, heavy, weight, etc.) metaphors were commonly employed by individuals. Analysis of a Spanish language forum for cancer patients and survivors identified prolific use of '*Violence*' (e.g., battle, fight, war, warrior, soldier) and '*Journey*' metaphors to describe their individual experiences with cancer [129].

Figurative language is a common means for communicating the sensation of both acute and chronic pain [5, 82, 146, 147, 197]. Given the lack of an adequate objective means to express pain and to measure pain, verbal expression of pain is an important method of communication. Individuals resort to metaphor and describe pain in terms of concrete experiences of physical damage (e.g., 'a *stabbing* sensation', 'a *burning* pain', 'an ear *piercing* sound', 'muscles *seizing* up', etc.) [197]. In a survey of women with endometriosis it was observed that elaborate metaphorical descriptions were often resorted to as a means for expressing pain [23]. A study of the textual descriptions of the personal experience of various types of chronic pain revealed prolific use of metaphor. Such as elaborately describing their pain with experience of physical damage (e.g., '*like I'm being hit with a sledge hammer every minute of the day*', '*like driving a knife into my bones and muscles and twisting it*') [146].

An important tool for intervention of mental health conditions is psychotherapy. The abstract nature of mental health experiences means that the expression of emotions, states and self-image can be difficult to accurately describe. The use of figurative language to help these illuminate experiences has been well documented [56, 173, 198, 220]. Therapists have noted the importance and positive impact of figurative language usage

by clients in the therapeutic process. Formal protocols and models have been devised for the purpose of identifying, affirming and elaborating on the metaphorical expressions used by clients to describe their personal experiences with mental health [103, 208, 220].

Despite considerable focus on the relationship between health and figurative language, there is little focus on the figurative usage of health concepts to convey emotions and opinions in the absence of actual personal health experience (i.e, 'My kids give me a *headache*', 'this video gave me *eye cancer*', 'I have baby *fever* really bad right now'). One such analysis found that 30% of all tweets containing the keyword *'ebola'* or *#ebola* were deemed to be sarcastic in nature in an analysis by [152]. Whether this tendency towards sarcasm in tweets is particular to the disease Ebola or whether it is a common trend in tweets about other diseases has not been addressed and is a point of investigation in Part I of this thesis.

2.2.1 Public Health Surveillance

Public Health Surveillance (PHS) is defined by the World Health Organization as 'the continuous, systematic collection, analysis and interpretation of health-related data needed for the planning, implementation, and evaluation of public health practice' [155].

Traditional systems for PHS rely on the collection and analysis of structured data to calculate various rates of disease, such as incidence, burden and seasonality [230]. Depending on the specific system this data may come from health care providers, voluntary reports, diagnostic laboratories or other sources. This body of structured data is continuously analysed to identify changes that warrant intervention on behalf of public health agencies. The collection and aggregation of all the information from various sources has been identified as a significant bottleneck and imposes limitations on the speed at which PHS systems can identify and respond to emerging health events [20, 230].

Harvesting data from the internet in the form of news articles, search engine queries and various social media applications has the potential to circumvent this slow proliferation of information through traditional channels and provides opportunities for near real-time PHS systems [20, 88, 230]. A bottleneck when using these data sources is the complexity of the data, (i.e., volume, velocity, veracity, variety of data) which requires new tools and methodologies to accurately correct, store, aggregate and interpret the data.

Social media data (i.e., Twitter posts, Facebook posts etc.) has been collected and analysed for various public health applications [88, 90, 161, 209]. From monitoring disease incidence and outbreaks [20, 71, 87, 231], detecting adverse reactions to drugs

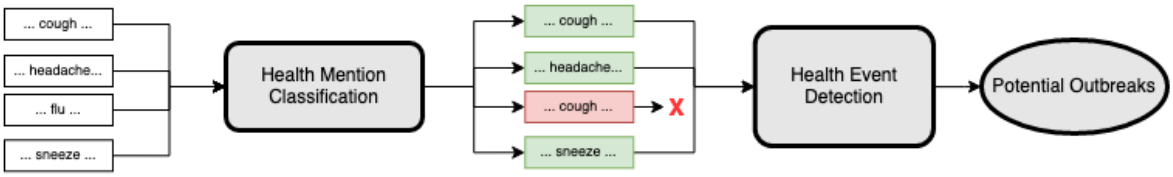

Figure 2.1: **Typical Epidemic Intelligence Pipeline**

[62, 151, 195], detection of mental health content through social media posts [38, 41, 57] and several others.

2.2.2 Epidemic Intelligence

Epidemic Intelligence systems are built to detect anomalous health events and issue early warnings for potential public health emergencies [22, 90, 158]. Much like other PHS systems, the type and source of data that these systems rely on is diverse and changing as new technologies emerge. Most important to this thesis are text-based Epidemic Intelligence systems that rely on social media as their main source of data. The typical framework for such systems involves two key steps incorporating NLP and time series analysis techniques to detect anomalous health events, see Figure 2.1 [90].

The detection of influenza and influenza like illnesses on social media are a common application of Epidemic Intelligence systems [88, 90]. One such early work describes the implementation of an algorithmic approach for the detection of tweets relating to influenza and provides a comparison of their results with official influenza rates across a singular flu season in the United States [20]. The authors found that the predictions of their system correlated with official rates of influenza provided by government health agencies, the correlations were statistically significant [20]. Along these lines, three years worth of influenza related tweets were used to predict the expected number of influenza patients in Japan [231]. The authors found that their predictions improved upon a simple baseline and showed strong performance across both rural and urban areas of Japan. The effectiveness of using Twitter signals to predict rates of disease incidence has also been shown on a smaller geographical scale. A statistical significant correlation was found between the amount of Tweets mentioning terms relating to influenza (i.e., headache, flu, coughing) and clinical data at a paediatric hospital in the United States [71]. Only tweets that were geo-tagged with locations that fell within the counties serviced by the hospital were considered in their analysis showing that Twitter can be an effective data source for predicting influenza occurrence on a local scale.

A common approach for the collection of data for Epidemic Intelligence systems is via keyword search, in some cases the keywords may simply be the disease or symptom terms [62, 98] or a set of terms enriched by medical ontologies [34, 37], topic models [29, 160] or other ad-hoc techniques. Irrespective of the particular strategy for generating keywords, the collection of tweets based on a limited set of keywords results in biased datasets with low precision (i.e. many false positives). The bias in these datasets may be introduced by an increased awareness of a particular disease outbreak (i.e., influenza). An increase in awareness of the disease results in an increase in tweets about the disease ('Hopefully I don't catch the *flu* this season'). These tweets not necessarily talking about the presence of the disease leading to an overestimation of disease prevalence. [20]. Overestimation bias of influenza incidence was observed on Google Flu Trends and is speculated to be a contributor to the discontinuation of the service [96, 111].

An important first step in any Epidemic Intelligence system that relies on informal sources of data is the detection of content related to health events, which could be in the form of a specific disease or symptom [90]. As previously mentioned keyword search produces low-precision datasets that captures mentions of health keywords in non-health contexts. Given the sheer volume of informal data generated on the web the task of filtering out those non-health related mentions is considerably important and intractable for manual work, as a result automated systems are crucial [90, 98]. This task has been presented as a binary sequence classification problem where the task classify if a sequence of text (such as a social media post) reports a health event [83, 86, 90, 97, 98]. This task is referred to in the remainder of this chapter as *Health Mention Classification* (HMC), see Figure 2.1. Statistical learning techniques have been proposed as a solution to the problem of differentiating between health keywords in health and non-health related contexts [90].

A method based on Long Short Term Memory (LSTM) [75] network and randomly initialized word embeddings was proposed to detect whether a tweet is related to a personal health experience or not (i.e., HMC) [85, 86]. The authors show that their method outperforms approaches that utilise hand-crafted features and traditional statistical learning methods (e.g., Logistic Regression, Decision Trees, Support Vector Machine, K-Nearest Neighbour) and suggest end-to-end deep learning architectures as a promising direction for addressing the bias introduced by non-health related mentions.

A classification framework consisting of a Logistic Regression classifier trained on pre-trained word representations was more recently proposed to address the HMC task [98]. Significant focus of this work was the distorting and partitioning of pre-trained

word representations. The distortion and partitioning operations, in combination with manually engineered syntactical features, outperforms several other methods on the HMC task. This method outperforms both traditional statistical learning methods with hand-crafted features (i.e., Logistic Regression, Rule-based classifier [109]), and end-to-end deep learning models (i.e., Convolutional Neural Network (CNN), LSTM-GRNN [218] and FastText [94]), and performs particularly well with limited data.

A more recent approach incorporated features from an unsupervised model for the detection of idiomatic utterances [120], into a classifier based on CNNs [83]. Experiments, indicated that their proposed approach resulted in an increase across a number of information retrieval metrics on a benchmark HMC dataset [98]. Despite motivating figurative language and HMC, the authors do not objectively measure for an improvement on tweets with figurative language, only the overall F1 score for the task, it is not clear how well their approach actually targets figurative language.

The mentioning of health concepts in a figurative sense has been claimed as a potential source of overestimation bias that has received scant attention [83, 98]. Words indicating symptoms and illnesses may be mentioned in figurative statements. Take for example the following tweets; *'The language called english they used in this text can cause one* ***Migraine****,* ***Dermatitis*** *&* ***Photophobia*** *altogether'*, *'Why is Jar Jar Binks trending? Is Twitter having a* ***stroke****?'*. These tweets are examples of hyperbolic statements that use disease words for purposes of exaggeration rather than to convey that any one is experiencing a health event related to the particular disease words. These examples indicate the bias introduced by figurative language that may pose challenges. This bias results in a signal that is an over-estimation of the prevalence of a particular health keyword. However, there is scant research on the impact of figurative language on the HMC task.

A recent benchmark for HMC, [98], contains limited coverage of health concepts. The dataset introduced in that paper, PHM2017[1], only focuses on diseases not symptoms, a limitation to Epidemic Intelligence. Another limitation is the small number of diseases covered and the type of diseases covered. Firstly, Alzheimer's Disease and Parkinson's Disease are degenerative diseases that predominantly impact the older population, many of whom are not avid users of social media. The choice of stroke and heart attack are also interesting, the utility of a system that monitors social media for instances of these severe and life threatening health events is not that high. An individual suffering one of these events is not likely to post on social media before contacting emergency services.

[1]https://github.com/emory-irlab/PHM2017

The expansion of this dataset to incorporate symptoms and annotated for figurative language is a focus of this chapter. As is the impact of figurative language on models for HMC given that the analysis of figurative language by computational means has proven challenging to classification problems for tasks in sentiment analysis, machine translation among other tasks [89, 187, 204, 233].

Keyword	Count	FP	HMP
Alzheimer's	1,924	0.070	0.143
Cancer	1,995	0.101	0.175
Cough	1,976	0.487	0.222
Depression	1,971	0.242	0.342
Fever	1,987	0.438	0.358
Headache	1,961	0.374	0.552
Heart attack	1,987	0.663	0.123
Migraine	1,964	0.147	0.617
Parkinson's	1,810	0.043	0.097
Stroke	1,983	0.282	0.147
Totals	19,558	-	-
Means	1955.8	0.285	0.278

Table 2.1: **HMC2019 Statistics**

Keyword is the keyword mentioned in tweet, ***Count*** is the number of tweets, ***FP*** indicates the proportion of tweets that mention the keyword in a figurative sense. ***HMP*** indicates the proportion of tweets that mention the keyword in a health context.

2.3 HMC2019

HMC2019[2] is built upon the foundations of an existing English language Twitter dataset for HMC that covers six different keywords related to particular diseases (i.e., *Alzheimer's disease*, *cancer*, *depression*, *stroke*, *heart attack* and *Parkinson's disease*) and contains approximately 7 thousand labeled examples [98]. Twitter is an ideal source of data for Epidemic Intelligence systems given the content in an individual Tweet, the widespread global usage of the platform and the real-time nature of the data makes Twitter an ideal source of data for Epidemic Intelligence systems[98]. To extend the relevance of this dataset for Epidemic Intelligence systems additional health related keywords are used as query terms for the Twitter API [3]. Four new keywords (i.e., *cough*, *fever*, *headache*,

[2]https://github.com/biddle-r/HMC2019

[3]https://developer.twitter.com/en/docs/twitter-api

Keyword	Class	Tweet
Alzheimer	HM	Sorry I'm MIA, I'm dealing w/ some very grim family news about my gpa w/ bad **Alzheimer's**.
	FM	You are delirious, either that or you evidently have **Alzheimer's** Disease which makes you mentally ill to serve as President...
	NHM	Metabolomic-guided Discovery of **Alzheimer's Disease** Biomarkers from Body Fluid
Cancer	HM	A3 Three years ago when I was diagnosed with **cancer**, it gave me a learning experience I wouldn't have had otherwise. #weirded
	FM	There'll never be peace in our country until @CBS looses it broadcasting license. They are a **cancer** to a civil society.
	NHM	Breast **cancer** prediction model developed for Hispanic women
Cough	HM	I need home remedies for a dry **cough** pls and thank you.
	FM	I will, without fail, return this debt that- *__cough__* *__cough__*
	NHM	**Cough** And Cold: Causes And Remedies
Depression	HM	I'm so sick of being sad for no reason (like yeah, **depression** is like that I get it) but could I just be sad about regular things like girls or money or something?
	FM	Twitter is basically **depression** hour but it's 24/7
	NHM	People who complain online are more likely to suffer from anxiety, **depression**, and stress.
Fever	HM	So, my brother has **fever** and currently bed-ridden. He calls me up asking, "Is it okay to take a bath?"
	FM	Why do I have baby **fever** rn?
	NHM	You now need your yellow **fever** card to travel to the UAE.
Headache	HM	I've had this **headache** for more than 6 hours now wow.
	FM	my **headache** is bigger than loona's discography
	NHM	**Headache** in your Jaw? What is causing it? https://t.co/YiJFrujRdl
Heart Attack	HM	I have not intentionally been a little cryptic over the past few days but I just wanted to clear things up. On Thursday morning July 18 I suffered a **heart attack**.
	FM	You who burned the meatballs and gave the entire building a **heart attack** I aksdkskkak I got a cut on my finger bc of you
	NHM	Evolutionary Gene Loss May Help Explain Why Only Humans Are Prone To **Heart Attack**
Migraine	HM	My body hurts and I have a **migraine**
	FM	The language called english they used in this text can cause one **Migraine**, Dermatitis & Photophobia altogether
	NHM	I just published If you feel challenged by **Migraine**, consider joining local research studies
Parkinson's	HM	Very hard trying to play board games with a grandad who's got **Parkinson's**
	FM	Y'all ever play Smash bros drunk?... it's like playing operation with **Parkinson's** lmaooo
	NHM	My story about fantastic dance classes run by @Balletboyz to help people with **Parkinson's disease**
Stroke	HM	@CharlesMBlow My dad died in his sleep of a **stroke** at 53.
	FM	Why is Jar Jar Binks trending? Is twitter having a **stroke**?
	NHM	saw a beautiful husky earlier, went to **stroke** it and it wasn't even fazed and walked off

Table 2.2: **HMC2019 Examples**

Keyword is the health keyword mentioned in tweet. ***Class*** indicates tweet label, HM = health mention, FM = figurative mention, NHM = non-health mention ***Tweet*** displays the tweet, keyword word in **emphasis.**

migraine) are added to the set of query terms more than twelve thousand new tweets are collected for the dataset. Tweets were collected during July and August 2019.

In Table 2.2 examples of tweets containing figurative mentions and health mentions of keywords as labelled during manual annotations are shown. Two annotators, native English language speakers, were given the dictionary definitions of figurative and literal language and asked to annotate a tweet into one of three categories. One class (NHM) indicated that the keyword was used in a non-health related context and was used in a literal sense (e.g., '*How to cut your risk of **Heart Attack** in half*'). Another class (HM) indicated that the keyword was used to indicate a health event (e.g., '*my **headache** is getting worse*'). Another class (FM) indicated that health keyword was used in a figurative sense (e.g., '*this guy is a literal **cancer** on my soul*'). The inter-annotator agreement was high, with a Cohen's kappa of 0.87. This annotation scheme differs from **PHM2017** and other datasets for HMC in that it a label for figurative mentions of health and disease words is added. It is important to note that whilst this class is a type of non-health related tweets, and a subset of the NHM class. This label provides a means to produce quantitative evidence of the extent and impact of figurative language on the HMC task.

Statistics relating to the **HMC2019** are shown in Table 2.1. From this table, it can be see that on average keywords are mentioned figuratively 28.5% of the time whilst being mentioned as an actual health event 27.8% of the time. This indicates that on average the health related keywords in **HMC2019** are mentioned figuratively in a similar frequency to which they are mentioned as actual health events. This suggest that Epidemic Intelligence systems that rely on mentions of symptoms and diseases on Twitter to track health events may be impacted by the bias introduced by figurative language. Without filtering out these figurative mentions these systems are getting a biased signal from the Twitter stream due to the considerable amount of figurative usage of particular health related concepts.

It is worth noting that ratio between health mentions and figurative mentions is not consistent across keywords further complicating the issue. *Heart attack* (5.4 : 1), *Cough* (2.2 : 1), *Fever* (1.2 : 1) and *Stroke* (1.9 : 1) have figurative mention to health mention ratios of greater inequality, indicating that they are more frequently mentioned in figurative expressions than as actual health mentions. Whilst *Alzheimer's* (0.5 : 1), *Cancer* (0.6 : 1), *Depression* (0.7 : 1), *Headache* (0.7 : 1), *Migraine* (0.2 : 1) and *Parkinson's* (0.4 : 1) all have figurative mention to health mention ratios of lesser inequality, indicating that they are more frequently mentioned as actual health events then in a figurative expression.

2.4 Health Mention Classification and Figurative Language Bias

The aim of this section is to further understand the relationship between figurative language and HMC task by exploring the impact of figurative mentions on the accuracy of models trained for HMC. Prior research has identified that these figurative mentions are potentially detrimental to accuracy of HMC models [83, 90, 98]. One of these works strongly alludes to the potential negative impacts of figurative mentions however, does not quantify these impacts or target figurative mentions in their methodology [98]. Conversely, the methodological contributions in one of these works is motivated by the problem of figurative mentions of health related keywords and directly targets these mentions in their model [83]. However, this work does not explicitly show that the improvements in accurate classification of health mentions is a result of better figurative mention classification or just a byproduct of adding another complex module to their model.

Keyword	**F1**	**P**	**R**	$\mathbf{b}_f$	**HMP**	**FMP**
Alzheimer's	0.658	0.544	0.844	0.346	0.143	0.070
Cancer	0.626	0.499	0.849	0.141	0.175	0.101
Cough	0.728	0.594	0.943	0.534	0.222	0.487
Depression	0.705	0.562	0.950	0.512	0.342	0.242
Fever	0.769	0.638	0.970	0.724	0.358	0.438
Headache	0.808	0.682	0.994	0.797	0.552	0.374
Heart attack	0.455	0.325	0.777	0.884	0.123	0.663
Migraine	0.860	0.758	0.995	0.415	0.617	0.147
Parkinson's	0.621	0.506	0.816	0.172	0.097	0.043
Stroke	0.664	0.525	0.915	0.539	0.147	0.282
Total	**0.278**	**0.285**	**0.689**	**0.563**	**0.905**	**0.506**

Table 2.3: **Baseline HMC results**

Keyword is the keyword mentioned in tweet. ***F1*** is F1 score, ***P*** is precision score, ***R*** is the recall score, $\mathbf{b}_f$ is the proportion of false positives that are figurative mentions, Note: All metrics averaged over 10-folds. ***HMP*** shows the proportion of tweets that mention the keyword in health context. ***FMP*** shows the proportion of tweets that mention the keyword in the figurative context.

Simple classification experiments are designed to quantify the impacts of figurative language on classifiers trained for HMC. Specifically, an analysis of the relationship between predictive performance and proportion of figurative language using a traditional

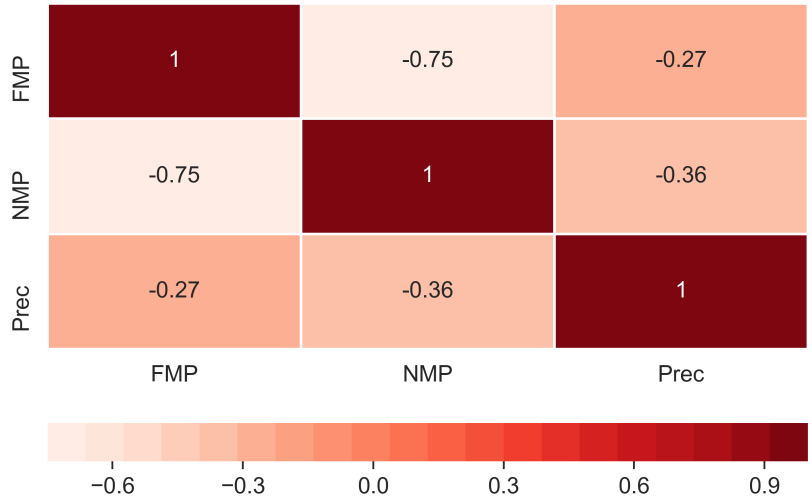

Figure 2.2: **Correlation Matrix Heatmap**

Matrix shows Pearson's correlation coefficient. ***FMP*** is proportion of figurative mentions, ***NMP*** is the proportion literal non-health mentions, ***Prec*** is the precision metric.

baseline classifier. The traditional classifier, consists of a Logistic Regression classifier trained on the **HMC2019** dataset with each tweet being represented by a TF-IDF weighted vector of unigrams and bigrams [130]. A standard 10-fold cross validation scheme is followed to ensure results are repeatable an not a result of randomness. A number of information retrieval metrics such as F1 score, precision and recall to represent the accuracy of the classifier, these metrics are standard evaluation metrics used in empirical studies on HMC. [90, 130]. See equations 2.1, 2.2, 2.3 for metric formulations, where fp stands for false positive, fn for false negative, tp for true positive and tn for true negative respectively. Additionally, a custom metric is proposed to measure the bias of figurative language on the HMC task, see eq. 2.4 where fp_f indicates the number of false positives that are also labelled as figurative mention tweets (FM). As mentioned in Section 2.2, datasets created via keyword sampling are low precision datasets thus improving precision (i.e., lowering false positive rate) is a key step to better health mention classifiers. The metric proposed by the author, b_f, indicates the bias introduced by figurative mentions and quantifies the impact these mentions have on the false positive rate and precision.

$$P = \frac{tp}{tp + fp} \tag{2.1}$$

$$R = \frac{tp}{tp + fn} \tag{2.2}$$

$$F1 = 2 * \frac{P * R}{P + R} \tag{2.3}$$

$$b_f = \frac{fp_f}{fp} \tag{2.4}$$

The results from these experiments are presented in Table 2.3. From the b_f metric it can be observed that figurative mentions make up the majority (50.6%) of false positives across the entire dataset. The extent to which figurative mentions are incorrectly identified as health mentions varies considerably between health keyword; 14.1% for *Cancer* to 88.5% for *Heart Attack*. The results from this table indicate figurative language is a source of false positives. However, false positives are just one term in calculating the precision metric (see eq 2.1). From a focus on the precision metric in Table 2.3, it can be seen that *Heart Attack* is often mentioned figuratively (66.3%) and that the traditional classifier achieves the lowest precision score of 0.325. Conversely, *Migraine* has a lower than average rate of figurative mentions (14.7%) and the traditional classifier achieves the highest precision score of 0.758.

Correlation analysis is performed to gain more insight into the relationship between precision and figurative mentions of health keywords. Formally, Pearson's correlation coefficient is computed using three variables: (i) the precision score (Prec) (ii) The proportion of tweets with a figurative health mention (FMP) and (iii) the proportion of tweets that mentions the health keyword in neither figurative nor health-related context (NMP). These correlations are visualized via heat map, (see, Figure 2.2). Results show a minor negative correlation, -0.27, between the proportion of figurative mentions and precision. There is a stronger negative correlation, -0.36, between mentions that are neither figurative nor health related and precision. This suggests that while figurative mentions of disease and symptom words are challenging for HMC classification the non health mentions of disease and symptom words are also challenging.

The experiments showed that the majority of false positives were a result of figurative expressions of health keywords. Additionally, considerable variation in usage of keywords was observed (i.e., heart attack often used figuratively, Parkinson's disease rarely used figuratively). Analysis of correlation showed that when the rate at which a keyword was mentioned figuratively increased that the precision of the HMC classifier decreased. The presence of this negative correlation allows us to conclude that figurative language is contributing to a biased signal. This motivates for a focused effort on reducing this bias via better detection of figurative mentions of health concepts.

2.5 Word Representations for Health Mention Classification

$$R_F = \frac{tnfm}{fm} \tag{2.5}$$

The main goal of the experiments detailed in this section is to empirically test the impact of word representations on the task of HMC. The motivation for this question is research that shows variation in the kind of information encoded in different layers of different word representations [123, 170]. To achieve this goal, classification experiments are performed on **HMC2019** using a variety of different word representations. The experimental setup is similar to that presented in Section 2.4. However, a new metric (see eq. 2.5) is introduced for these particular experiments that quantifies the ability for a classifier to correctly identify figurative mentions as *not* being a health mention. In this equation $tnfm$ is the number of true negative figurative mentions and fm is the number of figurative mentions, this metric is essentially Recall but for the figurative class only. Similar to the previous experiments, multiple information retrieval metrics are reported ($F1$, P, R, b_f), Logistic Regression is as classifier and 10-Fold cross-validation is performed.

For these experiments a variety of popular techniques for computing word representations are utilised. For contextual word representations, ELMo[168][4] and BERT[43][5], whilst for non-contextual word representations word2vec[138][6] and GloVe[166][7]. Given that experiments have shown variation in the information encoded in various layers of

[4]https://github.com/allenai/allennlp
[5]https://github.com/huggingface/transformers
[6]https://code.google.com/archive/p/word2vec/
[7]https://nlp.stanford.edu/projects/glove/

both ELMo and BERT [123, 170], individual layers as well as different combinations of these layers are experimented with as features in the following experiments.

A number of preprocessing steps were performed on all tweets before the computation of word representations;

- Convert all tweets to lowercase
- Remove all punctuation characters
- Emojis were converted to string representations such as ':sad_face:' using an open source python library[8].
- User mentions (i.e., @) are replaced with the token '_usr_'
- URLs (i.e., @) are replaced with the token '_url_'
- Digits (i.e., 0-9) are replaced with the token '_d_'
- # and the text suffix are removed
- Repeated characters and words were normalized to two repeats to address exaggerations such as 'looool' and 'lool', and repeated emojis.
- Tweets were split into individual tokens based on whitespace characters for word2vec and GloVE.
- Model-specific tokenizer implementations were used to perform tokenization for ELMo and BERT.

Results from experiments are shown in Table 2.4. Non-contextual word representations (word2vec, GloVe) are considerably worse across $F1$, P, and R_F than contextual representations (ELMo, BERT). Unsurprisingly, these results indicate that contextual word representations are better suited for the HMC task, aligning with studies indicating the benefits of contextual word representations [43, 123, 169]. With respect to the different layers and combinations of the contextual representations, it can be observed that there is not a significant difference between the summations of layers and the final layers. However, as more layers are included in the contextual word representations the performance steadily increase for both BERT and ELMo.

Figurative recall, R_F, is lower than general recall, R, regardless of word representations used. Contextual word representations reduce the difference between figurative recall and overall recall considerably compared to non-contextual word representations. There is a difference of 0.792, 0.419, 0.159 and 0.167 between recall and figurative recall

[8]https://pypi.org/project/emoji/

Representation	**F1**	**P**	**R**	**R_F**
GloVe	0.503	0.343	0.940	0.148
word2vec	*0.635*	*0.486*	*0.915*	*0.496*
ELMo_Layer_0	0.681	0.541	0.917	0.597
ELMo_Layer_1	0.746	0.627	0.920	0.719
ELMo_Layer_2	0.757	0.646	0.915	0.749
ELMo_SUM	*0.757*	*0.648*	*0.912*	*0.753*
BERT_Layer_0	0.713	0.583	0.920	0.661
BERT_Layer_1	0.722	0.594	0.921	0.672
BERT_Layer_2	0.725	0.599	0.920	0.679
BERT_Layer_3	0.730	0.605	0.920	0.694
BERT_Layer_4	0.743	0.620	0.927	0.708
BERT_Layer_5	0.746	0.626	0.925	0.709
BERT_Layer_6	0.743	0.624	0.919	0.714
BERT_Layer_7	0.753	0.636	0.925	0.722
BERT_Layer_8	0.761	0.646	0.927	0.738
BERT_Layer_9	0.767	0.653	0.929	0.748
BERT_Layer_10	0.767	0.654	0.927	0.744
BERT_Layer_11	0.768	0.653	0.934	0.741
BERT_SUM4	**0.768**	**0.658**	**0.924**	**0.757**
BERT_SUM8	0.763	0.651	0.923	0.748
BERT_SUM12	0.763	0.650	0.924	0.743

Table 2.4: **Word-Representations and Health Mention Classification**

Representation indicates the representations used as features. ***F1*** is F1 score, ***P*** is precision score, ***R*** is recall, $\boldsymbol{R_F}$ is the recall for figurative mentions. Note: All metrics averaged over 10 folds, Table is sorted via F1 from lowest to highest.

for GloVe, word2vec, ELMo and BERT respectively. This suggests that contextual word representations help identify figurative mentions. However, there is still a difference between figurative recall and overall recall, suggesting that figurative mentions of health words remain a challenging phenomenon.

2.6 Health Mention Classification

This section provides details on the baselines, methodology and experiments for the detection of health mentions on **HMC2019**.

2.6.1 Baselines

A simple model based on Long Short-Term Memory Networks (LSTMs) [75], was proposed to detect whether a tweet was indicating a personal health experience [86]. The authors detail a number of generic pre-processing steps before utilising pre-trained non-contextual word representations to represent the individual tokens contained in a tweet. The authors showed that this classifier outperformed several other classifiers, (i.e., Desicion Trees (DT), Support Vector Machines (SVM) and K-Nearest Neighbour (KNN)) on the task of detecting whether a tweet was indicating a personal health related experience or not. In addition to the improvements displayed by this model, the authors also claim that the benefits of their lightweight feature engineering steps. The authors use their own implementation of this method as a baseline in experiments and refer to it as $Jiang_{LSTM}$ in the remainder of this chapter.

A model for detecting personal health mentions on Twitter with considerably more detailed and complex feature engineering than the $Jiang_{LSTM}$ model was recently proposed [98]. The authors combine lexical, syntactic, word representation, and distorted+partitioned word representation features with a Logistic Regression classifier for health mention prediction, this model is referred to as **WESPAD**, throughout the remainder of the chapter. For lexical and syntactic features the authors build dependency tress for each Tweet using a well-known parser for the Twitter domain [102]. Word representation features were represented via *word2vec*[138][9]. The authors showed that the inclusion of distorted and partitioned word representation features was beneficial to the task of HMC. These features are described in detail as they are the key contribution to their model. Their technique is motivated by what they refer to as 'noisy' regions in the word representation space. The word representation space is partitioned into clusters, then 'noisy' regions are identified by training a classifier on the word representations of a tweet to predict if the tweet contains a personal health mention. Once these regions have been identified they are filtered out whilst label information is encoded for tweets in non-noisy regions. Formally, the authors partition the word representation space using the K-means clustering algorithm into k clusters. For the 'noisy' region identification the authors define a classification function, $f(t_i)$, to predict the probability that a tweet contains a personal health mention. The authors then construct two feature matrices, P and N, that indicate the class of a tweet if the tweet is not in a 'noisy' region and the cluster, k, the tweet belongs to (see eq. 2.6 and eq. 2.7). In these equations, α, is a hyperparamater that controls the threshold for noisy regions.

[9] Available at https://code.google.com/archive/p/word2vec/.

$$P_{ik} = \begin{cases} 1 & 0.5 + \alpha \leq f(t_i) \ \& \ t_i \in k \\ 0 & Otherwise \end{cases} \tag{2.6}$$

$$N_{ik} = \begin{cases} 1 & 0.5 - \alpha \geq f(t_i) \ \& \ t_i \in k \\ 0 & Otherwise \end{cases} \tag{2.7}$$

In addition to partitioning the word representation space, the authors also showed that distorting the representation space before partitioning had a considerable impact on the task of HMC. They use information gain to distort the word representation centroids in the context of the classification problem. Firstly the information gain IG_i of each word w_i is computed with respect to health mention and non-health mention labels for each tweet. A distorted tweet centroid representation dt_i for each tweet is then weighted using the information gain for each word, see Eq 2.8, where W_i is the word representation for word w_i.

$$dt_i = \frac{\sum_{i=0}^{n} IG_i * W_i}{\sum_{i=0}^{n} IG_i} \tag{2.8}$$

Another recent work directly addresses the problem of figurative language in the context of public health and HMC [83]. The authors rely on existing figurative language classifier ([120]) to predict the likelihood of figurative content in a tweet and incorporate this as a feature into a Convolutional Neural Network (CNN) based classifier. From their experiments, the authors concluded that their approach showed improvements, in information retrieval metrics, on a benchmark HMC dataset compared to a CNN trained on pre-trained word representations only. This method serves as a baseline and is referred to as **FeatAug+** for the remainder of this chapter. Despite improvements in predictive performance, the authors showed via an error analysis that **FeatAug+** still made errors with respective to the figurative usage of health keywords. Specifically, tweets that mentioned *heart attack* were often incorrectly classified, statistics from Table 2.1 showed that this keyword was mentioned figuratively most of the time.

2.6.2 BiLSTM+Senti

In this section, a model is introduced for detecting personal health mentions. **BiLSTM+Senti** is based on Bidirectional Long Short-term Memory Networks (BiLSTMs)

[66] that incorporates contextual word representations, see Section 2.5, and features that indicate the sentiment contained in a tweet. There are two key motivations in proposing this model; Firstly, context of a keyword is important in determining whether the tweet is indicating the occurrence of a health event, and that sentiment contained in a tweet may be also indicative of the class of the tweet. Annotators observed that context was essential in prescribing a class label to a particular tweet. This is intuitive but important for the HMC problem due to data collection strategy (i.e., keyword search). Additionally, the presence, location and intensity of sentiment was also utilised by annotators to help determine the class of a particular tweet. Consider the sentiment in these examples: *'Watching that video was like having eye **cancer**'* and *'**cancer** is **destroying** my sisters life in front my eyes, it is **devastating** to sit here and watch'*. In the initial example, there is a clear lack of sentiment in the context, the author is using the keyword to convey sentiment. This pattern was routinely observed during annotation. The latter example provides a stark contrast with respect to the sentiment in the context of the tweet, another pattern that was observed in honest health mentions.

BiLSTM+Senti text classifier is based on contextual word representations, Recurrent Neural Networks(RNN) and sentiment signals (see Figure 2.3). The model incorporates both word representations and distributions of sentiment to represent an individual tweet.

$$L,K,R = \begin{cases} L = w_i \forall w \in t | i < k \\ K = w_k \\ R = w_i \forall w \in t | i > k \end{cases} \tag{2.9}$$

A critical first step in the BiLSTM+Senti framework is preprocessing that follow those outlined in Section 2.5. In addition to these steps, a tweet partitioning step is performed that splits all tweets into three partitions. These partitions are based on the location of the keyword within the tweet; a partition containing the left context, a partition containing the keyword and and a single partition containing the right context (see eq.2.9). Where w_i represents word at index i in tweet t and k represents the position of the keyword in t. The motivation for this partitioning scheme is the desire to explicitly separate the context and the health keyword given the importance of context observed during manual annotation. This scheme may also capture tweets where there is a difference in sentiment between the partitions. A difference in sentiment within a sentence or short text has been shown to identify sarcastic intent [92]. For both the left

and right context tweet partitions, a sequence based on pre-trained word representations and a sentiment distribution is computed. The keyword partition is represented by the word representation and sentiment distribution for the keyword alone. The sequence of word representations for the left and right context were set to the largest possible sequence found in the data ensuring no tokens were lost; shorter sequences were padded to this length.

$$S = [x_1, x_2, x_3] \tag{2.10}$$

A sentiment distribution of a tweet is a vector of continuous values, $x_i \in [0, 1]$ with each value indicating a score computed to represent a particular sentiment signal, (see eq 2.10). There are a number of ways to compute sentiment signals from a text [134, 245, 251], three different approaches are utilised in this chapter.

$$x_p = \frac{1}{n} \sum_{k=1}^{n} p(w_k) \tag{2.11}$$

A simple approach is to use a pre-existing lexicon to lookup the scores for all the individual words in a tweet and average the scores for the whole tweet, (see eq. 2.11). Where n is the number of words, w_k is the word at index k and $p(w)$ is the polarity score of word t computed by lookup in the lexicon. The SentiWordnet lexicon is built by automated annotation of the syn-sets in WordNet [139][10] based on the amount of positive, neutral and negative sentiment contained within a syn-set [9][11]. The VAD lexicon[12] provides scores for valence, arousal and dominance rather than positive, neutral and negative like SentiWordnet. These values model different scales of affective meaning; valence models the pleasure/displeasure scale, arousal models the active/passive scale, and dominance models the dominant/submissive scale [143]. Several researchers have identified these scales as important dimensions for meaning [143, 234]. Sentiment distributions computed using both the SentiWordnet and VAD lexicons are used in experiments in this chapter. Another direction for computing sentiment distributions is to predict the distribution rather than using lexicons to compute values for individual tokens. ULMFit [79][13] trained on the Sentiment140 dataset [64][14]. The method

[10] https://wordnet.princeton.edu/
[11] https://github.com/aesuli/SentiWordNet
[12] https://saifmohammad.com/WebPages/nrc-vad.html
[13] https://www.fast.ai
[14] http://help.sentiment140.com/for-students

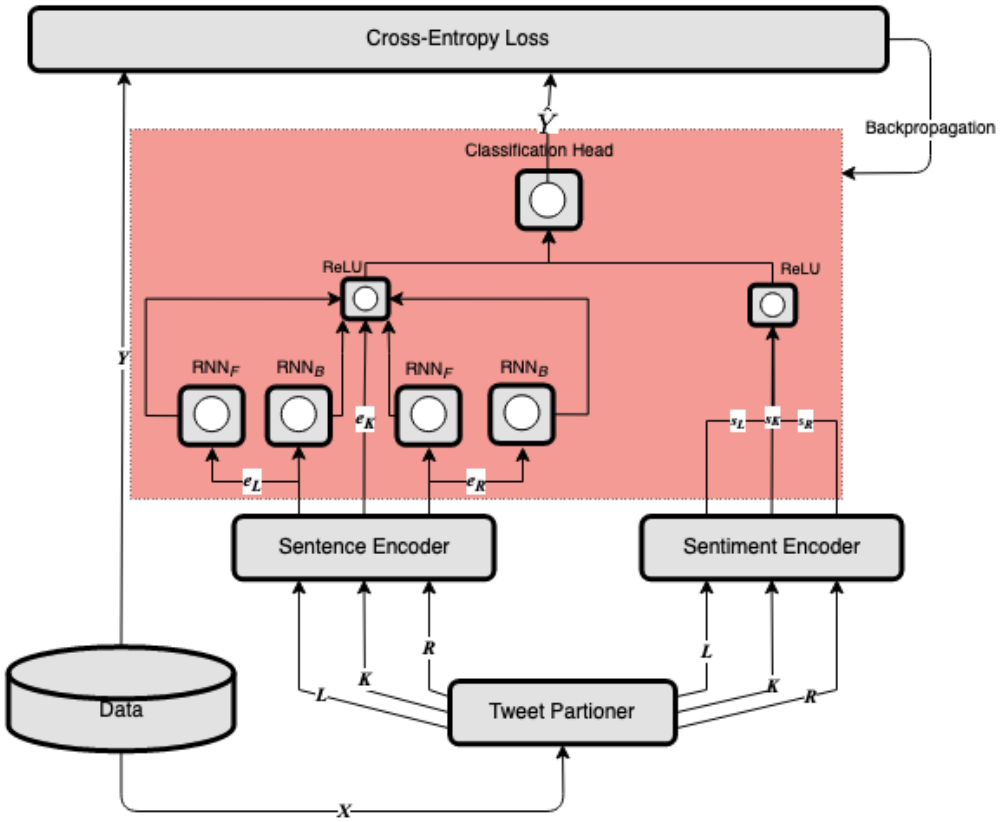

Figure 2.3: **BiLSTM + Senti**

The 9 different layers of the proposed model are bound by rectangles in this diagram.

for computing sentiment is denoted via subscripts for the remainder of this chapter; **BiLSTM+Senti**$_{WN}$ when sentiment distributions are computed using the SentiWordnet lexicon, **BiLSTM+Senti**$_{VAD}$ when sentiment distributions are computed using the VAD lexicons, **BiLSTM+Senti**$_{ULM}$ when sentiment distributions are predicted using ULMFit.

$$BiLSTM(e_i, j) = RNN(e_1 : j) \circ RNN(e_n : j) \tag{2.12}$$

The tweet partition word representations were combined into a BiLSTM architecture in a similar approach to prior research ([135]) with the addition of sentiment distributions. A detailed diagram is provided of the proposed model (see Figure 2.3). From this figure it can be seen that the first module in the proposed model is the Tweet Partitioner, see eq. 2.9. The partitions produced by this module are sent to Sentence Encoder and Sentiment Encoder modules. The Sentiment Encoder computes the sentiment distri-

bution of each of the three Tweet partitions (i.e., s_L, s_k, s_R). These three distributions are concatenated and sent to a ReLU activation unit. Meanwhile, the Sentence Encoder produces contextual word representations of all three tweet partitions, (i.e., e_L, e_k, e_R). These representations are sent to BiLSTMs, (see. eq. 2.12). Where $RNN(e_i, j)$ is used to represent an abstraction of a Recurrent Neural Network (RNN) which computes a vector representing the hidden state of token j in tweet i using the sequence of word representations, e_i, for tokens contained in tweet i. A BiLSTM[66] incorporates two RNNs, one in a forward mode and one in a backward mode. These modes refer to the order in which the sequence is processed, the forward run refers to the standard RNN forward run $x_1 : n$ (i.e., from beginning of sequence up to the specified index). The backwards run processes the sequence in reverse order, $x_n : i$ (i.e., from the end of the sequence to the specified index), $\circ$ is used here to denote vector concatenation. The $BiLSTM(e_i, j)$ is run for all tokens in the three tweet partitions (i.e., L, K and R) and these are all concatenated together resulting in a vector that contains hidden states for all tokens in a tweet. The penultimate step of the model consists of ReLUs being applied to the outputs from the BiLSTMs and the sentiment distributions. In the final layer, outputs from the ReLU are concatenated and feed into a softmax function in the last layer of the classifier. The model is optimised via cross-entropy loss, see eq. 4.2.

$$\mathscr{L} = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\Big] \tag{2.13}$$

2.6.3 Experimental Setup

BiLSTM+Senti and baseline models are trained and evaluated on the **HMC2019** dataset. The dataset is split into stratified train, development and test sets in a 70 to 20 to 10 ratio with stratification performed on label and keyword to ensure representative partitions for experimentation. The number of classes are reduced from three to two by treating the examples labeled as figurative mentions (FM) and non-literal non-health mentions (NHM) as the negative class and those examples labeled as health mentions (HM) as the positive class. The baseline models used in the experiments are those introduced earlier in this chapter (i.e., Jiang$_{LSTM}$, FeatAug$_{+}$, WESPAD and a Linear Baseline). The evaluation metrics reported, $F1$, P, R and R_F, see eqs. (2.1, 2.2, 2.3, 2.5), are all averaged across 10-folds of the training set. For all models grid-search is used to identify optimal parameters for all models, including method of sentiment

Model	**F1**	Δ%	**P**	Δ%	**R**	Δ%	$\mathbf{R}_F$	Δ%
Linear Baseline	0.768	-	0.653	-	0.934	-	0.741	-
FeatAug+	0.791	3.0	0.682	4.4	0.950	1.7	0.780	5.3
Jiang$_{LSTM}$	0.820	6.8	0.721	10.4	0.950	1.7	0.830	12
WESPAD	0.818	6.5	0.752	15.2	0.896	-4.1	0.851	14.8
BiLSTM+Senti$_{WN}$	0.812	5.7	0.716	9.6	0.940	0.6	0.876	18.2
BiLSTM+Senti$_{VAD}$	***0.829***	***7.9***	***0.756***	***15.8***	***0.920***	***-1.5***	***0.897***	***21.1***
BiLSTM+Senti$_{ULM}$	0.825	7.4	0.761	16.5	0.910	-2.6	0.893	20.5

Table 2.5: **Results of HMC experiments (BERT)**

Model represents the learning model, ***F1*** is F1 score, Δ% indicates the % change in metric over Linear Baseline, ***P*** is precision score, ***R*** is the recall score. $\boldsymbol{R}_F$ is the figurative recall. Note: All metrics averaged over 10-folds

Model	**F1**	Δ%	**P**	Δ%	**R**	Δ%	$\mathbf{R}_F$	Δ%
Linear Baseline	0.757	-	0.648	-	0.912	-	0.753	-
FeatAug+	0.764	0.9	0.642	-0.9	0.953	4.5	0.732	-2.8
Jiang$_{LSTM}$	0.776	2.5	0.663	2.3	0.938	2.9	0.775	2.9
WESPAD	0.805	6.3	0.744	14.8	0.878	-3.7	0.845	12.2
BiLSTM+Senti$_{WN}$	0.813	7.4	0.745	15.0	0.905	-0.8	0.888	17.9
BiLSTM+Senti$_{VAD}$	0.817	7.9	0.747	15.3	0.903	-1.0	0.898	19.3
BiLSTM+Senti$_{ULM}$	**0.817**	**7.9**	**0.745**	**15.0**	*0.907*	**-0.5**	**0.899**	**19.4**

Table 2.6: **Results of HMC experiments (ELMo)**

computation, type of word representations, learning rates, dropout, and other model specific parameters (e.g. α and k in WESPAD).

2.7 Results & Discussion

Results are presented in Tables 2.5,2.6 and 2.7. The first observation is that the highest scores for all models are found in Table 2.5 compared to the other two tables. This indicates that using BERT to compute word representations is beneficial across a broad range of models for HMC. This aligns with results seen in Section 2.5 that found that a simple linear baseline classifier benefited from BERT word representations. Focusing on the results of this table, it can be observed that the **BiLSTM+Senti$_{VAD}$** achieves the highest $F1$, P and R_F scores that are 7.9%, 15.8% and 21.1% above the linear baseline respectively. A minor reduction in recall, -1.5% compared to the linear baseline

Model	**F1**	Δ%	**P**	Δ%	**R**	Δ%	$\mathbf{R}_F$	Δ%
Linear Baseline	0.635	-	0.486	-	0.915	-	0.496	-
FeatAug+	**0.787**	**23.9**	**0.742**	**52.7**	**0.842**	**-8**	**0.853**	**72.0**
Jiang$_{LSTM}$	0.752	18.4	0.630	29.6	0.939	2.6	0.726	46.4
WESPAD	0.754	18.7	0.649	33.5	0.901	-1.5	0.724	46.0
BiLSTM+Senti$_{WN}$	0.753	18.6	0.637	31.1	0.922	0.8	0.820	65.3
BiLSTM+Senti$_{VAD}$	0.769	21.1	0.670	37.9	0.907	-0.9	0.849	71.2
BiLSTM+Senti$_{ULM}$	0.770	21.3	0.671	38.1	0.908	-0.8	*0.853*	*72.0*

Table 2.7: **Results of HMC experiments (w2v)**

	Jiang$_{LSTM}$		FeatAug+		WESPAD		BiLSTM+Senti	
Keyword	**F1**	$\mathbf{R}_F$	**F1**	$\mathbf{R}_F$	**F1**	$\mathbf{R}_F$	**F1**	$\mathbf{R}_F$
Alzheimer's	0.731	0.546	0.671	0.853	0.716	0.733	**0.735**	**0.940**
Cancer	**0.702**	0.891	0.644	0.904	0.682	**0.914**	0.684	0.864
Cough	0.822	**0.945**	0.783	0.934	0.79	0.941	**0.831**	0.865
Depression	0.726	0.586	0.716	**0.822**	0.747	0.699	**0.749**	0.816
Fever	0.846	0.848	0.843	**0.886**	0.838	0.84	**0.862**	0.848
Headache	0.906	0.776	0.878	0.773	0.901	0.785	**0.915**	**0.826**
Heart attack	**0.713**	0.917	0.591	0.921	0.686	0.913	0.705	**0.923**
Migraine	0.912	0.68	0.905	0.778	0.914	0.68	**0.926**	**0.837**
Parkinson's	**0.679**	0.739	0.608	0.848	0.66	0.841	0.675	**0.940**
Stroke	0.777	0.828	0.727	0.864	0.789	0.873	**0.792**	**0.917**

Table 2.8: **Results by Keyword**

Keyword refers to keyword. ***F1*** is F1 score, $\boldsymbol{R}_F$ is figurative recall.

is observed indicating that these improvements only cost a minor reduction in the sensitivity of the classifier. Overall, these results indicate that this model is better at detecting when a tweet is a health mention and when a tweet is not a health mention, further to this the figurative recall R_F indicates that the improvement observed is most notable with respect to accurately classifying figurative mentions as the negative class (i.e., non-health mentions).

Pertinent to figurative language understanding is a comparison between the R_F metric for FeatAug$_+$ and the **BiLSTM+Senti** variants as these two models were the only models that were motivated by understanding figurative mentions of health topics. The **BiLSTM+Senti** variants achieve either considerably higher (Tables 2.5 and 2.6) or the same (Table 2.7) R_F score. This result further provides further evidence that the proposed approach to focus on figurative language is successful in the context of HMC.

Similar results are observed in Table 2.6, particularly the $\Delta\%$ columns. Again, the **BiLSTM+Senti** models see the largest increases in $F1$, P, R_F with minor decrease in R compared to the other baselines. The result from 2.7 (i.e., word2vec as word representations) do not align with the results seen when using ELMo and BERT as representations. Interestingly, **FeatAug**$_+$ is the best performing model using word2vec, however, it should be noted that overall the scores across all metrics are lower compared to those seen in Tables 2.5 and 2.6. With respect to the method used to compute sentiment, results from experiments show that the use of SentiWordnet is the worst performing method. However, there is little difference in results between using the VAD lexicon and ULMFit to predict the sentiment distributions.

A breakdown of results by keyword mentions in Tweet is presented in Table 2.8. The strongest performing combination of model, word representation and sentiment method is presented in this table and the $F1$ and R_F is provided for each subset of the test set that contains tweets that only mention a particular keyword. These results provide further evidence that **BiLSTM+Senti** is better at the overall health mention task and dealing with figurative language in a health context. On seven of the ten keyword subsets **BiLSTM+Senti** achieves the highest $F1$ whilst on six of the 10 subsets **BiLSTM+Senti** achieves the highest R_F.

2.8 Error Analysis

In this section an analysis of misclassified examples is provided to improve understanding of the limitations of the proposed model and outline future research directions. The errors covered in this section are examples that were misclassified by **BiLSTM+Senti**$_{VAD}$, with BERT as word representations, on the test set of **HMC2019**.

A common theme amongst false positives were tweets that used health concepts for the purpose of exaggeration in hyperbolic expressions. Particularly, tweets that use complex realities in the formulation of the hyperbole (e.g. '*literally trying to not* ***cough*** *up my lungs from whats happening in my mentions*', '*just drank a kombucha for the first time and my* ***depression*** *is cured*'). These tweets are quite obviously excessive to us as readers but this exaggeration, and absurdity, is missed by **BiLSTM+Senti**$_{VAD}$. In the first example, the author mentions a serious, and violent, health reaction as a response to reading Twitter. The author is obviously exaggerating their feeling of unease, disgust or pain with a health experience that would be uncomfortable to experience. In the latter example the author is exaggerating that having a drink has cured a serious

health condition which is clearly absurd and obviously used for hyperbolic effect. It is interesting to note here that hyperbolic expressions were also used to exaggerate the experience of a health event (e.g. '*Ever had such a bad* ***headache*** *your brain feels like it's going to explode into a million pieces? Yep that's me rn.*'). Suggesting that the presence of hyperbolic expression alone is not enough to indicate that the author is not suffering from a health event.

Another error was based on the exploitation of surface patterns in the data resulting in aggressive classification based on words or phrases that were overwhelmingly found in tweets belonging to a particular class. A particular example of this related to the phrase *cough cough*, that is most frequently used in a figurative sense (e.g. '*the president is blaming mental health issues on gun violence, its not like there is a solution to this* ***couGH couGH*** *better gun restrictions*', '***Cough cough*** *Leaving Neverland accusers* ***cough cough***'. Despite being overwhelmingly used in a figurative sense this phrase also appeared in tweets that were mentioning health events(e.g. '*me the past week: hack* ***cough cough*** *splutter cough*', '*sick* ***cough cough*** *sneeze*'). However, in these cases **BiLSTM+Senti**$_{VAD}$ does not take into account the context in which the phrase was mentioned and classifies them as not being health mentions.

The errors provided in this section appear to showcase that **BiLSTM+Senti**$_{VAD}$ seems incapable of dealing with hyperbolic phrases and tends to ignore the context in which some phrases appear (i.e., phrases in tweets that predominantly belong to a single class)

2.9 Conclusion

This chapter described the details of an empirical study on the expression of figurative language in online user-generated content related to health topics (i.e., symptoms and diseases). Specifically, a study on figurative expressions of disease and symptom words on Twitter and how these expression impact public health applications that use Twitter data as an input signal.

The procedures for collection and annotation of a dataset for the study of health mentions on Twitter, **HMC2019**, were described. An important aspect of the annotation procedure **HMC2019** was the identification and annotation of figurative language resulting in a dataset that can be used to study the phenomenon of figurative language on social media, partially satisfying Research Objective i).

An exploratory analysis of **HMC2019** showed that, on average, symptom and disease

words were mentioned in a figurative sense more frequently then when they were used to actually convey the experience or existence of a particular disease or symptom. This analysis helped to answer Research Question i) and satisfy Research Objective ii) (see Section 1.3 and 1.4), specifically that *figurative language occurs frequently on social media in the context of symptom and disease words*.

This use of figurative language is problematic for public health applications that monitor Twitter for disease and symptom incidence (i.e., Health Mention Classification (HMC)). Experiments were designed to quantify the impact of these figurative expressions on models for HMC, finding that the majority of false positives were a result of figurative expressions of disease and symptom words. This result provided evidence for the need to address this bias caused by figurative expressions of disease and symptom words. These results provided answers to Research Question ii) and partially satisfied Research Objective ii), by demonstrating that NLP *classifiers do not accurately distinguish between figurative and non-figurative expression related to symptom and disease words*.

Algorithmic efforts to address this issue were also detailed in this chapter. The introduction of a text classification model, **BiLSTM+Senti**, based on contextual word representations, Recurrent Neural Networks (RNNs) and sentiment signals was introduced. Experiments designed to probe the ability of this model to address the bias introduced by figurative language were proposed. The results of these experiments showed that **BiLSTM+Senti** was able to address this bias and detect figurative expressions of disease and symptom words better than all other baselines models. This resulted in better predictive performance on the overall task of HMC compared to all other baseline models. The model proposal, experiments and subsequent results partially answered and satisfied Research Question iii) and Research Objective iii), by *providing evidence that better incorporation of sentiment signals can improve the detection of figurative mentions within the context of text classifiers for symptom and disease words*.

However, an error analysis found a number of problematic errors made by **BiLSTM+Senti** that would remain a bottleneck to more accurate classification of health mentions on Twitter. An observed error pattern of concern was the undetected use of hyperbolic expression of symptom and disease words for the purpose of exaggerating the opinion of an author, as opposed to actually expressing the existence of a disease or symptom ('This is the worst joke I've read the entire week, y'all are nothing but a **migraine**.').

The experience of a disease and the accompanying symptoms is unpleasant and

individuals on Twitter invoke these health experiences to exaggerate their current opinions. The appearance of hyperbolic expressions and the errors made by classifiers on these expressions is a key finding that led to a concentrated focus on hyperbolic expressions in the remaining chapters of this thesis.

Part II

Hyperbole

CHAPTER 3

HYPERBOLE

3.1 Introduction

This chapter contains details describing the collection, annotation and exploratory analysis of datasets relating to a particular type of figurative expression (i.e. hyperbole). The content in this chapter addresses the research questions and objective of this thesis as follows:

i. The collection and annotation of tweets focusing on hyperbolic expressions, **HyperTwit**, provides a resource for the study of figurative language on social media (Research Objective i).

ii. The creation of a synthetic test suite for detecting hyperbolic expressions, **HyperProbe**, is introduced to get a better understanding of the limitations of NLP models when detecting hyperbolic expressions (Research Objective i).

iii. An exploratory data analysis of **HyperTwit** provides a quantitative description of how figurative language occurs on social media in terms of the prevalence, intentions and diversity of expression (Research Question i).

The content in this chapter is presented as follows;

- Section 3.2 motivates the importance of hyperbole as a core topic in this thesis and provides review of literature on hyperbole.

- Section 3.3 describes an existing benchmark dataset, **HYPO**, for the computational exploration of hyperbole.
- Section 3.4, introduces the **HyperTwit** dataset. This dataset is a key contribution to computational understanding of hyperbole to emerge from this thesis.
- Section 3.5 details an exploratory analysis of **HyperTwit** that seeks to answer questions relating to the expression of hyperbole on Twitter.
- Section 3.6 describes a synthetic dataset, **HyperProbe** for behavioural testing of hyperbole detection models. This dataset is a key contribution to computational understanding of hyperbole to emerge from this thesis.
- Section 3.7 concludes the chapter.

The datasets and exploratory data analysis presented in this chapter sets the foundation for Part II of this thesis and motivates the importance of hyperbole as a focus point in the study of figurative language on social media.

3.2 Hyperbole as Important Figure of Speech

The author is interested in studying the nature of hyperbole on social media and the computational detection of hyperbole for several reasons;

- The commonness of the figure of speech particularly in informal settings (i.e., online user-generated text) increases the importance of computational methods that can process hyperbolic text
- The observation that hyperbole has been understudied compared to other figures of speech particularly within the field of NLP compared to metaphor, sarcasm and irony
- Given the predominantly connotative nature of hyperbole and the ability for hyperbole to be used to express both a positive and negative sentiment, understanding hyperbole is important for affective computing applications (i.e., sentiment analysis)
- Findings from Part I showed that hyperbolic usage of health concepts often went undetected by text classifiers trained to identify figurative usage of health concepts

3.2.1 Hyperbole

Hyperbole is one of the many common figures of speech used for figuration as well as metaphor, simile, irony, sarcasm and several others. As mentioned in Section 1.2, the exact definition of the figurative devices and the classification of figurative utterances has been the source of debate for centuries. With respect to hyperbole, some scholars have treated hyperbole as a sub-type of metaphor or a sub-type of irony rather than treating hyperbole as a unique figure of speech [26, 114, 185, 212, 238]. However, the author shares the view that hyperbole is in fact a unique figure of speech that displays characteristics that are not shared by any other figure of speech and it should be treated separately from other figures of speech [24, 26].

The definition of hyperbole from this point of view is that the figure of hyperbole is defined by an intentionally excessive contrast between utterance meaning and reality along a semantic scale to convey an evaluation (i.e., '*this computer takes like* ***500 years*** *to load a web page*', '*his room is the size of a* ***shopping mall***', '*she put the team* ***on her back and carried them*** *to a win*') [24, 26, 33, 133, 145]. By deliberately expressing this contrast an author of a hyperbolic utterance is conveying a positive or negative evaluation of the state of affairs that they have embellished. In the previously provided examples, the computer is *frustratingly* **slow** to load a web page, the bedroom is *disappointingly* **small** and her **contribution** was *impressive*. The connotative nature of hyperbole and the potential for complex and varied expression of sentiment heighten the importance of understanding hyperbole for computing applications that interpret the affective content in text (i.e., sentiment analysis).

The overwhelming majority of hyperbolic expressions are connotative, which is a key feature of hyperbole that differentiates it from a plain metaphor or simile [26, 33]. Metaphor and simile for example, may be expressed without evaluative intent but rather with the intent to improve understanding [26] (e.g, '*these chips taste like spicy chicken*', '*a whippet is like a small greyhound*'). A hyperbolic simile is one in which the likeness is obviously exaggerated to convey an evaluation ('*these chips taste like heaven*', '*a whippet runs like the wind*'). The use of hyperbole to convey both positive and negative evaluations differentiates it from ironic language. A key feature of irony is the communication of negative disassociative evaluations, whilst similar evaluations can be achieved via hyperbole, hyperbole is used to convey positive evaluations [26].

Despite these differences, these figures also share similarities and frequently co-occur, fueling the debate over the respective definition and boundaries between the figures. A corpus study found that hyperbole was found in 80% of all examples where

figures of speech were found to co-occur and was found to be the second most frequently occurring figure of speech [106]. Other studies have also noted the high prevalence of hyperbole, particularly in informal settings [26, 133, 145]. One such informal setting is the internet, with the New York Times proclaiming a '*death by internet hyperbole*' [14] and the Guardian arguing that '*Exaggeration is the official language of the internet*' [21] regarding the amount of hyperbole in online content.

However, despite the commonness of hyperbole and the frequent co-occurrence with other figures of speech, hyperbole has received little attention relative to other figures of speech [24, 26, 165, 193]. Most relevant to this thesis is the observation that the computational study of hyperbole has been overlooked compared to computational studies on other figures of speech [3, 101, 222].

There are several resources for the computational study of irony, sarcasm, metaphor and simile. Including annotated datasets for the study of irony on Reddit [233] and Twitter [11, 100]. Several Twitter datasets have been created via hashtag supervision (i.e., '#sarcasm', '#not' to indicate presence of sarcasm) for the study of sarcasm [1, 12, 16, 119] as well as several resources for the study of metaphor and simile [48, 58, 69, 179, 205].

Comparatively there are few resources to study the phenomenon of hyperbole. The first work to introduce the computational task of detecting hyperbole and prove the feasibility of the task on a small dataset consisting of 700 simple idiomatic hyperboles was recently published [222]. An extension to this foundation work is a study of hyperbole in Mandarin Chinese, formed by compiling hyperbole from websites and research papers, resulting in a dataset of idiomatic hyperbole [101]. See [3] for a comprehensive survey of available resources for the computational study of figurative language.

3.2.2 Hyperbole Types

For this study on hyperbole the focus is on three key types of hyperbole, the extreme case formulation (ECF), quantitative hyperbole and qualitative hyperbole, see Table 3.1.

Extreme case formulations (ECF) are semantic formulations that invoke extreme descriptions of events or objects [174, 236]. ECFs are not limited to a singular grammatical pattern or word class and as such can be formulated in a myriad of ways [236]. However, a typical example of an ECF is a sentence containing an extreme description via an adjective (*entire*, *absolute*, *infinite*, etc.), adverb (*never*, *always*, etc.), quantifier (*none*, *all*, etc.) or indefinite pronoun (*nobody*, *everybody*, etc.) [49, 150], see Table 3.1 for examples.

Type	Example
ECF	Her smile is **absolutely perfect**
	All you **ever** do is complain
	He seems to have **infinite** excuses
	Everybody knows the answer already
Quantitative	I am so hungry I could eat a **million** pizzas
	It is like **1000** degrees today
	That skirt probably costs a **billion** dollars
	I slept for like a **millisecond** last night
Qualitative	My stomach is on **fire**
	This song is **heaven**
	He is a **cancer** to your life
	She just talks so much **garbage**

Table 3.1: **Hyperbole Types**

Type indicates type of hyperbole. ***Example*** contains an example hyperbolic utterance, **emphasis** indicates key term associated with hyperbole type

The many functions of ECFs in communication have been well covered in the literature [49, 150, 174, 236]. Ranging from strengthening claims in order to pre-empt challenges (e.g., *'you have to buy it, it's **brand** new'*) [174], stating the morals of actions by virtue of commonness (e.g., *'it's fine, **everyone** does it'*) [174] and intentional non-literal descriptions in evaluative statements (e.g., *'the worst sandwich **ever**'*) [236].

A rich source of hyperbolic expressions in the non-literal and intentionally use of ECFs [26, 133, 145, 150, 236]. An analysis of conversations from the British National Corpus, revealed that the semantic concepts of absoluteness (*absolute*, *complete*, *entire*, *pure*, etc.), non-existence (*never*, *nobody*, *nothing*, *null*, etc.) and universality (*all*, *always*, *every*, *universal*, etc.) were the most (15.7%), equal second most (10.7%) and the fourth most (6.1%) common semantic categories for hyperbole [145]. Although not explicitly identified as ECFs in their analysis, the hyperboles related to these concepts are ECFs as they provide maximal or minimal descriptions to the objects or events to which they are describing (e.g., *the **entire** country was angry*, *you sit around **all** day doing **nothing***). In a corpus analysis of everyday conversation, extreme adjectives and adverbs such as *infinitely*, *endless* and *everywhere* were regularly used in hyperbolic expressions, again likely to be ECFs [133].

Quantitative hyperboles align with the objective-gradational dimension of hyperbole as defined in prior research on hyperbole [145]. The distinguishing feature of qualitative

Dataset	Source	Size	Annotations
HyperTwit$_K$	Twitter	6,150	Manual
HyperTwit$_R$	Twitter	3,750	Manual
HyperProbe	Manual	4,990	Manual
HMC2019	Twitter	19,558	Manual

Table 3.2: **Overview of Dataset Contributions in this Thesis**

hyperbole is the exaggeration of an *obvious* magnitude or magnitude to an extreme degree (e.g., '*she says a* ***million*** *words a minutes*', '*today has gone for like* ***100 hours already***'), see Table 3.1 for more examples. These hyperboles differ from ECFs in terms of the magnitude of contrast. In ECFs the contrast is via a maximal description whereas in a quantitative hyperbole the contrast is not maximal (e.g, '*this year has felt like an* ***eternity***' vs. '*this year has felt like a* ***decade***'). Corpus studies have shown that numerical expressions related to quantity and accumulation are rich sources of hyperbole [133, 145]. Specifically, quantity words such as '*dozens*', '*hundreds*' and '*millions*' and those relating to mass (i.e., *masses*, *tons*, *loads*, etc.,) were found to be prone to hyperbolic usage [133]. In a study of the British National Corpus (BNC), the semantic concept of time measure (i.e., *months*, *hours*, *weeks*, etc.) had the most occurrences of hyperbole of those categories related to numerical expressions [145].

Qualitative hyperboles align with the subjective-emotional dimension of hyperbole [145]. The distinguishing feature of qualitative hyperbole is a subjective evaluation made to an extreme degree (i.e., '*that play was* ***cancer***', '*those fries are cooked by* ***God himself***'), see Table 3.5 for more examples. A common method for constructing a qualitative hyperbole is through analogy via a simile or metaphor. The author takes the view that a hyperbolic metaphor or simile is one in which the analogy is patently absurd and predominantly for evaluative purposes rather than descriptive ('*these chips taste like spicy chicken*' vs. '*these chips taste like* ***heaven***'). Corpus studies show that qualitative hyperboles often provided negative evaluations [145]. Concepts such as frightfulness (21% of all evaluative hyperboles), physical loss (20%), sorrow and pain (17%), violence and destruction (12%) were common sources of negative hyperbolic evaluations. Qualitative hyperbole are often surrounded by loose language, requiring more context and reasoning to interpret, often involving complicated imagined realities [150] (e.g. '*I would rather be* ***french kissed by a rattlesnake***'). Qualitative hyperboles are an interesting point of focus due to their difference in form to ECFs and Quantitative hyperboles and their co-occurrence with other figures of speech such as simile and metaphor.

ID	Corpus	Text
1	Hyperbole	I know this place **like the back of my hand**
	Paraphrase	I know this place *well*
	Minimal Units	Your baby is already hairy **like the back of my hand.**
2	Hyperbole	Love you to the **moon and back.**
	Paraphrase	Love you so *much*.
	Minimal Units	The missions successfully went to the **moon and back.**
3	Hyperbole	Man your compassion is **greater than space**.
	Paraphrase	Man your compassion is *huge*.
	Minimal Units	Spacetime is more complicated **than space**.
4	Hyperbole	I went into the shop and we **cleared the shelves out**.
	Paraphrase	I went into the shop and we *bought a lot of things*.
	Minimal Units	At the restaurant they **cleared the shelves out** and put in some old tables.
5	Hyperbole	By the time Alf finishes that story, **his beard will be three inches longer**.
	Paraphrase	By the time And finishes his story, **a lot of time will have passed**.
	Minimal Units	In a few months, **his beard will be three inches longer**.
6	Hyperbole	Marriage is the **grave** of love.
	Paraphrase	Marriage is the *end* of love.
	Minimal Units	I have gone to visit the **grave** of a friend.

Table 3.3: **HYPO examples**

Hyperbole Corpus contains utterances deemed hyperbolic during annotation. **Paraphrase Corpus** contains a non-hyperbolic paraphrase of the original hyperbolic utterance. **Minimal Units Corpus** contains literal utterances that contain the tokens considered to hyperbolic in the original hyperbolic utterance in a non-hyperbolic context.

3.3 HYPO

The **HYPO** dataset [222] is a collection of utterances that are annotated for the presence of hyperbole. The utterances are a mix of manually composed examples and those sourced from various online sources ranging from news headlines, television scripts, love letters and advertisements.

Crowd workers were employed to provide annotations for the HYPO dataset. The workers were instructed to complete six tasks for each utterance. The first task was to ascertain whether the crowd workers thought the utterance was hyperbolic, resulting in a binary variable. For those utterances deemed to be hyperbolic, the crowd workers were then instructed to highlight the specific tokens they deemed to be hyperbolic. Then

the crowd workers were asked to further annotate the utterance by composing a literal paraphrase of the original hyperbole. Workers were also asked to indicate the type of hyperbole (i.e., quantitative or qualitative, creative or conventional).

The binary indicator variable from the first task was used to filter the data resulting in 709 hyperbolic utterances. This collection of sentences was denoted as the Hyperbole Corpus by the authors. A second corpus of data was created using the hyperbolic tokens as identified by the crowd workers. The utterances in this corpus were constructed by filtering the WaCKy corpus[1] for sentences that contained tokens identified as hyperbolic in a non-hyperbolic context. This was denoted as the Minimal Units Corpus by the authors. The non-hyperbolic paraphrase also made up another corpus, the Paraphrase Corpus. The construction of the dataset in this way meant that each hyperbolic utterance in the Hyperbole Corpus had two non-hyperbolic counterparts from the Paraphrase and Minimal Units corpora respectively, see Table 3.3.

3.4 HyperTwit

The **HyperTwit** dataset is an annotated collection of online user-generated texts from social media platform Twitter[2], consisting of approximately 10k Tweets annotated for presence of hyperbole, There are several motivations that inform the data collection and annotation for **HyperTwit**. Specifically;

- How prevalent is hyperbole on Twitter?
- How is hyperbole expressed on Twitter and what are the intended meanings of hyperbole on Twitter?
- How diverse, in terms of usage and intention, is hyperbole on Twitter?
- Can hyperbole expressed on Twitter be automatically detected?

3.4.1 Data Collection

Data is collected via two sampling strategies:

i. **Random Sampling**: randomly sample Tweets using the Twitter API[3]

[1] https://wacky.sslmit.unibo.it/doku.php
[2] https://twitter.com/
[3] https://developer.twitter.com/en/docs

ii. **Keyword Sampling**: query the Twitter API to return Tweets containing predefined keywords

For random sampling, the author randomly samples Tweets during February-March 2021. Random sampling allows us to estimate the prevalence and usage of hyperbole on Twitter over time, whilst keyword sampling allows us to compare the hyperbolic use of particular words on Twitter to prior findings from different communicative forms.

For the keyword sampling strategy, a list of 127 keywords motivated by prior research on hyperbole is compiled [145], see Table 3.4. The Twitter API is queried to return Tweets that mention these keywords during September 2020-March 2021.

After collecting Tweets from the Twitter API tweets are automatically filtered based on the following exclusion criteria:

- Exclude Tweets with any Twitter meta-characters (@, #, urls)
- Exclude Retweets, Quote and Reply Tweets
- Exclude Tweets with less than 4 words
- Exclude non-English Tweets

The motivation for this strict exclusion criteria is to reduce the complexity of collected Tweets and minimise signatures of the Twitter platform in the dataset. In addition to the Twitter-specific nature of meta-characters, tweets containing them are also ignored because they introduce extra context that may be required to correctly interpret the tweet (i.e., knowledge of a particular user and their Twitter activity in an @mention, knowledge of the event/topic/phenomenon represented by a particular # or the resource given by a URL). The motivation to exclude retweets and quote tweets is to help remove duplicates. Exclusion of reply tweets is due to the extra context needed to correctly interpret a tweet (i.e., the initial tweet being replied to by the reply tweet). Tweets with less than five words are also removed because tweets of this length can be vague and ambiguous (e.g., '*...No. the opposite.*', '*tbz no air*').

Manual filtering is also performed according to the following exclusion criteria:

- Exclude vague tweets
- Exclude multi-lingual tweets

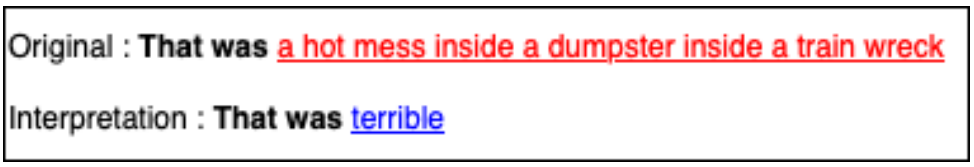

Figure 3.1: **Example Data**

Original is the source Tweet. ***Interpretation*** is a literal interpretation of the Tweet.

Vague tweets and incomprehensible tweets are considered unwanted noise and are removed from the data (e.g., '*meat machinery victim battery audience complete*'). A multilingual tweet may still be tagged as an English language tweet by Twitter therefore not excluded during the automatic filtering process (e.g., '*Pedazo final de Little Fires Everywhere*').

Upon completion of filtering 125 Tweets per day are randomly sampled for 30 days from those collected via random sampling, this subset of the data is referred to as $\textbf{HyperTwit}_R$ containing 3,750 Tweets. Also, 50 Tweets are randomly sampled per keyword from those collected via keyword sampling, this subset of the data is referred to as $\textbf{HyperTwit}_K$ containing 6,150 Tweets.

3.4.2 Annotation and Inter-Annotator Agreement Study

The annotation process follows that of Troiano [222]. Firstly the presence of hyperbole within a tweet is marked by the assignment of a binary label for that tweet. Then, for each hyperbolic Tweet, X, the annotators manually compose a literal interpretation of that tweet, Y, see Figure 3.1. The instructions for annotators are to perform minimal edits to X to remove the hyperbolic excess and capture the intended meaning of the utterance as understood by the annotator.

The agreement and similarity between annotators is examined when following the previously defined annotation task. Given the design of the task there are two aspects of annotation that are of interest. The inter-annotator agreement regarding the decision to annotate a tweet as hyperbolic or not, and the similarity of literal interpretations between annotators. A random sample of 200 tweets are collected to be annotated by three individuals familiar with the study.

Firstly, the agreement between annotators regarding the decision to label a tweet as hyperbolic or not is probed. Krippendorff's α[8, 107][4] is used to measure the agreement

[4]This metric is in the range of [-1,1] with -1 indicating disagreement, 0 indicating no consensus and 1 indicating complete agreement.

Semantic Concept	Type	Word List
Complete/Absolute	ECF	absolute, complete, entire, pure, whole
Non-existence/Nullity	ECF	impossible, never, no, nobody, nowhere
Perfection	ECF	perfect, flawless
Time Period	ECF	endless, eternal, infinite
Universality	ECF	all, always, every, everybody, everyone, everywhere
Veracity	ECF	definite, exact, undeniable
Quantity Words	Quantitative	zero, one, two, three, four, five, six, seven, eight, nine, ten, hundred, thousand, million, billion, trillion, load, heap stack, pile
Time Period	Quantitative	hour, day, week, month, year, decade, century
Dimensions	Quantitative	small, big, slow, fast, thin, thick, heavy, light, height, weight, tall, length, large, high
Measure	Quantitative	feet/foot, inch, metre, mile
Badness/Evil	Qualitative	bad, corrupt, evil, fraud, wicked
Chaos/Disorder	Qualitative	chaos, confusion, disorder, garbage, riot
Deadly/Hell	Qualitative	dead, hell, misery, murder, nightmare
Frightfulness	Qualitative	alarm, fear, panic, scared, shock
Physical loss	Qualitative	anxiety, autism, blind, deaf, insomnia
Pungency/Shrill	Qualitative	bitter, pierce, sharp, spicy, toxic
Sorrow/Pain	Qualitative	cancer, fever, headache, pain, sad, suffer
Violence/Destruction	Qualitative	attack, explode, fight, rape, ruin, wreck
Life/Heaven	Qualitative	dream, heaven, paradise, utopia, vital
Splendour/Beauty	Qualitative	attract, beauty, charm, grace, handsome
Magnificence	Qualitative	amaze, good, great, ideal, impress

Table 3.4: **Hyperbole term list**

Semantic Concept is the semantic concept as defined by Mora [145]. **Type** refers to the type of hyperbole as defined by Mora [145]. **Word List** is a list of the keywords in keyword list.

Type	Tweet
ECF	depending on people *leads you* ***nowhere***
	Academic failures are *worse than* ***all*** *other types of heartbreaks*
	they *ate* that concert up fr their vocals *BEYOND* ***FLAWLESS***
	Being a student in 2020 is *legit fucking* ***impossible***. At least it *can't get any worse*
Quantitative	Cozy *Levels On A* ***Million***
	Looks like it's gonna have to be a ***thousand****-cups-of-coffee* day
	if my laptop could take less than *500* ***years*** to load a photo that would be NEAT
	Wait its October already? How the hell did April *last ten* ***years*** and September *ten* ***minutes*** this year?
Qualitative	Nigeria is *a time bomb waiting to* ***explode***
	This referee needs to retire immediately. *Legally* ***blind***. ..Pathetic.
	Bruh if anyone ever told me that they would buy my art I would *drop* ***dead*** *on the spot*
	Working in the cryptocurrency world is *like working for a drug addict with bipolar* ***disorder***.

Table 3.5: **Hyperbole Types - HyperTwit**

Type indicates type of hyperbole as classified by [145]. ***Example*** contains tweet text, **emphasis** indicates key term associated with hyperbole type, *emphasis* indicates hyperbolic tokens

for this task . Analysis showed an α of 0.595 indicating moderate agreement between annotators on what constitutes a hyperbolic tweet. Troaino et al. Observed Agreement (A_o) was to measure inter-annotator agreement in another study of hyperbole [222], they calculate an A_o of 0.802, by comparison in this study an A_o of 0.816 is calculated. In another study of hyperbole annotation, Cohen's κ is used in their agreement study of hyperbole annotations, a Cohen's κ of 0.62 is calculated in that study. An average Cohen's κ of 0.638 between all possible pairings of annotators is computed for this inter-annotator agreement study. The inter-annotator agreement study on a random sample of HyperTwit data indicates similar levels of agreement to those achieved in other studies related to hyperbole.

To gain an understanding of the similarity of hyperbole interpretations an examination of the differences between the original hyperbolic tweet, X, and the literal interpretation, Y, provided by the different annotators is undertaken. An assessment of interpretation similarity is approximated by computing the semantic similarity between

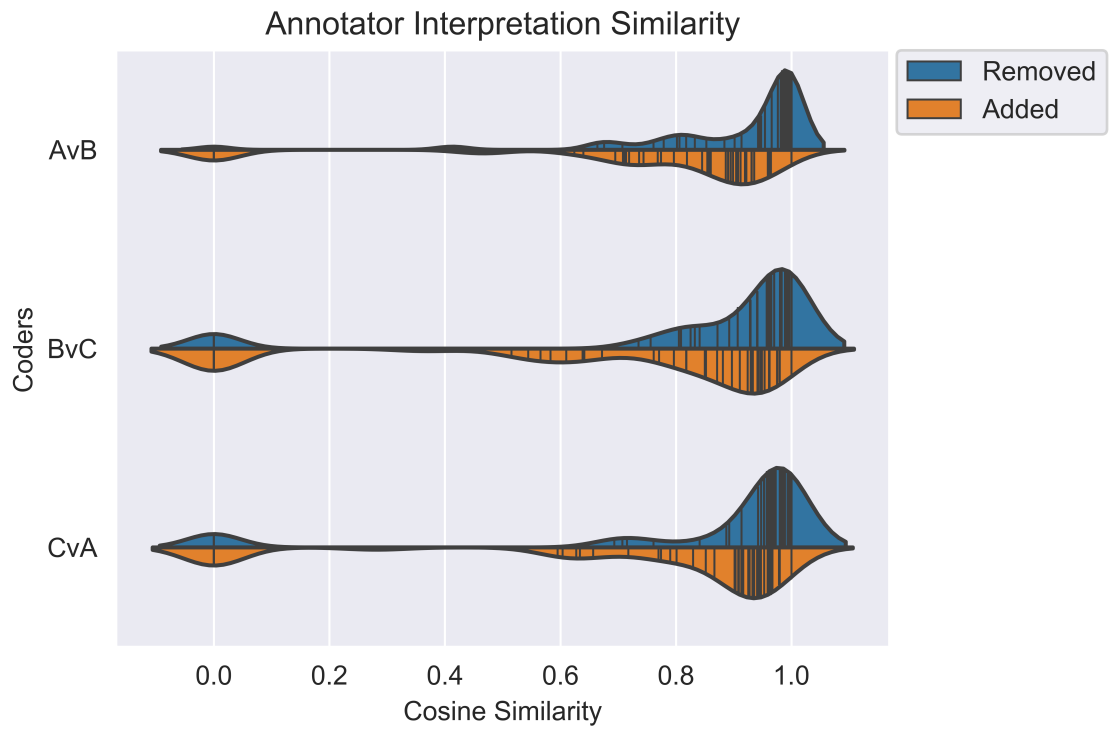

Figure 3.2: **Annotator Interpretation Similarity Distributions**

Distributions of cosine similarity words removed and added by different annotator pairs when interpreting a hyperbolic tweet.

the tokens removed and added during interpretations. For this analysis two separate bags-of-words R and A, see eq. 3.1 and eq. 3.2 are considered. The simplediff[5] library is used to compute these bags-of-words. This library provides an algorithm that uses a dynamic programming approach to compute the set difference and set intersections for strings of text.

Non-contextual dense word representations are used to represent the linguistic content in these bags of words. Due to the comparison occurring between tokens removed from, and added to, highly similar contexts by different annotators. Contextual word representations would encode information from the context of the original tweet biasing this similarity metric due to the highly similar context (i.e. higher similarity scores).

(3.1) $$R = X \notin Y$$

[5]https://github.com/paulgb/simplediff

$$A = Y \notin X \tag{3.2}$$

$$SIM_R(x,y) = \frac{e_R^x \cdot e_R^y}{||e_R^x|| \times ||e_R^y||} \tag{3.3}$$

$$SIM_A(x,y) = \frac{e_A^x \cdot e_A^y}{||e_A^x|| \times ||e_A^y||} \tag{3.4}$$

Each bag-of-words is represented via the averaged GloVe[6] [166] embeddings for each word in the bag denoted by e_R and e_A. The cosine similarity is computed between e_R for all hyperbolic tweets across all possible annotator pairs, likewise for e_A, similar to analysis in [78]. These metrics are denoted by SIM_R and SIM_A, see eqs 3.3 and 3.4, where e_R^x and e_R^y are the averaged GloVe representations of the tokens in the removed bags-of-words for annotator x and annotator y respectively, and e_A^x and e_A^y are the averaged GloVe representations of the tokens in the added bags-of-words for annotator x and annotator y respectively.

High semantic similarity is observed between all annotator pairings for e_R with mean SIM_R of 0.895, 0.837 and 0.835 respectively. This suggests that when annotators agree that a tweet is hyperbolic they identify similar tokens as contributing to the hyperbole. However, when it comes to the added tokens the semantic similarity is considerably lower with mean SIM_A of 0.733, 0.662 and 0.704 for e_A between all annotator pairs respectively. This suggests that interpretation of a hyperbole is more open-ended task then the hyperbole identification task, an intuitive result.

Kruskal-Wallis significance testing is performed on SIM_R and SIM_A between all annotator pairs. Finding no significant difference[7] between the cosine similarity for removed words between all the different annotator pairs, mean SIM_R of 0.895, 0.837 and 0.835 respectively. It also observed that there is no significant difference[8] for similarity of added words between all annotator pairs SIM_A of 0.733, 0.662 and 0.704 for e_A between all annotator pairs respectively. The outcome of the significance tests show consistency across different annotator pairs.

[6] https://radimrehurek.com/gensim/

[7] $p = 0.79$

[8] $p = 0.68$

3.5 Hyperbole on Twitter

This section details an exploratory analysis of the **HyperTwit** dataset to gain insight into the phenomenon of hyperbole. Specifically, answers to the following questions are sought;

- How prevalent is hyperbole on Twitter?
- How is hyperbole expressed on Twitter and what are the intended meanings of the hyperbolic expressions?
- How diverse, in terms of usage and intention, is hyperbole on Twitter?

3.5.1 Hyperbole Prevalence

HyperTwit consists of 9,900 labelled Tweets, of which 2,892 (29.2%) are hyperbolic. It is observed that 14.8% of Tweets in $\textbf{HyperTwit}_R$ are hyperbolic, whilst 39.0% of Tweets in $\textbf{HyperTwit}_K$ are hyperbolic. The prevalence of hyperbole in **HYPO** is not analysed given that the data collection procedure forbids such an analysis, see 3.3.

To get an idea of the prevalence of hyperbole on Twitter a comparison of the observed frequency of hyperbole in $\textbf{HyperTwit}_R$ is made against the observed frequency of hyperbole in the corpus study by [145]. In their study of conversational text the authors found 343 hyperbolic units amongst 52,208 words (0.007) of conversational text, in $\textbf{HyperTwit}_R$ 555 hyperbolic units are identified amongst 55,909 words (0.010) of Twitter text. A χ^2 test is performed and statistically significant difference[9] in the observed amount of hyperbole between the two datasets is observed.

This suggests that Twitter text may be more hyperbolic than general conversational text. The author was unable to compare hyperbole prevalence with several prior studies on hyperbole for various reasons. Both **HYPO** [222] and the data in [101] use deterministic sampling and do not attempt to provide an estimate of the prevalence of hyperbole. In contrast, [133, 145] focus on the hyperbolicity of particular hyperbolic phrases, not the prevalence of hyperbole in general.

Another point of inquiry is the variation in hyperbole prevalence over time. A Sharipo-Wilk test is performed to see if the daily proportions of hyperbole in $\textbf{HyperTwit}_R$ follow a normal distribution. The results[10] show that the daily proportions of hyperbole follow

[9] $\chi^2 = 36.53$, $p < 0.0001$
[10] $p = 0.848$

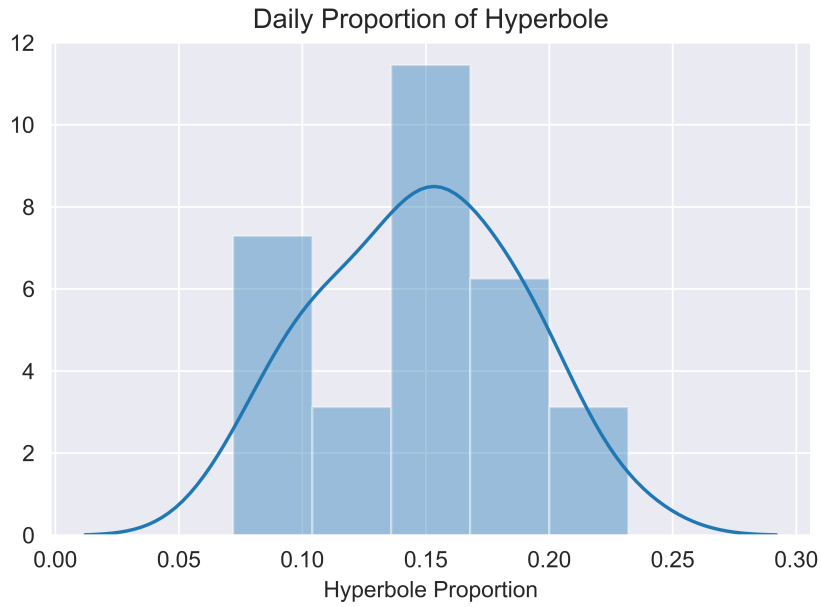

Figure 3.3: **Daily Proportion of Hyperbole**

. KDE plot of daily hyperbole proportions in **HyperTwit**$_R$

a normal distribution with strong statistical significance. This indicates a steady daily prevalence of hyperbole on Twitter, see Figure 3.3.

3.5.2 Word Hyperbolicity

In this section an investigation into the hyperbolic usage of words in **HyperTwit** is undertaken. To gain an understanding of the hyperbolicity of different words an examination of the differences between the original hyperbolic Tweet, X, and the literal interpretation, Y, provided by the different annotators, see Figure 3.1. For this analysis two separate bags-of-words R and K see eq. 3.1 and eq. 3.5 are considered. All words in Tweets are stemmed using the Snowball stemmer[11] before computing the bags-of-words. It is assumed here that a word was removed from a Tweet during the annotation because it contributed to the hyperbolic nature of that Tweet given the annotation instructions, see Section 3.4.2. Therefore, an estimation of the probability of removal can be used as an approximation of the hyperbolicity of that word. To measure the hyperbolicity of a word an estimate of the probability of removing a word from a Tweet is calculated by eq.

[11] https://www.nltk.org/

3.6, where r_i is the count of times word i was removed, α is a smoothing parameter[12] and f_i is the raw frequency count of word i across the dataset. This method of computing probabilities (i.e. Laplace Smoothing) has shown to be effective common when dealing with word counts in various NLP applications [30, 95, 131, 225]

$$S = X \cap Y \tag{3.5}$$

$$P(R|w_i) = \frac{r_i + \alpha}{f_i + \alpha 2} \tag{3.6}$$

Several words with high removal probabilities are identified in both subsets of HyperTwit indicating words that are prone to hyperbole on Twitter. In Tables 3.6 and 3.7, the Top 15 keywords by removal proportion in both subsets of **HyperTwit** are provided. The high removal proportion of these terms is notable, particularly in **HyperTwit**$_K$, Table 3.7. The valence, arousal and dominance (VAD) scores of words with high removal probability are computed by using a lexicon [144] to gain insight into the affect of these words. From the VAD scores it can be observed that all words contain extreme values ($x < .25$ and $x > .75$) in at least one of these affect dimensions except *heavy*, *steam* and *cure*, with both *heavy* and *cure* being close to extreme. With respect to valence, 6 and 7 of the words with extreme valence are negative in tables 3.6 and 3.7. This suggests a preference for negative hyperbolic expressions on Twitter. The extreme values of arousal in both Tables lean towards the active/stimulated dimension. With respect to dominance the extreme values are split between strong/weak and lean towards weak in Table 3.7. Whilst words with extreme dominance values lean towards the strong end of the scale in Table 3.6. These extreme values in the affect dimensions aligns with findings in hyperbole research that the overwhelming majority of hyperbole found in various corpus studies indicated strong sentiment, see Section 3.2.

3.5.3 Common Intentions

This section introduces an empirical analysis of the annotations to gain insight into the words commonly used in literal interpretations of hyperbole. For this analysis the bag-of-words A, see eq 3.2 , is considered. It is assumed here that words added to a non-hyperbolic interpretation capture the intended meaning of the hyperbole given the annotaion instructions, see Section 3.4.2. Stopwords are removed before computing the

[12] $\alpha = 0.1$

Word	$P(R\|w)$	**V**	**A**	**D**
planet	0.981	0.698	0.404	**0.832**
heaven	0.976	**0.896**	0.385	0.600
deadass	0.976	-	-	-
saddest	0.969	0.651	0.529	**0.917**
demon	0.969	**0.037**	**0.908**	0.509
heavy	0.969	0.250	0.454	0.600
absolute	0.922	0.526	0.510	**0.827**
insane	0.880	**0.062**	0.670	0.265
bomb	0.788	**0.167**	**0.912**	**0.750**
earth	0.744	**0.750**	**0.225**	0.614
toxic	0.744	**0.008**	**0.885**	0.492
trash	0.738	**0.163**	0.541	**0.154**
worst	0.723	**0.062**	0.704	**0.225**
perfect	0.711	**0.980**	0.471	**0.870**
brain	0.689	0.667	0.441	**0.823**

Table 3.6: **Top 15 words by Removal Probability - HyperTwit$_R$**

Word is the removed word. $P(R|w)$ is the estimated removal probability. ***V*** is the valence. ***A*** is the arousal. ***D*** is the dominance. Extreme values of **V**,**A**,**D** are in **boldface**.

Word	$P(R\|w)$	**V**	**A**	**D**
insane	0.986	**0.062**	0.670	0.265
ascend	0.969	**0.830**	0.620	**0.812**
retard	0.969	**0.194**	0.347	**0.222**
explode	0.969	0.277	**0.885**	**0.773**
slap	0.969	**0.100**	**0.804**	0.518
disease	0.969	**0.041**	0.539	0.333
dumpster	0.969	**0.229**	0.418	**0.101**
steam	0.969	0.469	0.413	0.421
toxic	0.929	**0.008**	**0.885**	0.492
everywhere	0.9	-	-	-
cure	0.866	0.719	0.377	0.623
nowhere	0.861	-	-	-
absolute	0.858	0.526	0.510	**0.827**
garbage	0.849	**0.188**	0.510	**0.192**
pure	0.849	**0.811**	0.306	0.652

Table 3.7: **Top 15 words by Removal Probability - HyperTwit$_K$**

Word is the removed word. $P(R|w)$ is the estimated removal probability. f_i is the raw frequency of word. ***V*** is the valence. ***A*** is the arousal. ***D*** is the dominance. Extreme values of **V**,**A**,**D** are in **boldface**.

Word	f_i	**V**	**A**	**D**
great	38	**0.958**	0.665	**0.810**
often	28	-	-	-
bad	27	**0.125**	0.625	0.373
many	24	-	-	-
good	22	**0.938**	0.368	0.534
much	20	-	-	-
frustrate	19	**0.100**	**0.809**	**0.243**
very	18	-	-	-
annoy	17	**0.094**	**0.765**	0.286
not	16	-	-	-
too	15	-	-	-
thing	15	0.449	**0.222**	0.26
people	14	0.604	0.400	0.500
terrible	13	**0.061**	**0.849**	0.604
hard	10	0.302	0.708	0.616

Table 3.8: **Top 15 words Added to Literal Interpretations - HyperTwit$_R$**

Word is the word added to a literal interpretation. f_i is the raw frequency count of the times added to a literal interpretation. ***V*** is the valence. ***A*** is the arousal. ***D*** is the dominance. Extreme values of **V**,**A**,**D** are in **boldface**.

raw frequency counts of words in all A, and show the top 15 most frequently added words in Table 3.9 and Table 3.8, the VAD scores for all words in these tables are also provided. Notably, there are several words in Tables 3.9 and 3.8 respectively that have no VAD score associated with them. These words have very little affect signal and are associated with specifying quantities (e.g., *many*, *much*, *too*, *very*, *often* and *few*). The words with VAD scores (*good*, *bad*, *great*, *frustrate*, *annoy* and *terrible*) show extreme values of valence suggesting that the common intention of hyperbole is to express strong sentiment. The extreme values of valence in the majority these words suggests that the common intention of hyperbole is to express a strong sentiment. This is intuitive and aligns with the theory of hyperbole as contrast to convey and evaluation [24, 26, 101, 145, 222]. This also shows the importance of accurate interpretation of hyperbole for Sentiment Analysis. Interestingly, 12 words are shared between the top 15 most frequently added words between both **HyperTwit$_K$** and **HyperTwit$_R$** subsets. This is a contrast to only 2 shared words in between the subsets in Tables 3.6 and 3.9 containing words with high hyperbolicity. This suggests that there is less diversity in the intentions of hyperbole than there is in the expression of hyperbole.

Word	f_i	**V**	**A**	**D**
many	127	-	-	-
good	123	**0.938**	0.368	0.534
terrible	97	**0.061**	**0.849**	0.604
frustrate	92	**0.100**	**0.809**	**0.243**
great	87	**0.958**	0.665	**0.810**
bad	85	**0.125**	0.625	0.373
not	74	-	-	-
too	73	-	-	-
very	71	-	-	-
much	59	-	-	-
long	49	0.541	0.353	0.543
annoy	45	**0.094**	**0.765**	0.286
often	43	-	-	-
few	37	-	-	-
amaze	35	**0.896**	**0.843**	**0.783**

Table 3.9: **Top 15 words Added to Literal Interpretations - HyperTwit$_K$**

Word is the word added to a literal interpretation. ***V*** is the valence. ***A*** is the arousal. ***D*** is the dominance. Extreme values of **V**,**A**,**D** are in **boldface**.

Hyperbolic Tweet	**Interpretation**
should be strike 3 but the umpire **is blind**	should be strike 3 but the umpire **missed it**
seem like the real thing but i was so **blind**	seem like the real thing but i was so **naive**
im **deaf & blind** to the **[expletive]**	im **unaffected** by the **lies**
about to **make a blind man see the light** in a few seconds	about to **reveal something** in a few seconds

Table 3.10: **Hyperbolic Expressions of 'blind'**

Hyperbolic Tweet is the original Tweet. **Interpretation** is the literal interpretation created during annotation. Various intended meanings are observed for hyperbolic expressions containing the word 'blind'.

Hyperbolic Tweet	Interpretation
This app is so **toxic** haha	This app is so **nasty** haha
You ever finally let go of a **toxic** person and gain so much clarity	You ever finally let go of a **nasty** person and gain so much clarity
Leaving a **toxic** relationship is so painful but so necessary	Leaving a **nasty** relationship is so painful but so necessary

Table 3.11: **Hyperbolic Expressions of 'toxic'**

Hyperbolic Tweet is the original Tweet. **Interpretation** is the literal interpretation created during annotation. A singular intended meaning for hyperbolic expressions containing the word 'toxic' is observed.

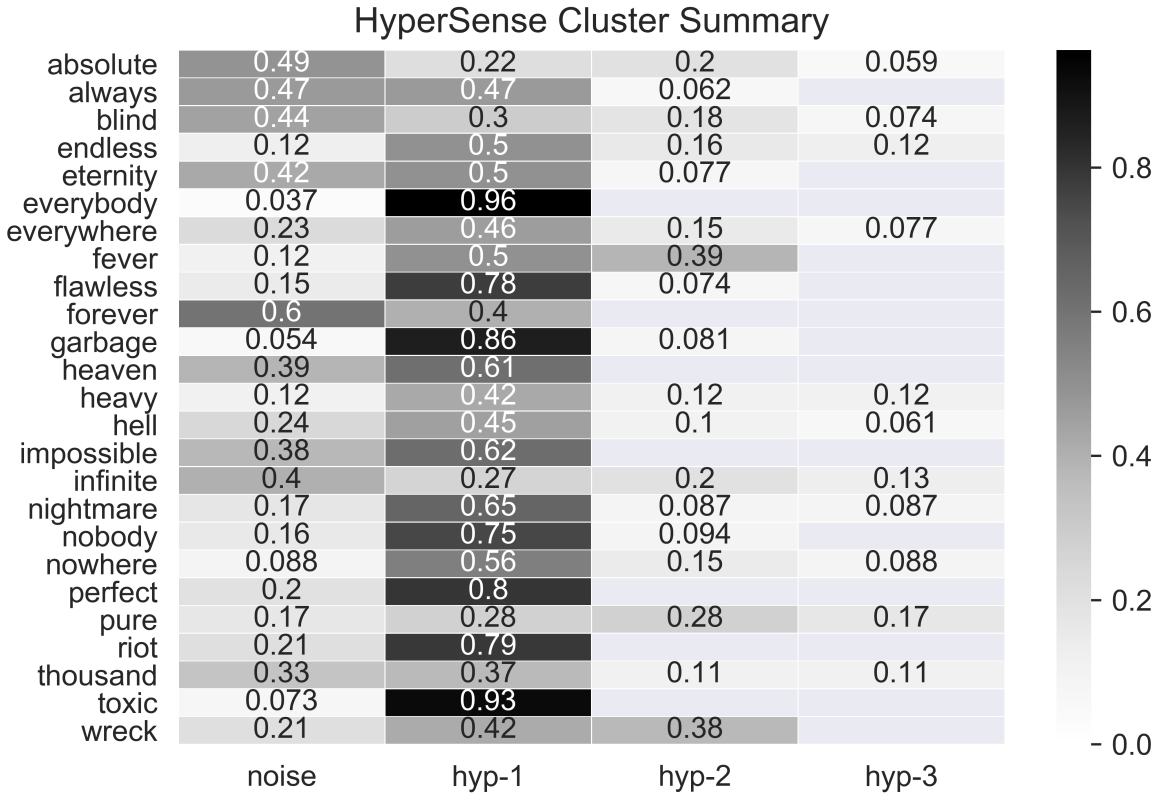

Figure 3.4: **Hyperbole Cluster Summary Heatmap**

Rows represent the keyword removed from a hyperbolic Tweet. Columns represent different clusters (*Noise* cluster contains unique/novel hyperboles. *hyp-1* is the largest non-noise cluster. *hyp-2* is the second largest non-noise cluster. *hyp-3* is the third largest non-noise cluster). Values represent the proportion of examples belonging to a cluster for all hypoerbolic examples of that keyword.

3.5.4 Diversity of Hyperbole

In this section an exploration of the diversity of hyperbolic expression in **HyperTwit** is undertaken. For example, the word *blind* appears in various hyperbolic expressions in **HyperTwit** with different intended meanings, see Table 3.10 in Appendix. Conversely, limited variation in hyperbolic expressions using the word *toxic* can be observed, see Table 3.11.

To probe the diversity of hyperbole expression manual clustering is performed based on both the hyperbolic expression and intended meaning. Keywords with high hyperbolicity from $\textbf{HyperTwit}_{KEY}$ were selected and for each keyword hyperbolic tweets where the keyword was removed were manually clustered. Similar examples are defined as those in which both the hyperbolic expression and the intended meaning are similar, see Tables 3.10 and 3.11 in Appendix. A summary of the manually identified hyperbole clusters are presented in Figure 3.4.

From this summary it can be seen that for a number of keywords, more than 75% of examples belong to a single hyperbole cluster (i.e., *everybody*, *flawless*, *garbage*, *nobody*, *perfect*, *riot*, *toxic*). This is explained by users parroting the same hyperbolic expressions for these keywords in similar contexts with similar interpretations (i.e., ***toxic*** -> ***nasty***/***unkind***, ***riot*** -> ***angry***/***annoyed***).

Several keywords are identified where at least a third of the examples belonged to the noisy cluster with no similar examples (i.e, *absolute*, *always*, *blind*, *eternity*, *forever*, *heaven*, *impossible*, *infinite*, *thousand*). Suggesting novel or complex hyperbolic uses of that particular keyword. Such as in elaborate hyperbole (***It is statistically impossible to have a bad day*** *if you start it listening to Pretty U by Seventeen*), unique adjective or adverbial phrases (i.e., *started an* ***infinite hate train***) as well as the combination of multiple hyperbolic expressions (i.e., *the connect in my stat class is* ***absolute dooki*** *im boutta kms*).

Keywords with several hyperbole clusters of considerable size can be observed (i.e., *absolute*, *blind*, *everywhere*, *fever*, *heavy*, *pure*, *thousand*, *wreck*.) Indicating various established hyperbolic uses for those keywords (e.g., *this show is a* ***train wreck*** -> *this show is* ***terrible*** , *i am an* ***emotional wreck*** -> *i am* ***upset***, *i want you to* ***wreck me*** -> *i* ***really*** *want you*). A manual cluster analysis of hyperbole identified considerable diversity in hyperbole usage on Twitter. This has implications for the automated detection of hyperbole, with the ability to deal with novel hyperbole a key requirement.

3.6 HyperProbe

The **HyperProbe** suite consists of synthetic test sentences generated to probe the behaviour of hyperbole detection models. The suite is created to expose the models to the three key types of hyperbole[145]; **Extreme Case Formulations**, **Qualitative Hyperbole** and **Quantitative Hyperbole**. The creation of the test sentences can be described by a general procedure consisting of four main steps;

i. **Word List Creation**

- Creation of word lists that will feature in the generated test sentences

ii. **Sentence Template Creation**

- Creation of sentence templates

iii. **Test Sentence Generation**

- Generation of test cases using CheckList[191][13] from the sentence templates and seed word lists

iv. **Manual Annotation and Assessment**

- Assessment of semantics and grammar of generated test sentences
- Annotation of generated test sentences

This general procedure is followed for the generation of the majority of test sentences contained in HyperProbe. However, in some exceptional cases this procedure produced poor results and different strategies for test generation were required.

3.6.1 Extreme Case Formulation Tests

Extreme Case Formulations are an important type of hyperbole that function unlike other expressions of hyperbole, see Section 3.2 for a discussion on ECFs. The ability to detect and interpret ECFs is a fundamental requirement for a hyperbole detection and interpretation model. To design test sentences to directly probe this ability, ECF words from Table 3.4 in Section 3.4 are grouped by part-of-speech category to form seed word lists. Grouping these words by part-of-speech categories results in four main categories (adjectives, adverbs, determiners and indefinite pronouns), for which tests are designed specific to each category.

[13] https://github.com/marcotcr/checklist

Test Name	Test Sentences	Hyperbole Proportion
ECF-Adjectives	73	0.520
ECF-Adverbs	39	0.590
ECF-Determiners	42	0.452
ECF-Indefinite Pronouns	27	0.556
Qual-Adjectives	306	0.284
Quant-Dimensions	43	0.488
Quant-Time Period	811	0.667
Quant-Time Period Quantities	108	0.555
Quant-Intrinsic Quantities	3460	0.476

Table 3.12: **HyperProbe Test Statistics**.

Test Name is the name of test, ***Test Sentences*** is the number of sentences in test, ***Hyperbole Proportion*** the proportion of sentences in the test that are hyperbolic.

3.6.1.1 Adjectives

A set of test sentences is designed to answer the following question; *can a model identify the hyperbolic and non-hyperbolic usage of adjectives in extreme case formulations?* The use of adjectives in hyperbolic ECFs was observed in HyperTwit (i.e., '*words are flowing out like* ***endless*** *rain in to a paper cup*', '*It is* ***impossible*** *to dislike Russell Wilson*', '*my* ***entire*** *life is just that one joke that goes over everyones head*'). From the list of ECF terms in Table 3.4 the following adjectives make up the seed list: *absolute*, *complete*, *definite*, *endless*, *entire*, *eternal*, *exact*, *flawless*, *impossible*, *infinite*, *invariable*, *invincible*, *perfect*, *pure*, *unconditional*, *undeniable*, *whole*. Two sentence templates are designed for generation of test sentences incorporating these adjectives, see Table 3.13. All adjectives (**{JJ}**) are drawn from the ECF adjective list previously defined in this section. Verbs ({VB}) are drawn from a user-defined verb seed list that contains simple linking verbs only (i.e., *is*, *was*). Likewise, the determiners ({DT}) are also drawn from a user-defined list containing simple determiners (i.e., *the*, *a*, *an*, *this* etc.). Note that regardless of test, {MASK} tokens are always infilled by a pre-trained language model, CheckList uses RoBERTa. Test sentences are then manually assessed to reject nonsensical sentences generated by CheckList and to annotate valid sentences, see Section 3.4.2 for details. Upon completion of assessment and annotation there were 73 test sentences, 38 (52%) of which were labelled as hyperbolic, see Tables 3.12 and 3.13.

Template	Example
{DT}{MASK}{MASK}{VB}{**JJ**}	the dishonest words are **endless**
{DT}{**JJ**}{MASK}{VB}{MASK}	the **endless** combinations are daunting

Table 3.13: **ECF Adjectives Test Sentence Templates**

Template shows templates as provided to CheckList, ***Example*** is an example sentence generated by CheckList.

Template	Example
{DT}{MASK}{MASK}{**RB**}{VB_a}	the code was **never** cracked
{DT}{MASK}{MASK}{**RB**}{MASK}	the good times **always** roll
{DT}{MASK}{VB_l}{**RB**}{MASK}	the dog was **never** silent
{DT}{MASK}{MASK}{VB_l}{**RB**}	the drug problem is **everywhere**

Table 3.14: **ECF Adverbs Test Sentence Templates**

Template shows templates as provided to CheckList, ***Example*** is an example sentence generated by CheckList.

3.6.1.2 Adverbs

A set of test sentences is designed to answer following question; *can a model identify the hyperbolic and non-hyperbolic usage of adverbs in extreme case formulations?* The use of an adverb to express a hyperbolic extreme case formulation was frequently observed in HyperTwit (i.e., '*Being a decent human being has gotten me* ***nowhere***', '*All I see is fake love* ***everywhere***'). From the list of ECF terms in Table 3.4 the following adverbs seed list is compiled: *always*, *everywhere*, *never*, *nowhere*. Four sentence templates are created to incorporate these adverbs into a sentence, see Table 3.14. All adverbs (**{RB}**) are drawn from the ECF adverb list previously defined in this section. Linking verbs and action verbs are introduced into the sentence templates for adverb testing. Linking verbs ({VB_L}) are drawn from a user-defined verb seed list that contains simple linking verbs only (i.e., *is*, *was*). Whilst, action verbs ({VB_a}) are drawn from a user-defined verb seed list that contains various action verbs only (i.e., *burning*, *falling*, *flying*, etc.). Determiners ({DT}) are also drawn from a user-defined list containing simple determiners (i.e., *the*, *a*, *an*, *this* etc.). Upon completion of generation, assessment and annotation there were 39 test sentences, 19 (45%) of which were labelled as hyperbolic, see Tables 3.12 and 3.14.

Template	Example
{DT}{MASK}{MASK}{**DT**}{MASK}	The moral of **every** story
{**DT**}{MASK}{MASK}{IN}{MASK}	**all** comments may be removed
{**DT**}{MASK}{VB}{MASK}{MASK}	**every** person will be disappointed

Table 3.15: **ECF Determiners Test Sentence Templates**

Template shows templates as provided to CheckList, ***Example*** is an example sentence generated by CheckList.

3.6.1.3 Determiners

A set of test sentences is designed to answer the following question; *can a model identify the hyperbolic and non-hyperbolic usage of determiners in extreme case formulations?* Extreme determiners were commonly observed in hyperbolic expressions in HyperTwit (i.e., '*My dorm so damn far from **every** fucking thing*', '*he says SO MANY WORDS containing **NO** INFORMATION*'). From the list of ECF terms in Table 3.4 the following determiner seed list was created: *all*, *no*, *every*. Three sentence templates are created to incorporate these particular determiners into a sentence, see Table 3.15. Two seed word lists are defined for determiners, one containing generic determiners ({DT}) (i.e., *the*, *a*, *an*, *this* etc.) and one containing the determiners of interest (**{DT}**) (i.e., *all*, *no*, *every*). Verbs ({VB}) are drawn from a user-defined verb seed list that contains simple linking verbs only (i.e., *is*, *was*). A manually defined list is defined from which to draw prepositions ({IN}) (i.e., *in*, *on*, *at*, etc.). Upon completion of generation, assessment and annotation there were 42 test sentences, 23 (59%) of which were labelled as hyperbolic, see Tables 3.12 and 3.15.

3.6.1.4 Indefinite Pronouns

A set of test sentences is designed to answer the following question; *can a model identify the hyperbolic and non-hyperbolic usage of indefinite pronouns in extreme case formulations?* The usage of indefinite pronouns to express hyperbolic extreme case formulations was frequently observed in HyperTwit (i.e.,' *i can guarantee **nobody** gives a single shit about ur zodiac sign*', '***Everybody** got an attitude this morning ! Ok I respect it*'). From the list of ECF terms in Table 3.4 the following indefinite pronoun seed list is compiled: *everyone*, *everybody*, *nobody*, *no one*. Two sentence templates are created to incorporate these indefinite pronouns into a sentence, see Table 3.16. Indefinite pronouns (**{PRON}**) are drawn from the ECF seed list (i.e., *everyone*, *everybody*, *no one*, *nobody*). Determiners

Template	Example
{DT}{MASK}{MASK}{MASK}{**PRON**} {**PRON**}{IN}{DT}{MASK}{VB}{MASK}	The child answers to **nobody** **everybody** in the house is bored

Table 3.16: **ECF Indefinite Pronouns Test**

Template shows templates as provided to CheckList, ***Example*** is an example sentence generated by CheckList.

({DT}) are also drawn form a user-defined list containing simple determiners (i.e., *the*, *a*, *an*, *this* etc.). Verbs ({VB}) are drawn from a user-defined verb seed list that contains simple linking verbs only (i.e., *is*, *was*). A list from which to draw prepositions ({IN}) (i.e., *in*, *on*, *at*, etc.) is manually defined. Upon completion of generation, assessment and annotation there were 27 test sentences, 15 (55%) of which were labelled as hyperbolic, see Tables 3.12 and 3.16.

3.6.2 Qualitative Hyperbole Tests

Qualitative hyperboles are an important type of hyperbole that are defined by an intentionally excessive qualitative contrast, (see Section 3.2 for a further discussion on qualitative hyperbole). The ability to detect and interpret qualitative hyperbole is a fundamental requirement of a hyperbole detection and interpretation model. Qualitative terms from Table 3.4 in Section 3.4 were used to form seed word lists for generating tests for qualitative hyperbole. Unlike ECFs, the qualitative terms in Table 3.4 predominantly function as adjectives. Given this observation only a single adjectives test is implemented for qualitative hyperbole.

The rationale behind these tests is to design sentences that allow us to answer the following question; *can a model identify the hyperbolic and non-hyperbolic usage of adjectives?* The use of extreme adjectives to provide exaggerated descriptions was a common and varied method of expressing hyperbole in HyperTwit (i.e., '*Dawg this ballpark is so* ***garbage***', '*Dating legit be a* ***headache***', '*Kyrie Irving is fucking* ***cancer*** *lmao*'). From the list of qualitative terms in Table 3.4 list containing 54 adjectives is compiled. Six sentence templates are defined to incorporate the adjectives into a sentence, see Table 3.17. All adjectives (**{JJ}**) are drawn from the qualitative adjective list previously defined in this section. Verbs ({VB}) are drawn from a user-defined verb seed list that contains simple linking verbs only (i.e., *is*, *was*). Likewise, the determiners ({DT}) are also drawn form a user-defined list containing simple determiners (i.e., *the*,

Template	Example
{DT}{MASK}{MASK}{VB}{MASK}{**JJ**}	an idea that is very **wicked**
{DT}{MASK}{VB}{**JJ**}	The argument is **confusing**
{DT}{MASK}{VB}{MASK}{**JJ**}	The wine is very **bitter**
{DT}{MASK}{MASK}{VB}{**JJ**}	the oil residue is **toxic**
{DT}{**JJ**}{MASK}{VB}{MASK}	A **great** story was completed
{DT}{**JJ**}{MASK}{VB}{MASK}{MASK}	The **shocking** speech was poorly prepared

Table 3.17: **Qualitative Adjectives Test Sentence Templates**

Template shows templates as provided to CheckList, ***Example*** is an example sentence generated by CheckList.

a, *an*, *this* etc.). Upon completion of assessment and annotation there were 306 test sentences, 87 (28%) of which were labelled as hyperbolic, see Tables 3.12 and 3.17.

3.6.3 Quantitative Hyperbole Tests

The understatement or overstatement of quantitative values could be considered the prototypical example of hyperbole given the trivial identification and interpretation of the hyperbole (i.e., 'im so hungry i could eat **1000** pizzas'), (see Section 3.2 for futher discussion on quantitative hyperbole). Accurate detection and interpretation of quantitative hyperbole is a required capability for computational models that process hyperbole. The process for generating test sentences varied compared to that of the other tests in HyperProbe. A key reason for this was the difficulty in generating sufficiently varied test sentences when following the general four step procedure for test generation. Specifically, the range of topics covered by the test sentences generated via CheckList was limited to financial and business topics (i.e., dollars, units, stocks, etc.,).

3.6.3.1 Quantitative Comparisons

A set of test sentences are designed to address the following question; *can a model identify plausible and non-plausible comparisons of objects along quantitative dimensions?*. Note, that this test does not specifically target hyperbolicity but rather plausibility of object comparisons. An understatement or overstatement of quantitative values was an identified pattern of hyperbole expression in HyperTwit, (see Section 3.5).

From the list of terms in Table 3.4 list containing all words relating to quantitative dimensions (i.e., *big*, *small*, *light*, *heavy*, *thin*, *thick*, etc.) is compiled. In addition to

Template	Example
{MASK}{MASK} is as {**JJ**} as {MASK} {MASK}	her eyes are as **blue** as the ocean
{MASK}{MASK} is {**JJR**} than {MASK}{MASK}	that building is **taller** than I thought

Table 3.18: **Quantitative Dimensions Test Sentence Templates**

Template shows templates as provided to CheckList, ***Example*** is an example sentence generated by CheckList.

this, the comparative form of each word is used (i.e., *bigger*, *smaller*, *lighter*, etc.). Two sentence templates are designed to incorporate these words into a sentence, see Table 3.18. These two sentence templates are more specific than the templates defined for other tests with several fixed natural language words and phrases (i.e., 'is a', 'than'). All adjectives (**{JJ}{JJR}**) are drawn from the adjective list previously defined in this section. Upon completion of assessment and annotation there were 43 test sentences, 21 (48%) of which were labelled as hyperbolic, see Tables 3.12 and 3.18.

3.6.3.2 Time Periods

A set of test sentences is designed to answer the following question; *can a model identify hyperbolic expressions of time periods?*. The overstatement and understatement of time periods was a pattern of hyperbole expression observed in HyperTwit (i.e., '*4* ***years*** *of Trump has seemed like 40* ***years*** *of hell.*', '*This past 12* ***hours*** *has been one of the longer* ***decades*** *of my life*' '*I only been at work for 30****mins*** *and it feel like I been here for* ***hours***'). The design of these test sentences differs from the four step procedure for sentence generation followed in Section 3.6.1 and 3.6.2.

A list of terms relating to time periods from Table 3.4 is established (i.e., *hour*, *day*, *week*, *today*, *yesterday*, etc.), from which all **{NN}** in the following sentence templates are drawn from. Also, a list of proper nouns relating to time periods (i.e., *Monday*, *Tuesday*, *January*, *February*, etc.), from which all **{PROPN}** in the sentence templates are drawn from. An additional list of comparative adjectives ({JJR}) is created containing only *less* and *more* to create test sentences comparing the length of different time periods. Four sentence templates are created see Table 3.19. The generation of test sentences for these sentence templates is based on all combinations of words in both the **{NN}**, **{PROPN}** and {JJR} lists. This results in 811 test sentences, 541 (67%) of which were considered to be hyperbolic, (i.e., an understatement or overstatement of the length of time in a time

Template	Example
that **{NN}** lasted a **{NN}**	that **day** lasted a **month**
{PROPN} lasted a **{NN}**	**December** lasted a **year**
that **{NN}** lasted {JJR} than a **{NN}**	that **day** lasted less than a **month**
{PROPN} lasted {JJR} than a **{NN}**	**December** lasted more than a **year**

Table 3.19: **Quantitative Time Periods Test Sentence Templates**

Template shows templates as provided to CheckList, ***Example*** is an example sentence generated by CheckList.

periods).

To further probe the understanding of time periods four additional sentence templates are created, see Table 3.20. Cardinal numbers are introduced in these templates to generate test sentences that specify quantities of time periods. Numeric {CD_N} and alpha {CD_A} cardinal numbers are used to generate test sentences. Constraints are employed to avoid an excessive number of test sentences when these generating sentences using these templates:

i. **Constraint on Time Period Pairs:** All templates contain two time periods, the first of which is filled with all time periods in the **{NN}** and **{PROPN}**, the second time period is only filled with next smallest time period by duration (i.e., minutes and seconds, hours and minutes, days and hours, etc.). This avoids redundant test sentences that are already covered by the 811 test sentences previously generated for time periods.

ii. **Constraint on Cardinal Numbers:** The choice of cardinal numbers is limited by the orders of magnitude that bound the equality comparison between the two time periods. (e.g. 1 *hour* = 60 *minutes* : that *hour* lasted {JJR} than **10/100/ten/one hundred** *minutes*, 1 *week* = 7 *days*: that *week* lasted {JJR} than **1/10/one/ten** *day*(s))

This process resulted in 108 sentences, of which 60 (55%) were considered to be hyperbolic. A hyperbolic test sentence in the context of this test is one in which the comparison of time periods is excessive (e.g., that day lasted more than a week).

3.6.3.3 Intrinsic Values

A set of test sentences is designed to answer the following question; *can a model identify plausible and non-plausible ranges of quantitative values?*. A common expression of

Template	Example
that **{NN}** lasted {JJR} than a {CD$_N$} **{NN}**	that **month** lasted more than 100 **days**
that **{NN}** lasted lasted {JJR} than a {CD$_A$} **{NN}**	that **month** lasted less than one hundred **days**
{PROPN} lasted {JJR} than a {CD$_N$} **{NN}**	**May** lasted more than 100 **days**
{PROPN} lasted lasted {JJR} than a {CD$_A$} **{NN}**	**April** lasted less than one hundred **days**

Table 3.20: **Quantitative Time Periods (Numeric) Test Sentence Templates**

Template shows templates as provided to CheckList, ***Example*** is an example sentence generated by CheckList.

hyperbole is the overstatement or understatement of quantitative values (i.e., '*is the chargers team doctor still that dude with like a **billion** pending lawsuits?*', '*I enter 10 **billion** giveaways and end up winning absolutely NOTHING*'). Understanding the valid distribution of values for objects along various quantitative dimensions is a desired behaviour of a hyperbole detection and interpretation model. To test this capability data is leveraged from research that looks at constructing quantitative distributions of objects via large scale internet crawling and data processing [51]. The authors evaluate a subset of their data across 4 quantitative dimensions (currency, length, mass and speed) via crowd-sourcing and this evaluations are transformed into simple sentences.

The original template for the question posed to the crowd workers was as follows; '*Does the {MEASUREMENT of a/an {OBJECT fall within the range of {NUMBER {UNIT?*'. Four sentence are created templates to represent the questions and answers in a single test sentence that alings with other sentences in **HyperProbe**, see Table 3.21. The first step in the transformation process is to extract the necessary data values from the original data (i.e., *{MEASUREMENT*, *{OBJECT*, *{NUMBER*, *{UNIT*). This process is straightforward except for the extraction of *{NUMBER* due to the need to get the minimum value and maximum value from this number. Once extracted these values are used to infill the templates for both the minimum and maximum value of the specified range in *{NUMBER*, see 3.22. The accepted answer is a majority vote of all answers provided by the crowd workers. If the accepted answer is that the value does not fall within a reasonable range for that object along that quantitative dimension then this can be labelled as a hyperbole. After transformation of 3458 sentences, 1646 (47%) of were considered to be hyperbolic (i.e., an understatement or overstatement of the intrinsic

Dimension	Template
Currency	That {OBJECT cost {NUMBER {UNIT to buy
Length	That {OBJECT is {NUMBER {UNIT long
Mass	That {OBJECT weighs {NUMBER {UNIT
Speed	That {OBJECT is travelling {NUMBER {UNIT

Table 3.21: **DoQ-Intrinsic Sentence Templates**

Dimension Quantitative dimension, ***Template*** shows transformation template from DoQ.

Dimension	Example	Label
Speed	That motorcycle is travelling one hundred kilometers an hour	0
Speed	That motorcycle is travelling one thousand kilometers an hour	1
Mass	That lizard weighs one hundred kilograms	0
Mass	That lizard weighs one thousand kilograms	1
Length	That kitchen is one centimeter long	1
Length	That kitchen is ten meters long	0
Currency	That footwear costs ten dollars to buy	0
Currency	That footwear costs ten million dollars to buy	1

Table 3.22: **DoQ-Intrinsic Test Sentence Example**

Dimension Quantitative dimension, ***Example*** is an example sentence generated by CheckList.

value of an object along a quantitative dimension).

3.7 Conclusion

This chapter provides the background and motivation for the computational study of hyperbole and how it fits within the broader thesis topic of computational understanding of figurative language. Key points discussed include:

- Hyperbole is one of the most common figures of speech, particularly in informal situations, heightening the importance of understanding the phenomenon on social media
- Hyperbole has received considerably less attention in both the linguistics and NLP communities relative to metaphor and irony. Hyperbole is relatively misunderstood

as are the best approaches for the computational study of the figure.

The introduction of both **HyperTwit** and **HyperProbe**, which provide more benchmarks for evaluating the ability of NLP systems to accurately detect hyperbolic expressions. This contribution satisfies research objective i), as these annotated datasets allow for the *Assessment of the capabilities of existing NLP methods in detecting and interpreting hyperbole on social media and idiomatic forms of hyperbole*

A significantly greater prevalence of hyperbole was observed on Twitter compared to that found in corpus studies on hyperbole in different communicative forms (i.e., conversational English). Hyperbole was commonly used on Twitter to convey strong sentiment, which highlights the importance of understanding hyperbole for computational tasks concerned with identifying affective content in text (i.e., sentiment analysis).

A detailed look at the diversity of hyperbolic expression on Twitter identified examples of hyperbolic expressions that were simply parroted by different Twitter users but also a number of novel, elaborate and specific hyperbole. This finding indicated that adapting to novel hyperbole expressions is a key challenge in computational analysis and understanding of hyperbole.

These observations provided evidence for research question i) and satisfied research objective ii), specifically the findings provided quantitative evidence of how *figurative language occurs on social media in the context of hyperbolic expressions*.

CHAPTER 4

TOWARDS COMPUTATIONAL HYPERBOLE DETECTION

4.1 Introduction

This chapter details efforts to develop models for the computational detection of hyperbolic language, with a focus on online user-generated content (i.e. Twitter posts). The content in this chapter addresses the research questions and aims of this thesis, (see Sections 1.3 and 1.4), in the following ways:

i. The evaluation of several existing models for hyperbole detection to *assess the adequacy of existing NLP models on the detection of figurative language on social media*. (Research Question ii, Research Objective ii)

ii. The proposal and empirical evaluation of several new models for hyperbole detection seeks to provide answers on *how to improve the computational detection of figurative language on social media*. (Research Question iii, Research Objective iii)

iii. Detailed error analysis that focuses on the explainability of model decisions seeks to *identify how models for the detection of figurative language can be improved further* (Research Question iii, Research Objective iii)

iv. Cross-domain experiments provide insights on *how the expression of hyperbole differs on social media in comparison to the occurrence of figurative language in traditional forms of communication* (Research Question i, Research Objective i)

The chapter is structured as follows:

- Section 4.2 provides the motivation for models that can detect hyperbole in online content, a review of literature on figurative language and hyperbole detection is covered also covered.
- Section 4.3 details the implementation of various baseline methods for hyperbole detection. The proposals, and implementation details, for two novel approaches for hyperbole detection are also introduced for the first time in this thesis.
- Section 4.4 outlines several experiments that test the accuracy of models trained and evaluated in various settings (i.e., in-domain, out-of-domain). The results and of these experiments are also presented in this section.
- Section 4.5 concludes the content in this chapter.

4.2 Motivation and Methods for Hyperbole Detection

4.2.1 Motivation for Hyperbole Detection

The computational detection of hyperbolic content in text has many benefits for various tasks in Information Retrieval (IR) and Natural Language Understanding (NLU) [3, 101, 222]. Such as better understanding and generation of hyperbolic expression leading to better experiences with chat-bots [222], to improved sentiment analysis and recommender systems that rely on social media [3]. The importance of hyperbole, and other figures of speech, in health communication, see Part I, heightens the importance of computational methods that can detect hyperbole. Further, a key finding from Chapter 2 in this thesis was that hyperbolic expressions of health concepts remained a challenge for both the existing and proposed text classification models.

The prevalence of hyperbole in online content also motivates the importance of methods that can automatically detect hyperbole. As shown in Chapter 3, over a 30-day period an average of 15% of Twitter posts collected for **HyperTwit** contained hyperbolic content. This hyperbolic content was often difficult to interpret and contained strong sentiment bearing opinion. A holistic understanding of discourse on Twitter requires accurate interpretations of hyperbole and the computational detection of hyperbolic content is a stepping stone to achieving this.

The task of detecting the presence of hyperbolic content in a short fragment of text has been posed as a supervised binary sequence classification task [101, 222], similar to other approaches for the detection of other figures of speech [3, 90]. A lack of datasets on

hyperbole and few efforts to study the computational detection of hyperbole has resulted in a gap in the literature on computational understanding of figurative language, see Chapter 3.

4.2.2 Methods for Figurative Language Detection

Methods for the detection of hyperbole have seen similar approaches to the detection of other figures of speech. The most common approach to the detection of figurative language follows the traditional NLP pipeline approach to text classification in a supervised setting. Generally, features are manually engineered based on the linguistic markers of a particular phenomenon (i.e., irony) then combined with general representations of textual content common to many NLP approaches [3, 11, 92, 222].

4.2.2.1 Feature Engineering

The feature generation step of the traditional figurative language classification pipeline has been the central focus of many methodologies for figurative language detection [3, 89]. The features in this stage are often motivated by findings from cognitive linguistics on the mechanisms and cues humans use to identify the particular phenomenon (i.e., metaphor) of interest. As a result, this step of the traditional pipeline approach has seen more variation than most other steps in the pipeline approach.

A feature often used in these pipelines is based on exploiting the presence of linguistic patterns that are common to the figure of speech of interest. The presence of particular sarcasm patterns have been introduced as features in sarcasm detection pipelines. Such as the appearance of a positive verb within the context of a negative situation [192], beginning a phrase with an interjection [16], an interjection followed by "I" [2] or an interjection followed by an intensifier [3]. The presence of words with opposite sentiment within a sentence is a pattern that has been exploited for irony detection [99]. The presence of hyperbolic markers have been used in model pipelines for the detection of sarcasm and irony. Such as the appearance of intensifiers, punctuation, interjections, and contiguous sequences of words with strong sentiment content [10, 16, 25, 224]

Incongruity is an important feature in some figures of speech and as such has been used as a feature in various figurative language detection pipelines [3, 91, 246]. The maximum semantic distance between pairs of words in a sentence and the minimum semantic distance of word-pairs in a sentence have been used to represent the semantic incongruity for detecting humorous figures of speech [246]. The incongruity of sentiment

has also been used as a feature, particularly for irony and sarcasm and has been represented as a feature in various ways [91, 141]: a frequency count of consecutive word pairs with opposite polarity, the frequency counts of words with positive and negative polarity or the length of longest sequence of words with contiguous polarity or lack of polarity.

Unexpectedness and ambiguity are two features often used in various pipelines for figurative language detection [11, 127, 189, 222]. Unexpectedness has been used for hyperbole detection by computation of the pairwise cosine similarity between the representations of all word pairs in a sentence [222]. The use of semantic relatedness between words in a sentence was used a measure of contextual imbalance for detection of ironic social media posts [188]. Various features were computed that indicate the number of possible senses of words within a sentence as a measure of ambiguity for irony detection [11]. Homophones and homographs are examples of ambiguous language usage [39, 223]. The annotation of these phenomena and the training of Naive Bayes and SVMs to detect these linguistic phenomena are key contributions of a work focusing on the detection of humorous figures of speech [226].

Features to represent sentiment have played a key role in many approaches to figurative language detection [3]. Frequency counts of the positive and negative words in a piece of text is a common feature in many models [1, 3, 91, 119, 122, 222]. These polarity of individual words has been computed by looking up sentiment lexicons such as SentiWordNet[9], Linguistic Inquiry and Word Count (LIWC) [219], General Inquirer [216], WordNet-affect [214] and several others. In addition to using lexicons to represent sentiment, the prediction of sentiment signals using various classification frameworks for figurative language detection has also been explored. Such as the use of CNNs trained on existing corpora to detect the signals of sentiment (i.e. positive, neutral, negative), emotion (i.e., anger, disgust, sadness, fear, joy, surprise) and personality type (i.e., openness, conscientiousness, extraversion, agreeableness, and neuroticism) within text for sarcasm detection [175]. More specific features have been computed for sentiment such as sentiment conflict and sentiment transitions for humour detection [122]. This work looks at the sentiment of various elementary discourse units found using a discourse parser [53] and how that sentiment changes between various discourse units throughout a short text. The idea of sentiment conflict is similar to idea of sentiment incongruity, which has been a key feature in methods for sarcasm detection [1, 3, 89, 91].

4.2.2.2 Word Representations

The type of general purpose representations used in models for detecting figurative language are more generic. Many approaches rely on the bag-of-words model using various n-gram sizes and weighting methods to generate sparse representations of linguistic content [52, 54, 59, 121, 159].

Pre-trained dense word representations have been utilised for various figurative language detection models. A number of features that look at the pairwise similarities between the pre-trained dense representations of words for sarcasm detection showed the value of such word representations for sarcasm detection [93]. The authors experiments with various methods for computing dense word representations such as GloVe[166], word2vec[138] and Dependency-based representations[115]. Pre-trained representations, Doc2Vec [112], were shown to be effective detecting satirical news in larger text sequences [184]. Pre-trained dense word representations have also been used as features for hyperbole detection [101, 222], see Section for more details. Researchers have experimented with computing their own dense word representations for their particular task rather than using general pre-trained representations. Random initialisation and training of dense word representations using LSTMs was explored for irony detection [242], this approach was used by the best performing method on an irony detection task at Semeval [227].

Large scale pre-trained language models have been successful across a wide variety of NLP tasks [125, 194] and have seen similar success in the detection of figurative language [13]. The combination of representations computed by BERT with features engineered to capture various linguistic phenomena (e.g., unexpectedness, abstractness, objectivity, etc.) was shown to be successful for hyperbole detection [101]. A model for the detection of metaphorical verbs was proposed that utilised representations computed by BERT [211]. A hierarchical model based on BERT was proposed to detect sarcasm given a pair of short texts (i.e. context and response) [213]. Probing experiments on the ability for BERT to differentiate the between plausible sentences and non-plausible metaphorical sentences showed that BERT was able to distinguish between these sentences and assign plausibility ratings similar to human annotations [163].

4.2.2.3 Learning Algorithms

A vast array of learning algorithms have been used in frameworks for figurative language detection with varying degrees of success. A common approach in the traditional pipeline

approaches to figurative language detection involved testing the pipeline with various learning algorithms for classification and choosing the best performing for the particular task. Support Vector Machines (SVM), Decision Trees (DT), K-nearest Neighbour (KNN), Naive Bayes, Logistic Regression, Latent Dirichlet Allocation (LDA) have commonly been tested as classifiers in many approaches for detection of various figures of speech [3, 12, 84, 89, 92, 141, 222]. The best performing learning algorithm is often dependent on the particular pipeline, dataset and figure of speech for which the task is focused on.

Approaches to hyperbole detection based on deep learning models have been also developed for figurative language detection such as irony detection [81], sarcasm detection [59] and metaphor detection [241]. Related to the detection of hyperbole, incorporation of deep learning models into an architecture for detecting hyperbole in Mandarin Chinese was shown to provide improvements in accuracy [101]. The authors also found that it was unclear if the hand-crafted features of Troiano [222] were actually effective when combined with deep learning architectures.

The addition of deep learning algorithms into figurative language detection frameworks has seen improvements to performance over the use of traditional learners. The detection of metaphor at the sequence level was proposed using both CNNs and LSTMs [241] whilst an approach to the detection of metaphor at the token-level utilising LSTM and CRF was proposed [177]. These approaches showed improvements over existing methodologies for the respective tasks and showed the utility of deep learning algorithms for metaphor detection. Similar to approached the metaphor detection, sarcasm detection has seen various approaches that rely on deep learning algorithms for classification. The detection of self-deprecating sarcasm employed LSTMs, a model based on CNNs was used for the detection of general sarcasm on Twitter, the combination of CNN+LSTM+fully connected neural network layers was highly successful at detecting sarcasm on Twitter. A comparison of various frameworks for hyperbole detection compared CNN, LSTM to traditional learning algorithms (i.e. LR, SVM, KNN, NB, DT, LDA) showed that the deep learning algorithms were able to outperform the traditional learners on detecting hyperbole in Mandarin chinese [101].

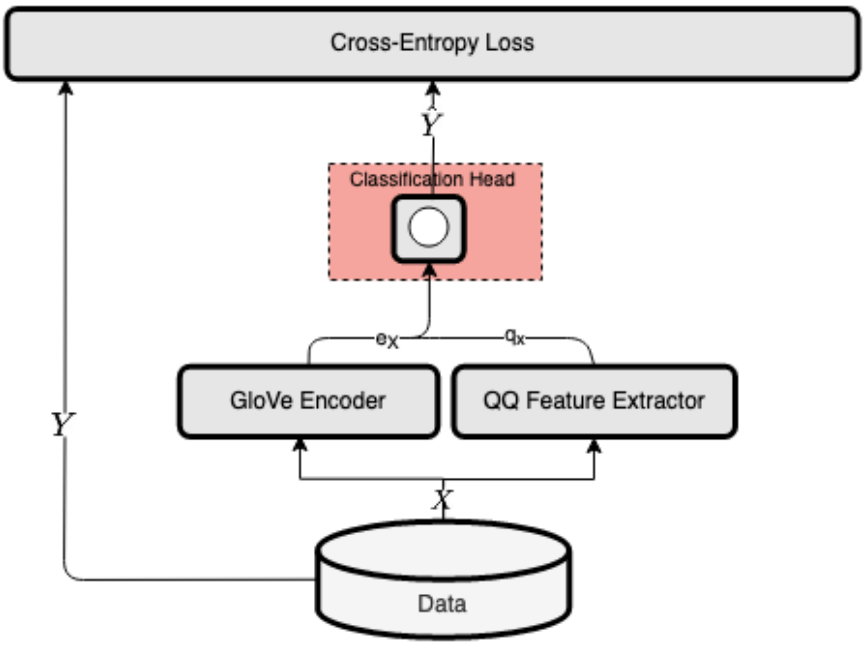

Figure 4.1: **LR+QQ Model Diagram**

The LR+QQ model contains a GloVe encoder, feature extractor module and linear classification head.

4.3 Methodology

4.3.1 Baselines

In this section several baseline models for hyperbole detection alongside implementation details are described.

In the foundational work on computational hyperbole detection an NLP pipeline style approach was proposed [222], (see Figure 4.1). The authors introduce a several manually engineered features motivated by cognitive linguistics research on the approaches humans use for detecting and understanding hyperbolic language.

The authors consider unexpectedness to be an important aspect of hyperbole and intend to measure this in a sentence by representing each word in an utterance by its non-contextual word representation (GloVe[166] and Word2Vec [138]) and compute the pairwise cosine similarity between the representations of all word pairs in the utterance. To measure the degree to which a word evokes mental imagery, denoted imageability, the authors turn to the MRC psycholinguistic database [36, 239] to compute an imageability score for all words in a sentence and then average these scores to get

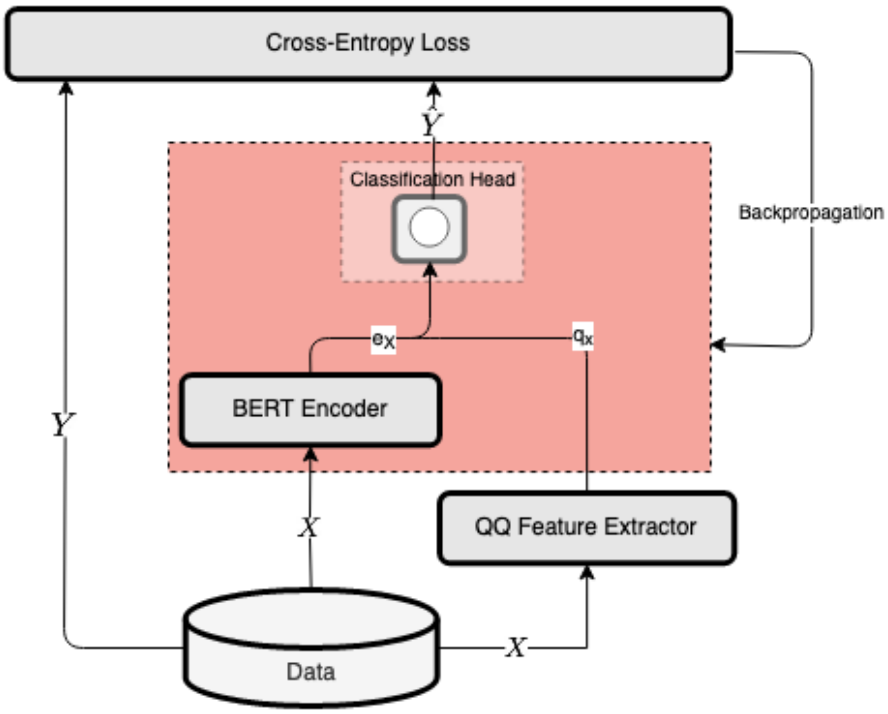

Figure 4.2: **BERT+QQ Model Diagram**

The BERT+QQ model contains a BERT encoder, feature extractor module and linear classification head.

an imageability score. To represent polarity, the authors use the TextBlob[1] toolkit to compute the sentiment for each word in a sentence then average all words in a sentence to compute a final polarity score. Also using in-built functionally of TextBlob, the authors compute whether a sentence represents a subjective or objective stance and refer to this as the polarity score. Lastly, the authors use the VADER[2] toolkit to quantify the intensity of the emotion within a sentence and refer to this as the emotional intensity score. These features are concatenated together and denoted as Qualitative and Quantitative (QQ) by the authors. A number of learning algorithms are used at the classification layer of their pipeline in experiments conducted on hyperbole detection such as Support Vector Machine, Nearest Neighbour, Decision Trees, Logistic Regression, Naive Bayes and Latent Dirichlet Allocation. With Logistic Regression and Naive Bayes proving to be the best choice for the classification layer as shown in their experiments. Two versions of this NLP pipeline are used as baselines in this chapter; **LR+QQ** refers to the pipeline

[1]https://textblob.readthedocs.io/en/dev/#
[2]https://github.com/cjhutto/vaderSentiment

with Logistic Regression at the classification layer and **NB+QQ** refers to the pipeline with Naive Bayes at the classification layer.

The NLP pipeline, particularly the manually engineered features, were utilised in a work on hyperbole detection in Mandarin Chinese [101]. These features were adjusted to compensate for language differences and incorporated into a deep learning framework. A pre-trained language model (i.e., BERT) is used as an encoder with this encoded representation combined with the QQ features and fed to a classification layer, (see Figure 4.2). This model is referred to as **BERT+QQ** in the remainder of the chapter, (see 4.2). A vanilla BERT baseline is included in experiments and referred to as **BERT** in the remainder of this chapter.

4.3.2 Affective Signals for Hyperbole Detection

Sentiment and affective signals have been an important feature in many approaches to the detection of figurative language, see Section 4.2.Findings from both Chapters 2 and 3 motivate the importance of affective signals in hyperbolic expressions. Particularly the observation that the affective content of words commonly used in hyperbolic expressions was contained strong affect, see Tables 3.6 and 3.7. From these tables it can be seen that words with high hyperbolicity (i.e., words prone to hyperbolic usage) also had extreme values in one or more of the affective dimensions of valence, arousal and dominance. A hypothesis here is that effective incorporation of affective signals into a hyperbole detection model would improve the ability of that model to accurately detect hyperbolic utterances.

For the calculation of affective signals, techniques are leveraged from research that introduced an annotated dataset, **4dEmotionsInTwitter**[3], of tweets labeled for the strength of arousal, valence, dominance and surprise in individual tweets [240]. The authors train Support Vector Machines to predict the strength of the signal for each of the four affective dimensions, based on the implementation of [7], showing that the affective content of a Tweet can be predicted[4]. These regressors were trained following a traditional NLP pipeline approach.

The feature engineering step of this pipeline consisted of various hand crafted linguistic features. The authors computed the average, minimum and maximum GloVe representation for all words in a Tweet for their general purpose dense representations. In addition to the dense representations, the authors also combined sparse representa-

[3]http://140.203.155.26/mixedemotions/ datasets/4dEmotionInTweets.tar.gz

[4]The results for surprise were poor so that dimension was excluded

tions in the form of frequency counts of all 1-gram to 4-grams in a Tweet. A number of hand-crafted features were also concatenated to these representations. Including the proportion of capitalised tokens and the proportion of words that being with a capital letter. The authors also computed average, min and max difference vectors based on the cosine similarity between all words in a Tweet and a GloVe representation of various emotions (i.e., the GloVe representation of the word "fear"). All these features were concatenated together and a Support Vector Machine was used as the final regression layer to predict the affective signal for each dimension [7].

The predictions for the valence, arousal and dominance produced by these regressors are used as affective signals for models in various configurations. Three models are introduced that incorporate affective signals into the modeling framework at various stages of the respective frameworks. **BERT+3dEmo** and **BERT+3dEmo**$_{AS}$ are both sequence classification models but differ in the way that the affective signals are incorporated. In **BERT+3dEmo**, the affective signals are concatenated with the output of the dense word representations produced by the encoder. However, in **BERT+3dEmo**$_{AS}$, the affective signals are introduced as special tokens in the original text sequence and this information is encoded in the dense representations. **BERT+3dEmo**$_{MT}$ is a multi-task classification framework where the affective signals are used as soft-targets with the model being trained to distill knowledge form the affective signals. Further descriptions of these three models will be covered in the following sections.

4.3.2.1 BERT+3dEmo

BERT+3dEmo is based on the traditional approach seen in many figurative language detection frameworks. In this approach the features are computed separately and combined just before the learning algorithm, a classification layer, in this case. BERT is used as an encoder to compute dense word representations of the Tweet content. The affective signals are concatenated and passed through a dense layer before being concatenated with the BERT representations then passed to a final classification layer, see Figure 4.3.

Formally, the logits for an individual Tweet are calculated according to

$$\hat{y} = \sigma(e_i \mathbf{W}^Y + a_i \mathbf{W}^Z + b^Y) \tag{4.1}$$

where e_i is the dense representation of tweet i computed by BERT [43], a_i is the affective signals of Tweet i as computed by the pre-trained regressors, $\mathbf{W}^Y$, $\mathbf{W}^Z$ an b^Y are learnable

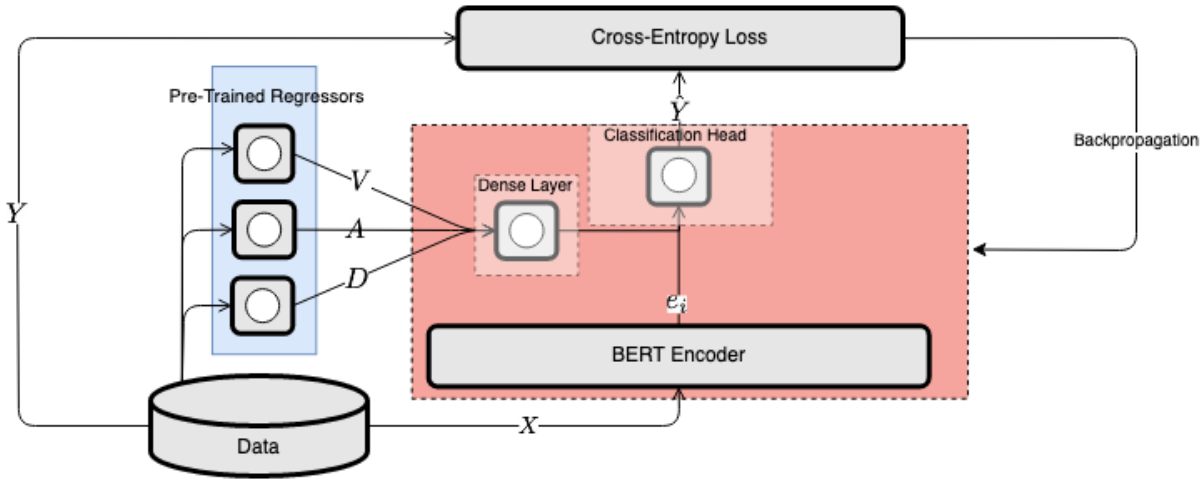

Figure 4.3: **BERT+3dEmo**

Model contains a BERT encoder, a dense layer and a a linear classification head. Signals from the pre-trained regressors a feed into a dense layer in this model configuration.

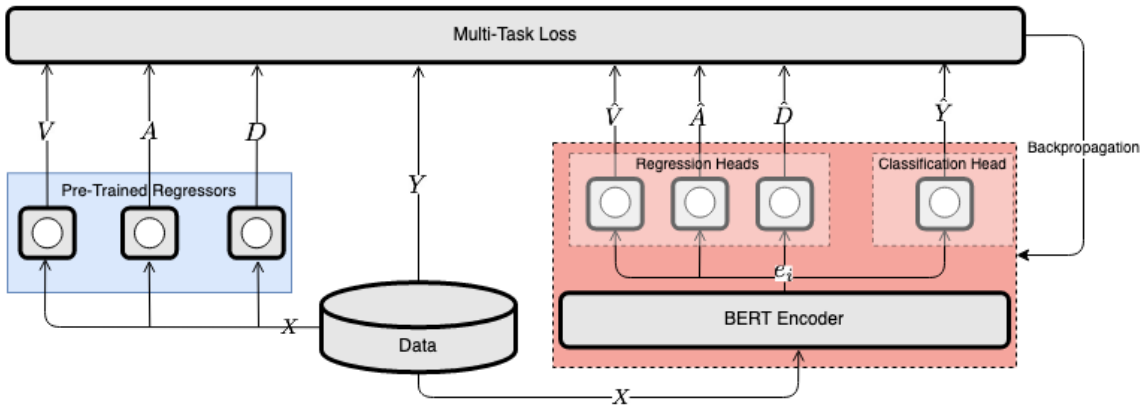

Figure 4.4: **BERT+3dEmoMT**

Model contains a BERT encoder, a linear classification head and multiple linear regressions heads. The aim is to distill knowledge from regressors pre-trained to detect the affective content in tweets.

parameters. The model is optimized via a cross-entropy loss calculated by

$$\mathscr{L}_c = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\Big] \tag{4.2}$$

4.3.2.2 BERT+3dEmoMT

Knowledge distillation is effective method for training models where the outputs of one model are the prediction targets of another model, often described as teacher-student relationship [31, 65, 172]. In **BERT+3dEmoMT**, the pre-trained regressors are given the role of teaching the student model to identify affective signals. This affective signal prediction is treated as an auxiliary task, as the main task of the model is still the binary classification of whether or not a sequence of text is hyperbolic, see Figure 4.4.

The main task of interest is the binary classification of whether a Tweet contains hyperbolic content or not. The goal of the auxiliary regression task is to predict values for the three affective dimensions of valence, arousal and dominance. Given that there are no annotations for valence, arousal and dominance in **HyperTwit** the is aim to distill knowledge from models pre-trained to predict the valence, arousal and dominance in tweets. The predictions for the valence, arousal and dominance produced by the regressors are used as soft targets for this model, see the pre-trained regressors (V, A, D) in Figure 4.4.

Formally, the logits for an individual Tweet are calculated according to

$$\hat{y} = \sigma(e_i \mathbf{W}^Y + b^Y) \tag{4.3}$$

where e_i is the dense representation of tweet i computed by BERT [43], $\mathbf{W}^Y$ and b^Y are learnable parameters. The three regression heads are also linear and the outputs of each head are calculated in a similar way (i.e., $\hat{V} = e_i \mathbf{W}^V + b^V$, $\hat{A} = e_i \mathbf{W}^A + b^A$, $\hat{D} = e_i \mathbf{W}^D + b^D$). The model is optimized via a multi-task loss calculated by

$$\mathscr{L} = \mathscr{L}_c + \lambda \mathscr{L}_r \tag{4.4}$$

$$\mathscr{L}_r = \frac{1}{N} \sum_{i=0}^{N} \left[(V_i \circ A_i \circ D_i - \hat{V}_i \circ \hat{A}_i \circ \hat{D}_i)^2 \right] \tag{4.5}$$

where $\mathscr{L}_c$ is the cross entropy loss, see eq 4.2, and $\mathscr{L}_r$ is the mean-squared error between the values predicted by the pre-trained regressors (i.e., V, A, D) and the values predicted by the model (i.e., $\hat{V}$, $\hat{A}$, $\hat{D}$), λ is a parameter to weight the importance of the auxiliary regression task.

4.3.2.3 BERT+3dEmoAS

BERT+3dEmoAS, is based on an early feature fusion approach that augments the original text sequence with special tokens that represent information [215, 248]. In this

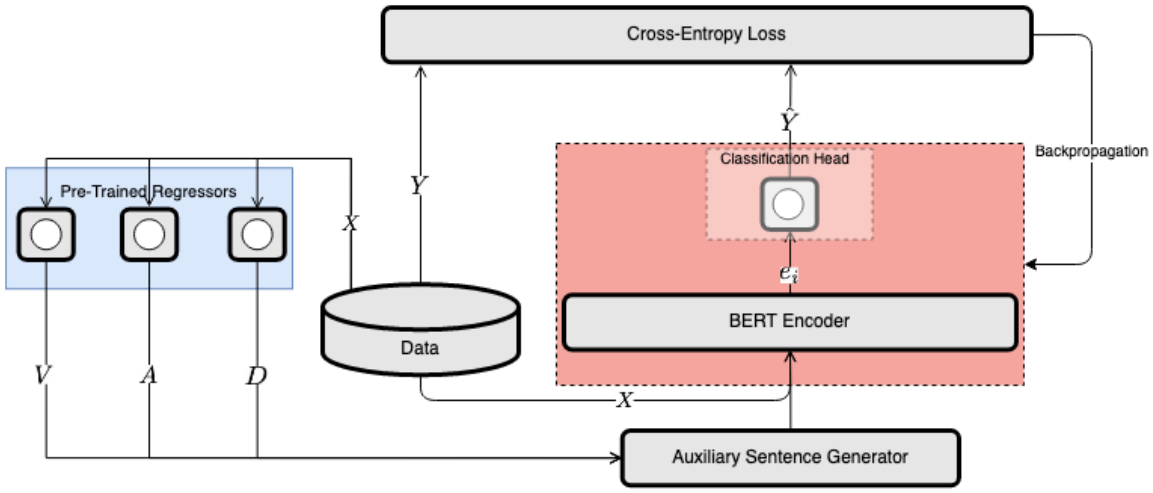

Figure 4.5: **BERT+3dEmoAS**

Model contains a BERT encoder, a linear classification head and multiple linear regressions heads. The affective signals generated from pre-trained regressors are incorporated into model via auxiliary sentence.

model, an auxiliary sentence is generated using special tokens to represent the affective signals as predicted by the pre-trained regressors, (see Figure 4.5). Given that the values for the affective dimensions are continuous values[5], values are placed into 5 equal width bins and 15 special tokens are assigned for each bin for each affective dimension. These special tokens are randomly initialized and updated during model training. The auxiliary sentence is a pseudo-sentence constructed by simply concatenating the special tokens that correspond with the valence, arousal and dominance signal as computed by the pre-trained regressors (i.e., '[LOW0] [HIGH1] [NORMAL2]'). The auxiliary sentence is appended to the original input sentence and separated via the special [SEP] BERT token.

Formally, the logits for an individual Tweet are calculated according to

$$\hat{y} = \sigma(e_i \mathbf{W}^Y + b^Y) \tag{4.6}$$

where e_i is the dense representation of tweet i computed by BERT [43], $\mathbf{W}^Y$ and b^Y are learnable parameters. The model is optimized via a cross entropy loss similar to **BERT+3dEmo**. It is important to note there that the representations of the the special tokens that make up the auxiliary tokens are being updated based on the cross-entropy loss.

[5] [0,1]

4.3.3 Privileged Information for Hyperbole Detection

Learning Under Privileged Information(LUPI) is a paradigm in machine learning that follows a teacher-student model where a teacher model provides information during training time to assist the student model [110, 162]. An important aspect of learning under this paradigm is the concept of privilege with respect to the information used during training. This information is considered privileged because it is only available at training time and not available at time of inference.

The source and type of privileged information (PI) is application dependent. In order to improve the detection of food in images, researchers provide a text list of ingredients in the food in the image as PI alongside various image features [136]. To improve the automated aesthetic ratings of image quality, researchers provide human ratings, (e.g., depth of field), as PI [202].

The proposal to use literal paraphrases of hyperbole as a source of PI is a novel approach hyperbole detection and one of the methodologies introduced in this chapter. It is hypothesised that this extra information will explicitly teach a model where the excessive contrast is within a hyperbole (e.g., '*his room is **the size of a shopping mall***' → '*his room is **very big***') as opposed to exploiting unrealted linguistic patterns.

The motivation for incorporating PI in the form of literal paraphrase of hyperbole is based on the prior observations for research on hyperbole from both a cognitive and computational linguistics point of view. Hyperbole, as defined, consists of an excessive exaggeration along a semantic scale. In the process of identifying hyperbole the identification of the semantic scale is an important step as is evaluating the plausibility of the contrast along that scale [24]. The literal paraphrases of hyperboles as provided in **HYPO**, **HyperTwit** and **HyperProbe** contain information on the semantic scale and an attenuation of the contrast to within a plausible range. Take for example the hyperbole and paraphrase pair ('*my bedroom is the size of a postage stamp*', 'my bedroom is too small'), see Figure 4.6. This literal paraphrase attenuates the implausible claim of a bedroom being the size of a postage stamp to the plausible statement that the bedroom is too small. Another along the same semantic scale is given where the hyperbole (*That bedroom is the size of a whole county*) is attenuated with a literal paraphrase ('that bedroom is so big') that preserves the semantic scale. Another set of examples is also provided in that figure, however along a different semantic scale. Additionally, prior experiments on hyperbole detection revealed that models trained on hyperbole and the literal meaning pairs (i.e., this meal tastes **like cancer**, this meal tastes *bad*) performed better at hyperbole detection compared to models trained differently, see [222] for full

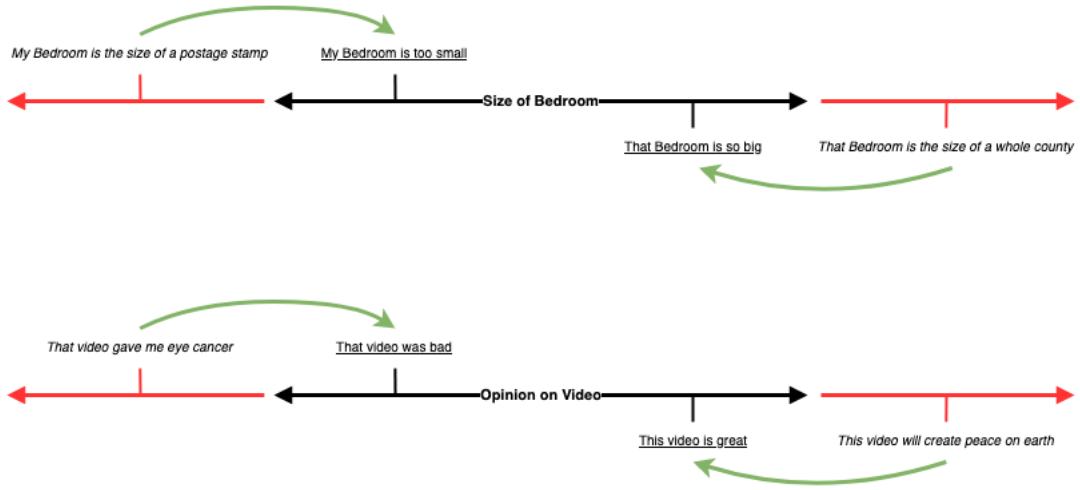

Figure 4.6: **Diagram of Hyperbole and Literal Paraphrase Examples**

This diagram shows hyperbolic expressions and literal paraphrases in terms of a contrast along a semantic scale. The visualisation of the semantic scale in this diagram is analogous to the visualisation of the number line. Hyperbolic expressions are to the extreme left or right of the semantic scale (red line) whilst plausible and non-hyperbolic contrasts are towards the center of the semantic scale (**black line**). The process of literally paraphrasing a hyperbole attenuates the contrast back to within the plausible range of the semantic scale.

details.

The key motivation for the treatment of literal paraphrases as privileged information is to explicitly teach a model when a word or phrase is being used in an excessive manner. Additionally, the literal paraphrase provides semantic information about the intention of the utterance. It is hypothesised that these literal paraphrases can be used to ground hyperbolic expressions to the intended literal meaning, as this intended meaning is often arrived at via common sense reasoning and world knowledge that is not completely encoded in the words contained in the utterance. Many recent research efforts in the NLP community have identified that commonsense reasoning and world knowledge is not well encoded into existing NLP models based on distributional semantics [17, 55, 181, 217].

The incorporation of the literal paraphrases is considered at both the dataset and model level. This is achieved at the dataset level by simply appending the literal paraphrases to the datasets and considering them to be non-hyperbolic samples. Two models are proposed, **BERT+PI**$_R$ and **BERT+PI**$_S$, to incorporate these literal paraphrases via triplet-loss.

$$\mathscr{L} = \frac{1}{n}\sum_{i=1}^{n}\left[||f(x_i^a) - f(x_i^p)||_2^2 - ||f(x_i^a) - f(x_i^n)||_2^2 + m\right] \tag{4.7}$$

Triplet loss has been used in various computer vision algorithms [47, 196]. The idea behind the triplet loss is to force an encoder $f(x)$ that maps x to a feature space R^d, to ensure small distances between all objects of the same class, whilst ensuring distances between pairs of objects from different classes is large [196], see eq. 4.7. Where x_i^a, x_i^p and x_i^n represent an *anchor*, *positive* and *negative* sample respectively, m is the margin enforced between positive and negative pairs and n is the number of objects. A canonical example of triplet loss usage is in facial recognition problems, where an anchor and a positive would be images of the same face but under different conditions (e.g., viewing angle, lighting, etc.) and the negative image would be of a different face entirely [72]. Following on from this canonical example the idea here is to use a triplet loss to differentiate between hyperbolic and literal language. By specifying a hyperbole as an anchor, a different hyperbole as positive and a manually composed literal paraphrase (i.e., privileged information) as a negative, then idea of hyperbolicity can be explicitly enforced on the representation space via the triplet loss.

4.3.3.1 BERT+PI$_R$

BERT+PI$_R$ is a multi-task classification model similar to **BERT+3dEmo$_{MT}$**. A sampling module is used to populate triplets for each tweet in the dataset. A pre-trained BERT model is used to encode dense representations for each tweet with these representations being passed to a linear classification layer. The representations of all three tweets (i.e., anchor, positive and negative) as computed by BERT are used in the computation of the triplet loss.

The methodology used for sampling is an important aspect of approaches that use contrastive losses, such as triplet loss [243, 247]. For **BERT+PI$_R$** the triplet sampling algorithm involves sampling based on label and the knowledge of hyperbole and literal paraphrase pair relations, see Algorithm 1 and see Table 4.1 for examples. This algorithm traverses through all text objects in the data setting each tweet as the anchor tweet. It is important to note here that a text object contains text, label, and either a hyperbole or interpretation of the text. The decision to assign positive and negative examples for each anchor depends on the label of the anchor. If the anchor is a hyperbole then a randomly sampled hyperbole is selected as a positive example. The negative example is then set to be the literal paraphrase of the positive, it is important to note that this literal

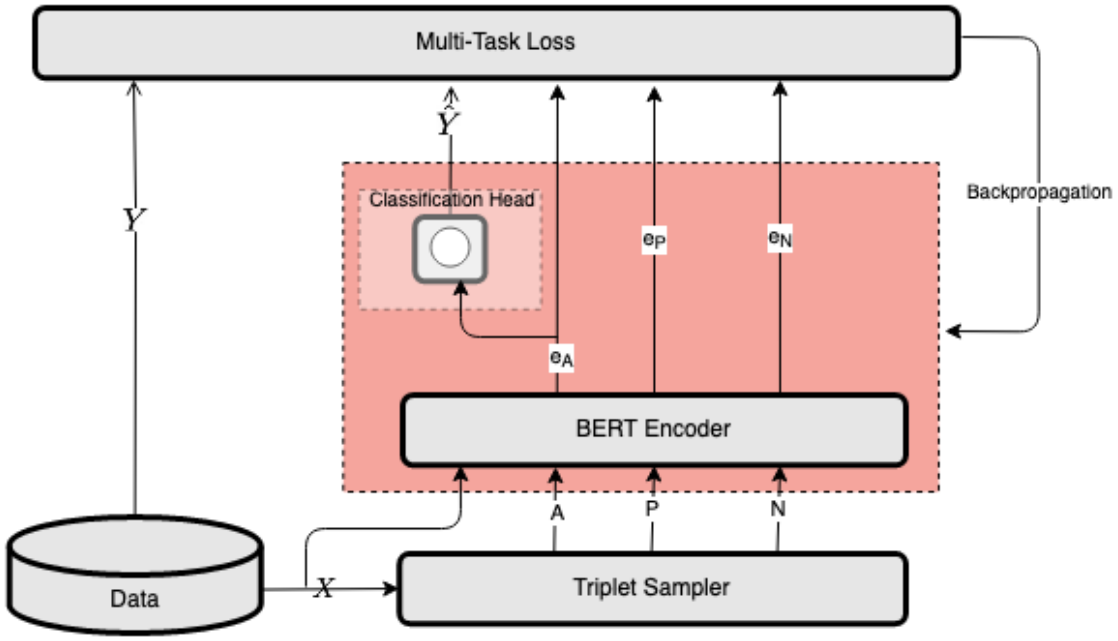

Figure 4.7: **BERT+PI**

Model contains a BERT encoder, a triplet sampler and a linear classification head.

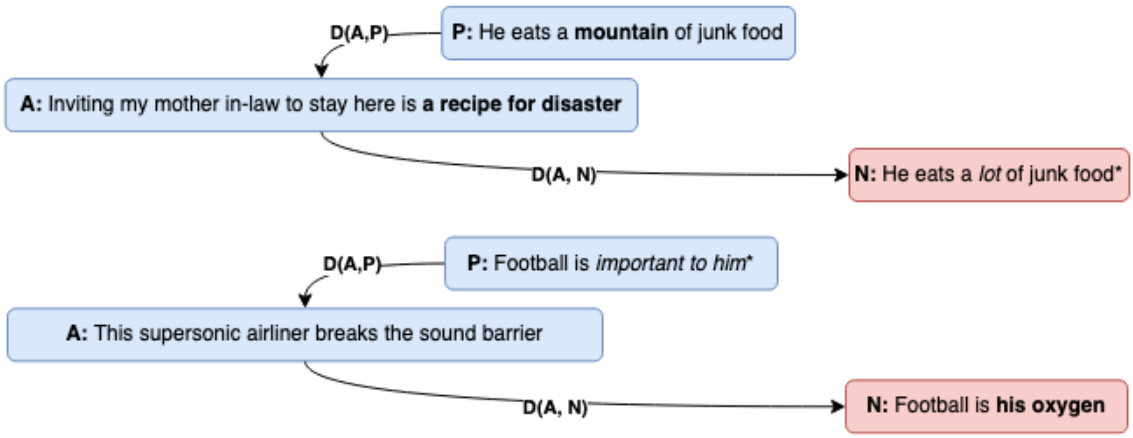

Figure 4.8: **Triplet Sampling Example**

The key idea is ensuring that the distance (i.e. D(A, P)) between an anchor (A) and a positive (P) example is less than the distance (i.e. D(A, N)) between an anchor (A) and a negative (N) example. Note: * indicates privileged information example.

paraphrase is an annotation artifact and as such the source of privileged information for this model. This sampling strategy ensures that optimization of the triplet loss forces a hyperbole to be closer to another hyperbole than its literal paraphrase in the representation space, (see Figure 4.8). If the anchor is not a hyperbole then a randomly sampled literal paraphrase is set as the positive example, again it is important to note here that these literal paraphrases are annotation artifacts thus privileged information. The negative examples is then set to be the hyperbole of the positive. The motivation here is to enforce the model to consider a non-hyperbolic text and a literal paraphrase to be closer in the representation space than the non-hyperbolic text and a hyperbole.

Formally, the logits for an individual Tweet are calculated according to

$$\hat{y} = \sigma(e_i^a \mathbf{W} + b) \tag{4.8}$$

where e_i^a is the dense representation of anchor example i computed by BERT, $\mathbf{W}$ and b are learnable parameters. The model is optimized via a multi-task loss calculated by

$$\mathscr{L} = \mathscr{L}_c + \lambda \mathscr{L}_t \tag{4.9}$$

where $\mathscr{L}_c$ is the cross entropy loss, see eq 4.2, and $\mathscr{L}_t$ is a triplet loss that is calculated between the BERT encoded representations of anchor, positive and negative examples, (see eq. 4.10). In this triplet loss D is the cosine distance, (see eq. 4.11), and m is a hyperparamater indicating the margin. λ is a parameter to weight the importance of the auxiliary regressions task.

$$\mathscr{L}_t = \frac{1}{Ns} \sum_{i=1}^{N} \sum_{j=1}^{s} \left[\max(D(e_i^a, e_{ij}^p) - D(e_i^a, e_{ij}^n) + m, 0) \right] \tag{4.10}$$

$$D(X,Y) = 1 - \frac{X \cdot Y}{||X|| \times ||Y||} \tag{4.11}$$

4.3.3.2 BERT+PI$_S$

BERT+PI$_S$ differs to **BERT+PI$_R$** only by way of sampling algorithm, specifically, the inclusion of sampling based on the semantic similarity of examples rather than random sampling, see Algorithm 2. This sampling algorithm contains similar sampling logic to that of **BERT+PI$_R$** with respect to the relationships between anchor, positive and

Algorithm 1 Semi-Random Triplet Sampling

Require: $D = [t_0, t_1, ..., t_n]$
Require: $s \in \mathbb{Z}^+$ ▷ Sampling Factor
  $H \leftarrow t \forall t \in D \mid t.\text{label} == 1$ ▷ t.label contains annotated label for t
  $P \leftarrow t \forall t \in D \mid t.\text{label} == 2$ ▷ N consists of literal paraphrases (i.e., PI)
  $S \leftarrow \emptyset$
  for $i = 0, i < |D|, i++$ **do**
    $a \leftarrow D_i$
    $T \leftarrow \emptyset$
    for $j = 0, j < s, j++$ **do**
      if a.label $== 1$ **then**
        $p \leftarrow$ sample(H) ▷ sample(X) draws a random sample from X
        if $p == a$ **then**
          $p \leftarrow$ sample(P)
        end if
        $n \leftarrow p$.par ▷ t.par is a literal paraphrase of t
      else if a.label $== 0$ **then**
        $p \leftarrow$ sample(P)
        if $n == a$ **then**
          $p \leftarrow$ sample(N)
        end if
        $n \leftarrow p$.hyp ▷ t.hyp is a hyperbolic expression of t
      end if
      T.insert($[a, p, n]$)
    end for
    S.insert(T)
  end for
  return S

Triplet Label	**Text**	**Label**
Anchor	This video gave me eye cancer	H
Positive	I have a **mountain** of work to do.	H
Negative	I have a **lot** of work to do.*	NH
Anchor	His excuse was not good enough.	NH
Positive	My bedroom is **too small**.*	NH
Negative	My bedroom is **the size of a postage stamp**	H

Table 4.1: **Semi-Random Triplet Sample Examples**

Triplet Label indicates the label of tweet within the sampled triplet. ***Text*** contains text (*indicates privileged information). ***Label*** indicates the label with respect to the overall hyperbole detection task (H = Hyperbole, NH= Non-hyperbole).

Algorithm 2 Similarity Triplet Sampling

Require: $D = [x_0, x_1, ..., x_n]$
Require: Sentence Encoder $f(x)$
Require: $s \in \mathbb{Z}^+$
  $E \leftarrow f(x) \forall t \in D$ ▷ Encode all examples
  $X_{n \times n} \leftarrow$ pairwiseCosineSimilarity(E) ▷ Similarity matrix
  $S \leftarrow \emptyset$
  for $i = 0$, $i < |D|$, $i++$ **do**
    $a \leftarrow D[i]$
    $T \leftarrow \emptyset$
    $ids \leftarrow$ argsort($X[i][:]$) ▷ Sort ids of examples by most similar to anchor
    $j \leftarrow$ len(ids) -1
    if t.label $== 1$ **then**
      while $j > 0$ and len(T) $< s$ **do**
        $c \leftarrow D[ids[j]]$ ▷ Get most similar example as a candidate sample
        if c.label $== 1$ **then** ▷ Candidate must share same label with anchor
          $p \leftarrow c$
          $n \leftarrow p$.interp
          T.insert($[a, p, n]$)
        end if
        $j \leftarrow j - 1$
      end while
    else if t.label $== 0$ **then**
      while $j > 0$ and len(T) $< s$ **do**
        $c \leftarrow D[ids[j]]$
        if c.label $== 2$ **then**
          $p \leftarrow c$
          $n \leftarrow p$.hyp
          T.insert($[a, p, n]$)
        end if
        $j \leftarrow j - 1$
      end while
    end if
    S.insert(T)
  end for
  return S

Triplet Label	**Text**	**Label**
Anchor	When the boy broke his toy he cried a sea of tears.	H
Positive	The small child was **drowning in her tears**.	H
Negative	The small child was *crying a lot*.*	NH
Anchor	I was very sad for having to leave my home.	NH
Positive	His sister will *be very angry* when she hears that.*	NH
Negative	His sister will **hit the roof** when she hears that	H

Table 4.2: **Similarity Triplet Sample Examples**

Hyperparameter	**Values**	**Optimal**
Dropout	0.1, 0.2, 0.3	0.1
Learning Rate	1e-04, 1e-05, 1e-06	1e-04
λ	0.25, 0.5, 1	0.25
Freeze embeddings	True, False	False
Frozen Transformer layers	0, 9, 10, 11, 12	10
Encoder	BERT, RoBERTa, BERTweet	BERT

Table 4.3: **Hyperparameter search**

Hyperparameter indicates the hyperparameter. **Values** indicates the different values used in the hyperparameter search. **Optimal** indicates the optimal parameter choice on average across the different models. Note: Not all parameters are necessary for all models (e.g., λ is only required for multi-task models (i.e., **BERT+3dEmo**$_{MT}$, **BERT+PI**$_R$, **BERT+PI**$_S$.

negative tweets and their respective classes. However, the positive example of a hyperbole anchor example is now chosen based on the similarity between these examples in BERT representation space using cosine distance. Likewise, a positive example of a non-hyperbolic anchor is also chosen based on the similarity of the examples in BERT representation space. From Tables 4.1 and 4.2 the different output of these two sampling algorithms is shown, most notably the relationship between anchor and positive examples.

4.4 Experiments and Results

Several experiments are designed to test the ability of models to detect hyperbole in multiple settings. Firstly, detecting hyperbole within the same domain to assess the

ability of a model to adapt to different hyperbolic expressions from a similar domain, (see Section 4.4.1). Detecting hyperbole from a different domain to which the model was trained on is also explored (see Section 4.4.2), the observations from Chapter 3 showed that hyperbolic expression varied between domains. These experiments are designed to provide insight on how much of an impact this has on hyperbole detection models. Lastly the ability of models to detect minimal synthetic examples of common forms of hyperbolic expressions on **HyperProbe** are examined (see Section 4.4.3).

4.4.1 In-Domain Hyperbole Detection

The first set of experiments seeks to answer the following question '*How accurately can a model detect hyperbole? Given that it was trained on hyperbole from the same domain?*' In particular, the focus is on two domains, online user-generated content and a generic domain. The **HyperTwit** dataset, (see Chapter 3), is used as a source of hyperbole in the domain of online user-generated content. The **HYPO** dataset, see Chapter 3, is used as a source of hyperbole in a generic domain (i.e., idiomatic hyperbole).

The datasets are split into train:dev:test partitions in a 80:10:10 ratio and a hyper-parameter search is conducted on the held out development set for numerous hyper-parameters including dropout, learning rate, λ, whether to freeze or train the encoder embeddings, which transformer layers to freeze in the encoder and encoder type, see Table 4.3. Validation accuracy is the criterion for choosing hyper-parameters across 3 runs per hyper-parameter configuration on the held-out development partitions containing approximately 600 and 140 examples for **HyperTwit** and **HYPO** respectively. Various configurations of freezing transformer layers and embeddings was also conducted during hyperparameter search due to the performance impacts of these decisions on downstream tasks [104, 113, 137]. The method for encoding examples into word representations is experimented with in the hyperparameter search using RoBERTa[6] [124]. BERTweet[7] [149] and BERT. Once optimal hyperparameters have been chosen, all models are trained for 6 epochs on the training set and the loss is monitored on the development set via an early stopping criteria to identify optimal model checkpoints. Three runs are conducted for each model and training dataset configuration to account for model variance.

Results of in-domain experiments for hyperbole detection show that models that incorporate privileged information outperform, with respect to $F1$ score, baseline models in both domains, see Tables 4.4 and 4.5. This is most clear on the experiments for

[6]https://huggingface.co/roberta-base

[7]https://huggingface.co/cardiffnlp/twitter-roberta-base

Model	**F1**	**Precision**	**Recall**
LR+QQ	0.710(-)	0.679(-)	0.745(-)
NB+QQ	0.693(-)	0.689(-)	0.696(-)
BERT	0.709(0.064)	0.711(0.077)	0.735(0.177)
BERT+QQ	0.671(0.086)	0.650(0.147)	0.765(0.246)
BERT+PI$_S$	0.768(0.009)	0.739(0.051)	0.804(0.051)
BERT+PI$_R$	**0.781(0.012)**	0.754(0.053)	0.814(0.039)
BERT+3dEmo	0.730(0.041)	**0.785(0.033)**	0.690(0.095)
BERT+3dEmo$_{MT}$	0.733(0.033)	0.697(0.103)	0.797(0.124)
BERT+3dEmo$_{AS}$	0.626(0.021)	0.485(0.004)	**0.886(0.099)**

Table 4.4: **Results from In-Domain Experiments - HYPO**

Model	**F1**	**Precision**	**Recall**
LR+QQ	0.583(-)	0.638(-)	0.537(-)
NB+QQ	0.579(-)	0.490(-)	0.706(-)
BERT	0.733(0.012)	0.718(0.055)	**0.755(0.057)**
BERT+QQ	0.732(0.014)	0.730(0.049)	0.736(0.034)
BERT+PI$_S$	**0.746(0.017)**	**0.769(0.010)**	0.725(0.037)
BERT+PI$_R$	0.745(0.018)	0.754(0.021)	0.736(0.046)
BERT+3dEmo	0.725(0.008)	0.758(0.120)	0.714(0.088)
BERT+3dEmo$_{MT}$	0.434(0.381)	0.565(0.492)	0.362(0.329)
BERT+3dEmo$_{AS}$	0.0 (0.0)	0.0 (0.0)	0.0(0.0)

Table 4.5: **Results from In-Domain Experiments - HyperTwit$_K$**

the domain of idiomatic hyperbole where a +0.71 (10%) increase in F1 is observed for **BERT+PI$_R$** over the best performing baseline (**LR+QQ**). A much smaller, +0.014, increase in F1 is observed for **BERT+PI$_S$** over the best performing baseline (**BERT**) for hyperbole in the domain of Twitter.

The highlighted text outputs generated by LIME[190] are provided, see Figures 4.9 and 4.10. There are two dimensions of information translated by this highlighting method from LIME. The colour indicates the class, blue highlights indicating the negative class (i.e., non-hyperbolic) and orange indicating the positive class (i.e., hyperbole). The intensity of the colours represents the extent to which that word contributes to the prediction made by a model. A dull highlight indicating a small contribution to predicting that particular class whilst a strong highlighting indicates a strong contribution to predicting that particular class. Figure 4.9 contains examples that indicate that the

BERT		BERT+PI$_R$	
LIME Word Weightings	**P(*h*)**	**LIME Word Weightings**	**P(*h*)**
Search engines are brainless entities.	.66	Search engines are brainless entities.	.18
Me, the wife of that boorish, brainless man.	.78	Me, the wife of that boorish, brainless man.	.74
Caterpillars grow wings and become butterflies.	.55	Caterpillars grow wings and become butterflies.	.01
The shelves contained crystal clear glasses.	.67	The shelves contained crystal clear glasses.	.04
There was no one there but the family.	.71	There was no one there but the family.	.20

Figure 4.9: **Model Explanation Comparisons - HYPO**

LIME Word Weightings indicate the importance of a word for classification, **orange** highlights indicate hyperbolic words, **blue** highlights indicate non-hyperbolic words. **P(*h*)** is the prediction probability that a sentence was hyperbolic, **red** indicates the wrong class (assuming a .5 decision threshold).

BERT		BERT+PI$_R$	
LIME Word Weightings	**P(*h*)**	**LIME Word Weightings**	**P(*h*)**
This policy will plunge the country into a chaos.	.20	This policy will plunge the country into a chaos.	.79
Imagination is the queen of truth.	.42	Imagination is the queen of truth.	.93
Every flavor is dynamite.	.35	Every flavor is dynamite.	.96
The lesson was taking forever.	.49	The lesson was taking forever.	.65
Moms are more powerful than the sea.	.38	Moms are more powerful than the sea.	.81

Figure 4.10: **Model Explanation Comparisons - HYPO**

increase in precision for **BERT+PI**$_R$ seen in Table 4.4 is a result of a better contextual understanding of hyperbole-prone ECF terms. The first two examples in particular highlight the understanding of the word *brainless* in both a hyperbolic and non-hyperbolic context that are correctly classified by **BERT+PI**$_R$ but incorrectly classified by baseline **BERT**.

It is interesting to note here that the addition of the QQ features does not result in a significant improvement in $F1$ score for either domain. With respect to prior research, the work that introduced the QQ features found that they did improve hyperbole detection accuracy when added as features to their detection pipeline [222] whilst, follow up research found that the results of incorporating QQ features were mixed [101]. The

Type	**F1**	**Precision**	**Recall**
ECF	0.720(0.226)	0.791(0.253)	0.681(0.237)
Quant	0.522(0.399)	0.550(0.433)	0.541(0.427)
Qual	0.484(0.372)	0.544(0.422)	0.485(0.404)

Table 4.6: **Keyword Results by Hyperbole Type on HyperTwit$_k$**

Type indicates the type of hyperbole. **F1**, **Precision** and **Recall** indicates the mean and standard deviation for all keywords of that type.

results from the in-domain hyperbole detection experiments show that the QQ features have a either a negligible or detrimental impact on the detection of hyperbole. A decrease in average $F1$ of 0.38 on **HYPO** and an average decrease of 0.001 in $F1$ on **HyperTwit**$_K$.

In Table 4.6 the mean F1, precision and recall scores for keywords grouped by hyperbole type on **HyperTwit**$_k$ are provided, see Table **??** for details on hyperbole types. High standard deviations in F1, precision and recall scores across the keywords for all types can be observed. It is hypothesized that this result relates to observations on hyperbole diversity presented in Chapter 3. Specifically, the observation that many hyperboles using particular keywords (i.e., '*everybody*', '*garbage*', '*toxic*' etc.) are simply parroted by different users on Twitter and would be easy to detect resulting in high F1 for those particular keywords. It was also observed that hyperboles using particular keywords (i.e., '*blind*', '*heaven*', etc.) contained many novel hyperbolic expressions that would be harder to detect resulting in lower F1 for those particular keywords.

The highest mean F1 for keywords are associated with ECF hyperboles (i.e., '*always*', '*never*', '*absolute*', '*everybody*' etc.) suggesting that ECFs are a relatively simple type of hyperbole to identify, see Table 4.6. However, there are also many tweets in both subsets that contain ECFs that seem reasonable as a factual statement labelled as hyperbole (e.g., 'The *only* regret I have about offline tekken is *never* making a regional top 8', 'That was the *quickest* I've ever seen a 2-0 lead blown').

Results from Table 4.6 indicate that quantitative hyperboles (i.e., '*hour*', '*zero*', '*stacks*', '*piles*' etc.) are harder to identify than ECF hyperboles. Several errors are identified that indicate an inability to identify hyperbole expressed via excessive quantitative concepts (i.e., duration) in Arabic numerals ('Either responds in *.00000000001 seconds* or in *84 years* with *zero* in-between', '*day 10393920829*: i still don't understand why jughead faked his own death'). There are also examples where likely factual or reasonable statements about quantitative concepts (i.e, duration) are incorrectly labeled as hyperbolic ('Lakers are *48 minutes* away from their 1st ring in a *decade*', 'Omg ive been on Twitter for *ten years* wow. i think i deserve a blue tick').

Qualitative hyperboles (i.e., '*corrupt*', '*amazing*', '*dead*', '*pain*', etc.) are also a difficult type of hyperbole to classify in **HyperTwit**, see Table 4.6. A hypothesis here is that qualitative hyperboles are hardest to identify, (i.e., low recall), because there is a difference in the way they are often expressed compared to the quantitative and ECF hyperbole. With quantitative and ECF hyperbole the semantic concept, along which the excessive magnitude is exaggerated, is scalar and obvious (i.e., time period, measure, quantity, universality, nullity, veracity etc.). Conversely, in qualitative hyperbole the semantic concept is often not scalar and unclear ('To make a tasty tequila sunrise. just whisk a teaspoon of *bitter* stout with green paint.', 'Only fans is a *cancer* to the west...Already in stage 2.'). These examples require more complex reasoning and world knowledge to identify if the utterance contains an excessive contrast with reality.

Results indicate that there is limited scope for further improvement to the modelling of affect signals and their incorporation into hyperbole detection models. Despite the improvements of **BERT+3dEmoMT**, there are hyperbolic tweets that are expressed without strong affective language, which **BERT+3dEmoMT** fails to detect ('mexico is *calling my name rn*', 'my parents want me to watch the superbowl *what dimension did i just land in*'). This observation, and others made in this discussion, suggest that the limited improvements are not a technical issue (i.e., ineffective incorporation of signals), but a reflection of the relationship between affect and hyperbole. It is the lack of reasoning and world knowledge that results in poor detection of hyperbole, not the inability of the models to utilise the affective signals.

With respect to the incorporation of privileged information in **BERT+PI** models. Despite achieving the best F1 on both datasets, the improvement is diminished for hyperbole in the Twitter domain compared to the idiomatic domain. A hypothesis for this is that the **HYPO** datasets is better suited to this given the trio of corpora (hyperbole, paraphrase and minimal units) aligns well with the triplet sampling and triplet loss used in the model. Particularly, the sentences in the minimal units corpus contain the hyperbolic units in a non-hyperbolic context ('The missions successfully went to the **moon and back**', 'Love you to the **moon and back**') providing hyperbolic words and phrases in a non-hyperbolic context. There is no parallel to the minimal units corpus in **HyperTwit**, however the keyword sampling strategy was designed to address this in the sense that many tweets were collected that mention hyperbole prone words in both a hyperbolic and non-hyperbolic context. However, this did not address cases when the hyperbolic tokens were not just the keywords used for sampling.

4.4.2 Cross-Domain Hyperbole Detection

A second set of experiments were designed to probe '*how well can a model detect hyperbole? Given that it was trained on hyperbole from a different domain?*'. These experiments also provide an insight into the differences in hyperbolic expression between the two domains. The hypothesis being tested here is that a hyperbole detection model trained to recognise hyperbole on the Twitter domain (i.e **HyperTwit**$_K$) would be able to achieve similar accuracy when evaluated on the idiomatic domain (i.e. HYPO) and vice versa. If this holds, then the assumption is that the expression of hyperbole between the two domains is similar. For these experiments model checkpoints from in-domain experiments are used (see Section 4.4.1). There are two settings for cross-domain experiments, using models trained on the Twitter domain to detect hyperbole in the domain of idiomatic hyperbole (i.e., **HYPO** test set), and using models trained on idiomatic domain to detect hyperbole in the Twitter domain (i.e., **HyperTwit**$_K$ test set).

Results from the cross-domain experiments are shown in Tables 4.7 and 4.8. The first observation is that models trained on Twitter domain and evaluated on idiomatic domain perform worse than models trained and evaluated on the domain of idiomatic hyperbole. Maximum scores of 0.619,0.673,0.634, for F1, precision and recall when cross-trained (see, Table 4.7), compared to max scores of 0.781, 0.785, 0.886 when trained in-domain (see, Table **??**). This is despite the **HyperTwit**$_K$ dataset being more than five times the size of the **HYPO** dataset.

The same observation can be made when the models are trained on the idiomatic domain achieve poor results when evaluated on the Twitter domain. Maximum scores of 0.550, 0.403, 0.944 for F1, precision and recall when cross-trained (see, Table 4.8), compared to max scores of 0.746, 0.769, 0.755 for in-domain experiments (see, Table **??**).

It is important to note here that these results cannot be attributed entirely to a difference in expression of hyperbole between the two domains. The problem of diminishing accuracy of NLP models when trained and evaluated across different domains is a well known problem, the study of this problem is referred to Domain Adaptation, (see [32, 118, 182] for further details). However, this result is enough to state that hyperbolic expression varies between the two domains, drawing conclusions about how these differences are manifested in the actual hyperbolic expression can not be drawn from this result. Additionally, The gains in metrics for **BERT+PI** models observed from the results on in-domain experiments were not observed in these experiments.

Model	**F1**	**Precision**	**Recall**
LR+QQ	0.519(-)	0.578(-)	0.471(-)
NB+QQ	0.520(-)	0.613(-)	0.451(-)
BERT	0.616(0.030)	0.604(0.070)	**0.634(0.039)**
BERT+QQ	**0.619(0.015)**	0.632(0.037)	0.608 (0.010)
BERT+PI$_S$	0.611(0.019)	**0.673(0.018)**	0.562(0.044)
BERT+PI$_R$	0.576(0.047)	0.643(0.070)	0.523 (0.035)
BERT+3dEmo	0.616(0.072)	0.66(0.066)	0.608 (0.182)
BERT+3dEmo$_{MT}$	0.303(0.280)	0.465(0.404)	0.235 (0.236)
BERT+3dEmo$_{AS}$	0.0 (-)	0.0 (-)	0.0 (-)

Table 4.7: **Cross-Domain Results - Idiomatic Hyperbole Evaluation**

Trained on Twitter hyperbole (i.e. **HyperTwit**$_K$), evaluated on idiomatic hyperbole (i.e. **HYPO**).

Model	**F1**	**Precision**	**Recall**
LR+QQ	0.551(-)	0.402(-)	0.876(-)
NB+QQ	0.529(-)	0.380(-)	0.870(-)
BERT	0.525(0.028)	0.377(0.009)	0.880(0.135)
BERT+QQ	0.496(0.073)	0.358(0.022)	0.838(0.261)
BERT+PI$_S$	**0.550(0.014)**	0.388(0.016)	**0.944(0.029)**
BERT+PI$_R$	0.539(0.009)	0.377(0.006)	0.943(0.043)
BERT+3dEmo	0.545(0.011)	**0.403(0.018)**	0.847(0.094)
BERT+3dEmo$_{MT}$	0.538(0.005)	0.380(0.015)	0.929(0.064)
BERT+3dEmo$_{AS}$	0.513(0.021)	0.357(0.006)	0.914(0.094)

Table 4.8: **Cross-Domain Results - Twitter Hyperbole Evaluation**

Trained on idiomatic hyperbole (i.e. **HYPO**), evaluated on Twitter hyperbole (i.e. **HyperTwit**$_K$).

4.4.3 HyperProbe Experiments

Experiments to probe the ability of hyperbole detection models to accurately detect minimal hyperbolic expressions are undertaken on the **HyperProbe** dataset, (see Chapter 3). It is important to note that no models are trained on any of the examples in **HyperProbe**, rather, they are treated as test sets only. The models used for evaluation are the same model checkpoints used for both in-domain and cross-domain experiments, that is, models trained on the **HYPO** train set. Results are presented in 5 sections, covering the ECF, qualitative hyperbole, quantitative dimensions, time periods and intrinsic quantity tests.

4.4.3.1 Results - ECF

The mean and standard deviation of all runs for various metrics on the ECF tests from **HyperProbe** is presented in Table 4.9. From this table it can be observed that models that incorporate affective signals and models that incorporate privileged information provide considerable improvements in detecting ECF hyperbole compared to the baseline models. These models are also considerably more stable with standard deviations in $F1$ for baseline models (BERT based), (0.337, 0.340), much higher than the models introduced by the author, (0.014,0.011,0.081,0.022,0.058).

From the explanations provided by LIME, see Figure 4.11, it can be observed that the inclusion of privileged information into **BERT+PI**$_S$ has resulted in a better contextual understanding of ECF keywords. Also, **BERT+PI**$_S$ understands that ECF keywords (i.e., *absolute*, *never*, *nobody*, *everybody*) are being used in non-hyperbolic sentences and correctly classifies these sentences non-hyperbolic. Conversely, the **LR+QQ** baseline does not appear to understand that ECF keywords are being used in non-hyperbolic sentences and incorrectly classifies the sentences as hyperbolic. The decision to classify these sentences as hyperbolic is strongly driven by the ECF keywords alone according to the LIME explanations, see Figure 4.11.

4.4.3.2 Results - Qualitative Hyperbole

The mean and standard deviation of all runs for various metrics on the test sentences designed to probe qualitative hyperbole from **HyperProbe** are shown in Table 4.10. From this table it can be observed that all models struggle to detect qualitative hyperbolic expressions, **BERT+PI**$_R$ achieves the highest $F1$ of only 0.527 with a sub-0.5 precision of 0.486. It is observed that many models display large standard deviations for recall, (i.e.,

Model	F1	Precision	Recall
LR+QQ	0.678(-)	0.747(-)	0.621(-)
NB+QQ	0.523(-)	0.690(-)	0.421(-)
BERT	0.490(0.340)	0.751(0.158)	0.516(0.453)
BERT+QQ	0.540(0.337)	0.721(0.184)	0.632(0.484)
BERT+PI$_S$	0.701(0.014)	0.756(0.033)	0.656(0.047)
BERT+PI$_R$	0.688(0.011)	0.706(0.057)	0.677(0.070)
BERT+3dEmo	0.656(0.081)	**0.814(0.119)**	0.576(0.152)
BERT+3dEmo$_{MT}$	**0.737(0.022)**	0.700(0.104)	0.800(0.107)
BERT+3dEmo$_{AS}$	0.650(0.058)	0.510(0.024)	**0.902(0.144)**

Table 4.9: **HyperProbe Results. Extreme Case Formulations**

LR+QQ		BERT+PI$_S$	
LIME Word Weightings	P(*h*)	LIME Word Weightings	P(*h*)
the absolute majority was significant	.69	the absolute majority was significant	.35
the exam result was absolute	.70	the exam result was absolute	.12
the dead will never return	.53	the dead will never return	.02
nobody in the group looked interested	.51	nobody in the group looked interested	.10
everybody in the audience was shocked	.68	everybody in the audience was shocked	.50

Figure 4.11: **LIME Explanations - HyperProbe (ECFs)**

Model	F1	Precision	Recall
LR+QQ	0.407(-)	0.333(-)	0.522(-)
NB+QQ	0.336(-)	0.400(-)	0.290(-)
BERT	0.278(0.275)	0.240(0.209)	0.401(0.497)
BERT+QQ	0.352(0.307)	0.255(0.227)	0.599(0.529)
BERT+PI$_S$	0.518(0.072)	**0.496(0.054)**	0.551(0.119)
BERT+PI$_R$†	**0.527(0.030)**	0.486(0.054)	0.590(0.089)
BERT+3dEmo	0.445(0.172)	0.480(0.062)	0.454(0.244)
BERT+3dEmo$_{MT}$	0.509(0.066)	0.416(0.081)	0.691(0.146)
BERT+3dEmo$_{AS}$	0.481(0.028)	0.325(0.012)	**0.932(0.117)**

Table 4.10: **HyperProbe Results (Qualitative Hyperbole)**

BERT+PI$_R$	
LIME Word Weightings	**P(h)**
the old car is a headache	**.83**
the main symptom is mild headache	**.08**
the mans breath is toxic	**.83**
the mushroom is not toxic	**.10**
the game is dead bad	**.76**
the internet traffic is bad	**.16**
this show is autistic	**.82**
the client is autistic	**.31**

Figure 4.12: **LIME Explanations - HyperProbe (Qualitative Hyperbole)**

BERT+PI$_R$	
LIME Word Weightings	**P(h)**
the man who is in fear	.86
the little guy is getting scared	.78
the black woman is very attractive	.62
the elderly man is completely deaf	.75
the old man is finally dead	.84
the old man was contemplating heaven	.82
the man is experiencing panic	.67

Figure 4.13: **LIME Explanations - HyperProbe (Qualitative Hyperbole)**

BERT+PI$_R$	
LIME Word Weightings	**P(*h*)**
the new album is great	.86
the smile is charming	.84
the playlist is great	.52
the game is bad	.67
the web site is really great	.73
the first one was bad	.67
the interview is really great	.88

Figure 4.14: **LIME Explanations - HyperProbe (Qualitative Hyperbole)**

0.529, 0.497, 0.244), suggesting that some of these runs are degenerating to outputting all positive class or all negative class predictions. However, the variances are at least stable in **BERT+PI**$_R$, allowing for a deeper dive into the performance of this model to gain insight into the detection of qualitative hyperbole.

Despite the sub-0.5 precision, worse than a random classifier, analysis of the predictions of **BERT+PI**$_R$ reveals some patterns in the decisions made by the model. With respect to correct decisions examples are identified where the model appears to understand when a word is used in a hyperbolic vs a non-hyperbolic context, see Figure 4.12. The model, whilst scoring a precision worse than random, appears to display patterns in decision making that do not appear to be random. Observing figures 4.13 and 4.14 evidence of the model displaying non-random decisions can be found. Particularly in these figures it can be seen that the model wrongly predicts sentences that contain common nouns (e.g., man, woman, guy) as being hyperbolic and gives a strong importance to those particular nouns for that classification decision. The hypothesis for these errors is that in the training set of **HYPO** there are several idiomatic hyperbolic expressions that use these common nouns (’*the old* ***man*** *is a dinosaur*’, ‘*It’s time to stop living like a dead* ***man***’, ‘*Manners make the* ***man***’, ‘*she has become an iceberg of a* ***woman***’, ‘*she was an unattainable* ***woman***’).

Another common, albeit general, error was the classification of benign evaluative sentences as hyperbolic, see Figure 4.14. These errors are difficult to explain as they

Model	F1	Precision	Recall
LR+QQ	0.615(-)	0.500(-)	0.800(-)
NB+QQ	0.565(-)	0.500(-)	0.650(-)
BERT	0.576(0.048)	0.463(0.001)	0.775(0.177)
BERT+QQ	0.552(0.183)	0.470(0.073)	0.733(0.379)
BERT+PI$_S$	0.590(0.088)	0.492(0.048)	0.750(0.200)
BERT+PI$_R$†	**0.615(0.005)**	**0.503(0.025)**	**0.800(0.087)**
BERT+3dEmo	0.539(0.096)	0.485(0.015)	0.633(0.208)
BERT+3dEmo$_{MT}$	0.608(0.040)	0.496(0.008)	**0.800(0.150**
BERT+3dEmo$_{AS}$	0.571(0.068)	0.464(0.006)	0.767(0.225)

Table 4.11: **Hyperprobe Results. Quantitative Dimensions**

appear trivial to a human reader but it is hypothesised that perhaps there is just too little context for the model to make a correct decision. Whatever the reason for these particular decisions, it is clear that the qualitative tests serve as a challenging benchmark for hyperbole detection models.

4.4.3.3 Results - Quantitative Hyperbole

The mean and standard deviation of all runs for various metrics on the test sentences designed to probe quantitative hyperbole from **HyperProbe** are provided in Table 4.11. From this table it can be observed that all models display a similar pattern of high recall (0.633 to 0.800) and low precision (0.463 to 0.503). Suggesting that false positives are a problem and the models are aggressive in the sense that they favour hyperbolic predictions over non-hyperbolic predictions.

From analysis of LIME explanations one particular decision pattern can be identified as the source of many false positives. When a determiner, particularly a possessive, appears as the first word in the following sentence template, *{MASK}{MASK} is as {JJ} as {MASK}{MASK}*, the model predicts a hyperbole, seemingly irrespective of the hyperbolic nature of the comparison begin made, see Figure 4.15. From the word highlights in the figure it can be noted that there is a strong influence towards a hyperbolic classification for the first word of a sentence when it is a possessive (i.e., *he*, *she*, *her*, *their*, *my*, etc.) and the words and phrases '*is*', '*as*', '*is as*' and '*as a*'. This contributes to the low precision because of the many literal statements in the test dataset for this particular sentence template. A hypothesis for this error is that the literal paraphrases of hyperbolic expressions that take this form remove many tokens from the original sentence. (i.e., '*He's*

BERT+PI$_R$	
LIME Word Weightings	**P(*h*)**
her brain is as small as a quarter	**.86**
Her hair is as thin as silk	**.84**
that bag is as heavy as a truck	**.71**
my heart is as heavy as the world	**.73**
his penis is as small as a mosquito	**.81**
his mouth is as big as a barn	**.87**
his body is thin as a mountain	**.73**
His beard is as thick as his mustache	.87
that bag is as heavy as a suitcase	.72
Her sister is as tall as her mother	.86
their hair is as long as a finger	.74

Figure 4.15: **LIME Explanations - HyperProbe (Quantative Dimensions)**

as *mad* ***as a*** *hippo with a hernia*' -> '*He's* ***very*** *mad*'). This could potentially contribute to the increased importance of particular words and phrases (i.e., 'is as' and 'as a') being considered hyperbolic because they were removed from the original sentence. It is also noted, that this sentence is a particularly common form of hyperbolic expression in the training data (i.e., '*There lived a man* ***as*** *big* ***as a*** *barge*' '*He has* ***as*** *many debts* ***as a*** *dog has fleas*', '*He's* ***as*** *mad* ***as a*** *hippo with a hernia*'. '*you look* ***as*** *white* ***as a*** *ghost*').

4.4.3.4 Results - Time Periods

The mean and standard deviation of all runs for various metrics on Time Period tests from **HyperProbe** are provided in Table 4.13. These test sentences were designed to probe the understanding of the length of time periods and comparisons between them. It can be observed from this table that most models perform very poorly. It is clear from these results that the models do not understand the plausible ranges of duration for periods of time. A peculiar result here is that **BERT+3dEmo**$_{AS}$ achieves relatively good

Model	F1	Precision	Recall
LR+QQ	0.343(-)	0.400(-)	0.300(-)
NB+QQ	0.194(-)	0.273(-)	0.150(-)
BERT	0.273(0.457)	0.431(0.374)	0.336(0.575)
BERT+QQ	0.443(0.407)	0.454(0.393)	0.475(0.502)
BERT+PI$_S$	0.228(0.283)	0.697(0.153)	0.183(0.258)
BERT+PI$_R$†	0.589(0.127)	0.673(0.038)	0.556(0.230)
BERT+3dEmo	0.327(0.360)	0.455(0.394)	0.301(0.386)
BERT+3dEmo$_{MT}$	0.438(0.296)	**0.697(0.046)**	0.410(0.393)
BERT+3dEmo$_{AS}$	**0.768(0.053)**	0.680(0.017)	**0.896(0.157)**

Table 4.12: **HyperProbe Results. Time Periods**

Model	F1	Precision	Recall
LR+QQ	0(-)	0(-)	0(-)
NB+QQ	0(-)	0(-)	0(-)
BERT	0.354(0.317)	0.309(0.269)	0.425(0.405)
BERT+QQ	0.418(0.358)	0.457(0.050)	0.593(0.524)
BERT+PI$_S$	0.301(0.284)	0.365(0.097)	0.366(0.474)
BERT+PI$_R$†	0.369(0.215)	0.400(0.086)	0.395(0.331)
BERT+3dEmo	0.270(0.311)	0.279(0.254)	0.309(0.418)
BERT+3dEmo$_{MT}$	0.381(0.272)	**0.548(0.149)**	0.462(0.454)
BERT+3dEmo$_{AS}$	**0.623(0.031)**	0.485(0.018)	**0.877(0.114)**

Table 4.13: **HyperProbe Results. Intrinsic Quantities**

F1, precision and recall scores despite degenerating to the all-negative classifier in other results.

4.4.3.5 Results - Intrinsic Quantities

The mean and standard deviation of all runs for various metrics on Intrinsic quantities tests from **HyperProbe** are provided in Table 4.13. These test sentences were designed to probe the understanding of quantitative attributes of objects. It can be observed from this table that most models perform very poorly on this test, similar to the time period test. It is clear from these results that the models have not learnt an understanding of the intrinsic quantitative values of objects.

4.5 Conclusion

The content in this chapter focused on the task of computational hyperbole detection and provided numerous contributions to the research questions and objectives of this thesis. A number of baseline models for hyperbole detection as well as various model configurations proposed by the author were evaluated under various experiment settings throughout this chapter. The results of these evaluations are summarised in two heat map visualisations, See Figures 4.16 and 4.17. Model configurations are represented along the y-axis with naming conventions following those as presented throughout the chapter. However, subscripts can not be represented via plotting library and are altered (i.e. BERT+PI$_R$ → BERT+PI-r). The following naming conventions are used for experiment settings:

- hypo = **HYPO** (Section 4.4.1)
- hyperT = **HyperTwit** (Section 4.4.1)
- ecf = **Extreme Case Formulation Tests** (Section 4.4.3)
- qual = **Qualitative Hyperbole** (Section 4.4.3)
- dims = **Qunatitative Dimensions** (Section 4.4.3)
- time = **Time Periods** (Section 4.4.3)
- quant = **Intrinsic Quantities** (Section4.4.3)
- hypo-hyperT = **HYPO** on **HyperTwit** (Section 4.4.2)
- hyperT-hypo = **HyperTwit** on **Hypo** (Section 4.4.2)

Figure 4.16 provides a visualisation of all models and and all experiment settings as evaluated in this chapter. This particular heatmap visualises the rankings for each model across each experiment setting from highest mean F1 to lowest mean F1. From this figure it can be observed that **BERT+PI$_R$** ranks among the top 3 models, in terms of mean F1, in 6 of the 9 different experimental settings under which all models were evaluated in this chapter. The model is also the top ranked model under 3 experimental settings. From this overview it is clear that this model is the best performing model on average across the different experimental settings. The next best model under this analysis is the **BERT+PI$_S$** which is among the top 3 ranked models across 5 of the 9 different

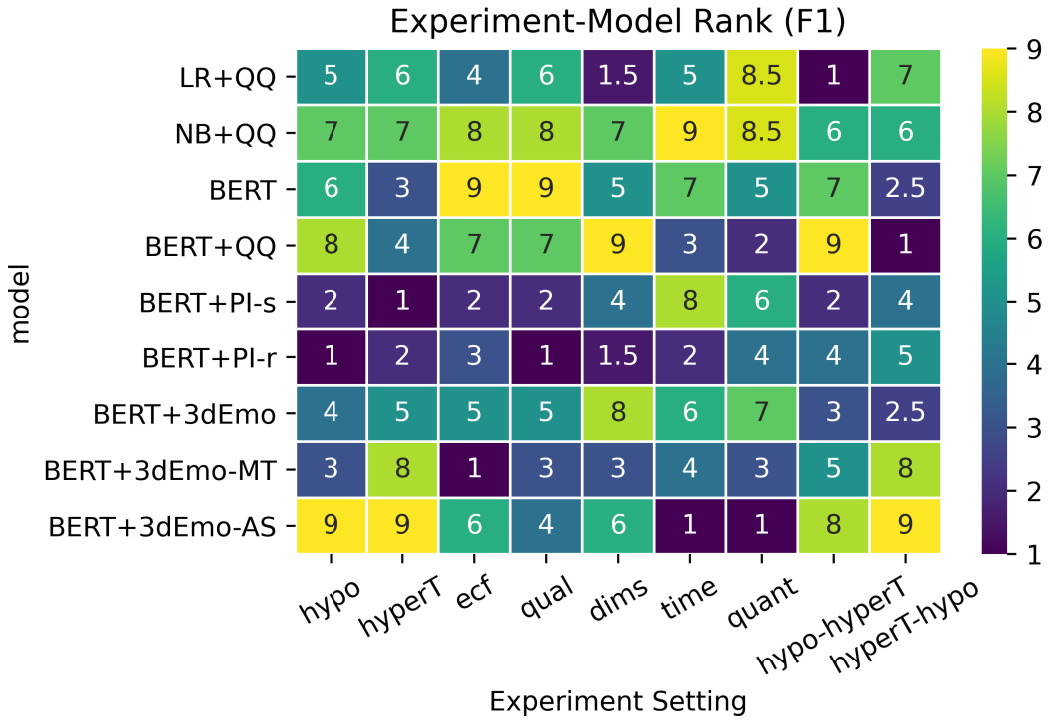

Figure 4.16: **Model-Experiment Rankings (F1)**

This heatmap visualises the ranking (mean F1 in descending order) of each model for each experiment setting as evaluated in this chapter. Models are represented along the y axis, experiment settings are represent along the x-axis (experiment settings refer to the individual evaluations for in-domain, cross-domain and hyper-probe experiments). **BERT+PI**$_R$ ranks in the top 3 models for mean F1 in 6 out of 9 experiment settings.

experiment settings. This model performs relatively poorly on the Time Period and Intrinsic Quantities experiment settings in **HyperProbe**. This suggests that this model is particular poor at dealing with quantities with respect to hyperbole. The top 2 model configurations across all experiment settings involve the incorporation of privileged information as proposed in Section 4.3. This showcases the utility of incorporating privileged information into a hyperbole detection model.

Figure 4.17 provides a visualisation of of the relative difficulty of the various experiment settings as introduced and evaluated in this chapter. There are some clear patterns in this visualisation that give a good indication of the difficulty of an experimental setting. The **HYPO** and **HyperTwit** datasets as well as the ECF tests in **HyperProbe** are the easiest experiment settings. With most models achieving a mean F1 in the top 3 of all experiment settings. **BERT+3dEmo**$_{MT}$ and **BERT+3dEmo**$_{AS}$ on **HyperTwit**

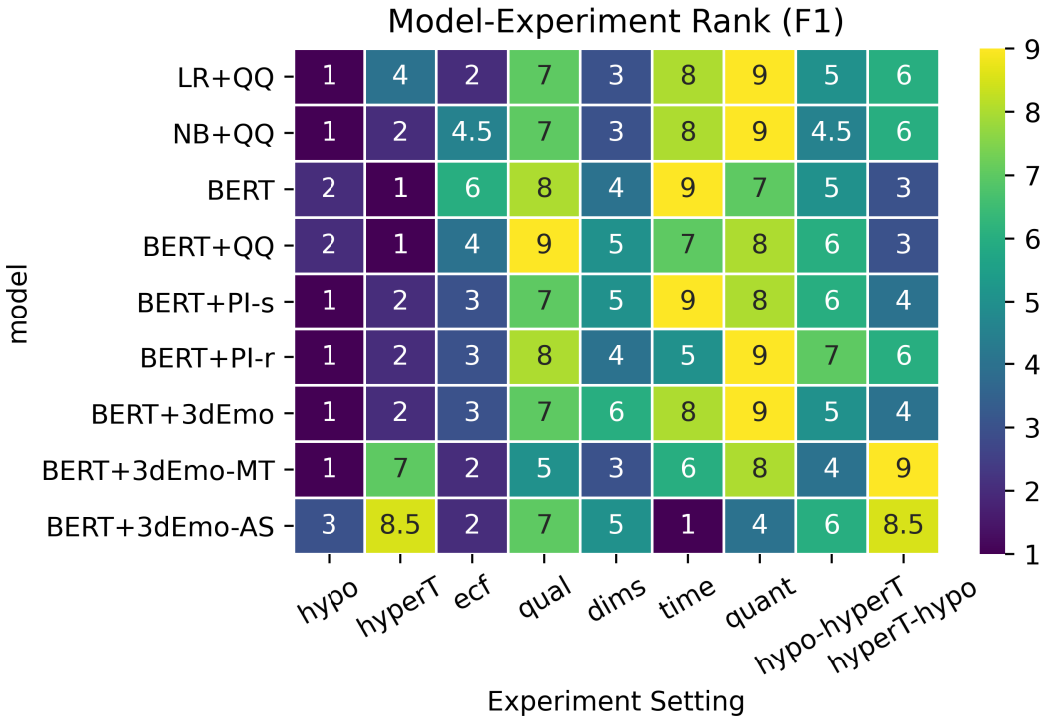

Figure 4.17: **Experiment-Model Rankings (F1)**

This heatmap visualises the relative difficulty (mean F1 in descending order) of the various experimental settings for all models as evaluated in this chapter. This diagram reveals that the **HYPO** dataset is the easiest model (i.e. highest F1 for that model) for 6 of the 9 models and is in the top 3 for all models, suggesting that it is the easiest experimental setting. Conversely, the Intrinsic Quantity test (quant) is the hardest experimental setting.

being outliers. This visualisation also indicates that the **HYPO** dataset appears to be the simplest dataset, which is intuitive given the idiomatic nature of the hyperboles in the dataset. Conversely, the Intrinsic Quantities, Time Period and Qualitative tests in **HyperProbe** are challenging datasets for most models.

These two figures help to summarise a number of key findings from the experiments and results in this chapter:

- Models that incorporate privileged information are the best performing models across a variety of experiment settings

- The models that incorporate affective signals are inconsistent across the various experiment settings

- The **HYPO** dataset is the easiest dataset across the models suggesting the relative ease of detecting idiomatic hyperbole
- The Intrinsic Quantities and Time Period tests in **HyperProbe** are relatively challenging datasets. Suggesting that models have a poor understanding of the range of plausible quantities required to understand if an extreme contrast is being expressed.
- The Qualitative test in **HyperProbe** is also challenging for most models. This suggests the models do not understand the hyperbolic use of adjectives when describing objects.

The evaluation of existing methods for hyperbole detection indicated that hyperbolic language on social media is a challenging phenomena. It was observed that hyperbole on Twitter was harder to accurately detect compared to idiomatic hyperbole suggesting more complex expression of hyperbole on Twitter. This finding partially answers research question i) and satisfies research objective ii) regarding *how hyperbole expressed on Twitter is different from idiomatic hyperbolic expressions*. This finding also provides evidence towards research question ii) and research objective ii) by showing that *existing NLP approaches to hyperbole detection result in poor accuracy, especially with respect to hyperbolic expressions as expressed on Twitter*.

The proposal and evaluation of models that incorporate affective signals through a variety of mechanisms and models that incorporate literal paraphrases as a type of privileged information provided an insight into how hyperbole detection models could be improved. Specifically, it was observed that the incorporation of literal paraphrases resulted in considerable improvements over baseline hyperbole detection models particularly on idiomatic hyperbole. This finding contributes to research objective iii) by *developing and evaluating machine learning algorithms on the task of hyperbole detection*.

A detailed error analysis identified that the promising results could be contributed to improvements in the detection of ECF hyperbole. The contrast in this type of hyperbole is generally encoded in a small number of tokens so the incorporation of literal paraphrases provides the necessary context to train a model. However, more complex types of hyperbole require significant editing during the process of paraphrasing and do not provide adequate context for detection. Identifying better annotation strategies for complex hyperbole is an important area of future research. The findings from this error analysis address research question iii) by providing insight on *how models for hyperbole detection can be improved moving forward*

CHAPTER 5

TOWARDS COMPUTATIONAL HYPERBOLE INTERPRETATION

5.1 Introduction

This chapter poses automatic hyperbole interpretation as a paraphrasing task and evaluates baseline models that attempt to address this problem. The content in this chapter addresses the research questions and aims of this thesis, (see Sections 1.3 and 1.4), in the following ways;

i. The evaluation of various paraphrasing models for hyperbole interpretation to *assess the ability of existing NLP models to interpret the intentions of hyperbolic expressions on social media*. (Research Question ii, Research Objective ii)

ii. Detailed error analysis of model interpretations seeks to *identify areas of further research for improving interpretation of hyperbole* (Research Question iii, Research Objective iii)

The chapter is structured as follows:

- Section 5.2 motivates the task of hyperbole interpretation, a review of literature of NLP approaches to similar problems are also covered.
- Section 5.3 details the implementation of various baseline methods for hyperbole interpretation.

Original (x)	:	that was a hot mess inside a dumpster inside a train wreck
Interpretation (y)	:	that was *terrible*
Removed (x_r)	:	[0,0,1,1,1,1,1,1,1,1,1,1]

Figure 5.1: **Example Data**

Original is the source text. ***Interpretation*** is a literal interpretation of the source. ***Removed*** is a binary sequence aligned with original, a value of 1 indicates that the word at the corresponding position was removed during interpretation.

- Section 5.4 describes the design of experiments to probe the similarity, fluency, semantic meaning and hyperbolicity of interpretations generated by various models.
- Section 5.5 presents the results of these experiments.
- Section 5.6 details a manual error analysis of automatically generated paraphrases.
- Section 5.7 concludes the chapter.

5.2 Natural Language Generation and Figurative Language

Hyperbole is an understudied figure of speech despite high prevalence and frequent co-occurrence with other figures of speech, see Chapters 3 and 4 for further details. Particularly, the computational study of hyperbole has been overlooked compared to computational studies on other figures of speech. Approaches to hyperbole detection, and general figurative language detection, were reviewed in Chapter 3.

Natural Language Generation (NLG) tasks relating to figurative language are scarce, more-so than Text Analytics approaches to figurative language (i.e., figurative language detection). The generation of figurative utterances is the most common NLG task related to figurative language. The approaches to figurative language generation are focused on sarcasm and metaphor generation, similar to to the predominance of these two figures as a focus of figurative language detection research.

Approaches to sarcasm generation focus on the reversal, or flipping, of sentiment and semantic or sentiment incongruity [27, 89, 142, 153]. A rule-based sarcasm generator was proposed that relied on eight different hand-crafted rules to generate a sarcastic utterance from a non-sarcastic input [89]. Several of these rules are based on the

flipping of sentiment, such as computing the sentiment of the input and generating output that is of opposite sentiment or computing the sentiment of a verb in the input and generating a situation with opposite sentiment as part of the output. To address incongruity, randomness is used to generate reasons and responses that are incongruous with the user input. The generation of outputs in this approach were based on regular expressions rather than statistical or neural language generation. A recent approach to sarcasm generation use neural NLG techniques to generate sarcastic utterances that rated higher than those generated by humans in manual evaluations [27]. This framework focused on sentiment reversal and semantic incongruity. The valence of an input sentence was reversed by replacing evaluative words with lexical antonyms ('this is **great**' → 'this is **bad**'). A language model for commonsense (COMET[19]) was used to generate scenarios based on the input and a pre-trained language model [124] computed the incongruity between the generated scenarios and the input. The most incongruous scenario was then appended to the valence reversed input.

Systems for the generation of metaphor and simile have been proposed ranging from rule-based systems to those based on statistical and neural language generation techniques. [28, 221, 229, 250**?**]. A recent proposal for metaphor generation was via controlled NLG using a pre-trained language model for sequence generation (BART) [116**?**]. This approach involved training a model on parallel pairs of literal and metaphorical sentences with a modification to the decoding objective to favour metaphorical replacements of verbs rather literal replacements. The modification of the decoding objective in a seq2seq model is a popular method for controlled text generation which will we cover later in this section [80, 128, 252]. A framework to embellish natural language via the injection of automatically generated similes was recently proposed [250]. This framework first predicts where a simile should be inserted into the original sentence, using BERT and a linear classification over the token sequence, then generating a simile that fits within the context of the predicted location for insertion. Experimental results show promising results for the feasibility of simile generation in context but further research required to refine the generation.

A common theme amongst the generation methods proposed for the various figures of speech is the iterative improvements to a few key heuristics particular to that figure of speech. Sarcasm generation methods focus heavily on the sentiment flipping operation and the contextual incongruity or metaphor models that focus on the transition of verbs from literal to metaphorical that. A focus on these heuristics has been empirically shown to produce reasonable or even state-of-the-art when dealing with a particular figure of

speech but lack generalising to to other figures. Specifically, the flipping of sentiment is a unique feature of sarcasm and models that focus on this heuristic are not applicable to hyperbole, metaphor, simile or other figures of speech.

The most relevant NLG research to the content in this chapter, however is the automated interpretation of figurative language rather than the generation. Much like the generation of figurative language this task has been formulated as a mono-lingual machine translation task (i.e., paraphrase generation) [18, 164, 203, 206].

The task of sarcasm interpretation has been formulated as a mono-lingual machine translation task [164]. The authors create a parallel dataset of 3000 sarcastic tweets with literal interpretations as produced by crowd-workers. A methodology is proposed by the authors that targets the sentiment flipping heuristic common to the sarcasm generation methodologies. Sentiment words are encoded according to a sentiment cluster and fed into a statistical machine translation algorithm that is trained to translate between opposing sentiment clusters. A de-clustering process then replaces the cluster label with a sentiment bearing word ('*i just* ***love*** *mondays #sarcasm*' → '*i just* ***hate*** *mondays*'). Experimental results showed that the proposed approach generated better interpretations of sarcasm than statistical and neural machine translation baselines, showcasing the potential for automated sarcasm interpretation.

Metaphor interpretation as a paraphrasing task has been limited to the translation of metaphorical verbs into literal verbs. One such work introduces an annotated dataset and proposes a metaphor interpretation model that estimates the probability of a replacement verb (i.e. the interpretation) co-occurring with other words in the context of the original text sequence. The dataset is a subset of the British National Corpus that is annotated for metaphor using the Metaphor Identification Procedure (MIP) [67] guidelines. The proposed model takes a sentence with a singular metaphorical verb within a literal context as input then generates and ranks a list of possible candidate replacement verbs. Hypernym relations from WordNet are used to filter candidate verb translations based on the overlap in shared concepts (i.e., hypernyms) between the metaphorical verb and the possible paraphrases. The remaining candidate translations are then ranked based on selectional association measure as defined by Resnik [186], the top ranked candidate is then selected as the literal interpretation.

Style Transfer is another mono-lingual machine task related to the work presented in this chapter with many different transfer tasks, associated datasets and models proposed. Such as the transfer of texts from informal to formal English [183], the transfer of texts in to Shakespearean style [244] and the transfer of product reviews from positive to

negative amongst several others [117]. Several of the figurative language frameworks outlined are based on style transfer architectures or share many similarities. However, the most related to the work in this chapter is a style method for neutralizing subjective bias in Wikipedia[1] text which is utilised during experiments undertaken in this chapter [178]. The author hypothesises that this task is similar to the neutralizing hyperbolic expression due to the attenuation of

The literal interpretation of hyperbole requires rich knowledge about the physical and non-physical world and an ability to reason with that knowledge, a long-standing goal in NLP research [19, 194]. Take for example the following hyperbolic expressions and possible literal interpretations;

- '*i would cut off all my limbs just to hear robert pattinson talk to me in a southern accent*' → '*i* ***want*** *to hear robert pattinson talk to me in a southern accent*'
- '*Sorry your password must contain the entire alphabet. your left foot. a theme song to a television show. and the blood of your enemies*' → '***These*** *password* ***requirements are excessive***'

A reader understands that 'cut off all my limbs' is used to express desire in this context because of our familiarity with this kind of exaggeration in daily informal language. The second example an example of complex and novel hyperbolic expression; as a reader we understand this as a hyperbole that is expressing dissatisfaction at overly complicated password requirements by our knowledge and experience with passwords. Automatically generating hyperbole interpretations is an important task given prevalence of hyperbole and the complex and varied intentions of hyperbolic expression. Additionally, the promising results of models that incorporated literal interpretations as privileged information, see Chapter 4, heightens the importance of automated literal interpretation. Manual composing literal interpretations is a labour intensive task, the automated generation of literal interpretations will allow for scaling up hyperbole detection.

5.3 Methodology

5.3.1 Hyperbole Interpretation

The formulation of the hyperbole interpretation task expands on hyperbole detection task. Let X be a short source text (i.e., tweet or sentence) and Y be a literal interpretation

[1] https://www.wikipedia.org/

of X, see Figure 5.1. Also let X_r be a sequence of binary values of length w, where w is the number of words in X. The values of X_r are computed by $d(X,Y)$, where d is a function that identifies words that are in X but not in Y (i.e., words removed from the hyperbolic text during literal interpretation) [2]. Given that Y is a literal interpretation of X, it follows that any words removed from X can be considered as contributing to the hyperbolic nature of X. Given this data the interpretation of hyperbole can be considered as a mono-lingual translation task with source sequence X and target sequence Y.

5.3.2 Baselines and Models

Two naive baselines are used to provide context for the automatic evaluation metrics and to gain insight into the process of paraphrasing a hyperbole as performed by expert annotators. The **Delete** baseline simply deletes the tokens that were removed during the literal paraphrase. Formally, $XOR(X_r)$ is used to mask the original sequence X, see Figure 5.1. This baseline preserves the non-hyperbolic content in the source text however, will often result in incomplete sentences. The **Delete** can provide insight into the amount of non-hyperbolic content preserved during the paraphrasing process. The **Copy** baseline simply copies the original source text X with no modifications and can provide a point of reference during evaluation.

A machine translation model is used to perform a round-trip-translation to generate a generic paraphrase (**RTT**)[148]. This model is based on a transformer architecture [228] and employs ensembled back-translation for data augmentation. Data for training consisted of various large parallel corpora (i.e. Newscrawl, CommonCrawl[3]) with another stage of fine-tuning performed on smaller domain specific corpora. Strong results were observed across four language directions (English → German, German → English, English → Russian, Russian → English), showing significant improvements over other automated translation systems and human translations. The motivation for using this model is to understand the similarities between a generic paraphrase and literal paraphrase of a hyperbole. Specifically, *is the hyperbolicity removed in a generic paraphrase?*. Further, *does this depend on type of hyperbolic expression?*.

Modular is a state-of-the-art model for style transfer, designed to detect and remove subjective bias from Wikipedia articles [178]. A two-step model that consists of a detection and edit modules. The detection module consists of a sequence tagger that utilises BERT combined with handcrafted bias features, see Figure 5.2, that is trained to detect

[2] https://github.com/paulgb/simplediff
[3] https://commoncrawl.org/

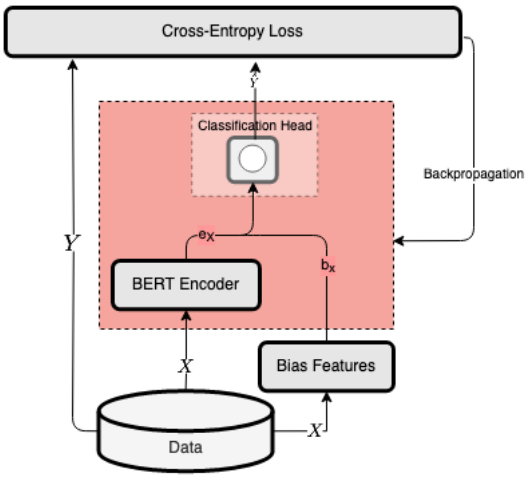

Figure 5.2: **Diagram of Tagging Module from Modular**

Tagging module consists of BERT as encoder, handcrafted bias-features and a linear classification layer.

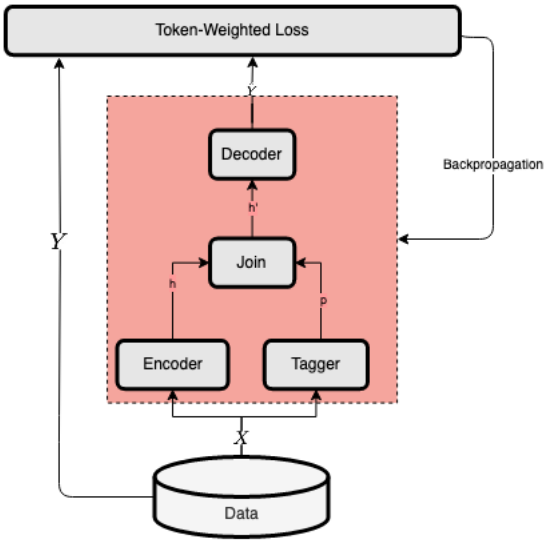

Figure 5.3: **Diagram of Modular**

Modular consists of a tagger for controlling language generation, encoder-decoder for sequence generation and a join embedding mechanism.

tokens that display subjective bias. The edit module consists of a sequence to sequence architecture designed with a BiLSTM encoder and an LSTM decoder. The edit modules is trained on Wikipedia articles that are flagged for subjective bias and the edited version where the subjective bias has been removed by human editors to address the subjective bias complaints. The key aspect of this model is the combination of these two modules where both models are pre-trained individually and then jointly fine tuned. In this joint fine-tuning operation the output of the detection modules is used to control the edit module, see Figure 5.3. The motivation for using this model is to understand the similarities between the automatic neutralisation of subjective bias and the hyperbole interpretation task. Specifically, *is the hyperbolicity removed when subjective bias is neutralised?*.

A hyperbole is often used to convey an excessive evaluation which is an extreme form of subjective bias. Examples of subjective bias from Wikipedia articles provided in the original paper by the authors were hyperboles (e.g., '*Go is the **deepest** game **in the world***'. The neutralizing of subjective bias and the interpretation of hyperbole are similar tasks: both are minimal translations (i.e., minimal edits) that target similar semantic content (i.e., evaluative/emotive content) with a similar goal (i.e., a reduction evaluative/emotive). However, the threshold for editing the evaluative/emotive content and the level of attenuation are different between the tasks. The neutralisation of subjective bias has a low threshold for evaluative/emotive content, whereas the threshold for removing hyperbole is towards the extreme end. The reduction operation when neutralising subjective bias is more extreme because all emotive/evaluative content must be removed, whereas the neutralisation of hyperbolic content is more of an attenuation operation that reduces the extreme contrast.

5.4 Experiments

Experiments are conducted to establish capabilities and limitations for the computational literal paraphrasing of hyperbole in online user-generated content. Details of evaluation metrics, baseline models, experimental setup and results of the experiments are provided. The specific mono-lingual machine translation task here is to generate a literal paraphrase of a hyperbolic source text. The original hyperbolic tweet, X, is treated as the source sequence and the literal paraphrase created by the annotators, Y, as the target sequence.

A combination of automated metrics and manual assessment are used to evaluate the

generated literal paraphrases. Despite the well documented weaknesses of automated metrics in evaluating the performance of machine translation systems, [132, 237], several of these metrics still provide insight given the specific nature of the problem addressed here. Recent research has argued for a reduction in the usage and reliance on **BLEU** [157] and **TER** [210] for evaluating machine translation systems [132]. However, both of these metrics are used for evaluation because the hyperbole paraphrasing task is a constrained translation problem where only minimal edits (i.e., perhaps deletion of a single word) may be all that is required to interpret a hyperbole. Both of these n-gram based metrics are good indicators of the amount of changes at the token level between sequences. The SacreBLEU implementation is used to calculate BLEU scores [176] [4]. An author implementation of Translation Edit Rate (**TER**) [210] relying on *simplediff* [5] to compute the edits between the generated literal paraphrase and the expert human literal paraphrase is used for **TER**. The $\mathbf{SIM}_R$ and $\mathbf{SIM}_A$ metrics are also reported. These metrics are based on semantic similarity between sentences computed using cosine similarity and word representations. These two metrics can be considered as variations of other semantic similarity metrics for machine translation [126, 237], see Section 3.4.2 for further details and motivations on these metrics.

For manual assessment of the generated interpretations, the **Fluency**, **Meaning** and **Hyperbolicity** metrics are introduced. **Fluency** is measured on a Likert scale [6]. Specifically, '*Is $\hat{Y}$ is more readable than Y*', where $\hat{Y}$ is the generated interpretation and Y is a ground truth interpretation. **Meaning** is an evaluation on the amount of meaning preserved between the generated output and input text. A Likert scale[7] is used to measure meaning. Specifically, '*Does $\hat{Y}$ mean the same as Y*'. **Hyperbolicity** is a comparison of the hyperbolic nature of a pair of texts. **Hyperbolicity** is measured on a Likert scale.[8] Specifically, '*Is $\hat{Y}$ is less hyperbolic than X*'. The manual assessment was performed by one of the authors and was blind with respect to the model responsible for generated interpretation (i.e. $\hat{Y}$).

Tweets from HyperTwit that were identified as hyperbolic during annotations, 2087 in total, were used for experiments. This data was split into training, development and test sets containing approximately 1669, 313 and 105 tweets respectively. The pre-trained implementation of the **RTT** model was used, this implementation is provided

[4] https://github.com/mjpost/sacrebleu
[5] https://github.com/paulgb/simplediff
[6] [-2,2] from Strongly Disagree to Strongly Agree
[7] [-2,2] from Strongly Disagree to Strongly Agree
[8] [-2,2] from Strongly Disagree to Strongly Agree

Model	**BLEU (↑)**	**TER (↓)**	**SIM$_R$ (↑)**	**SIM$_A$ (↑)**
Copy	50.542	0.312	-	-
Delete	**63.891 (2.251)**	**0.157**	-	-
RTT	24.658 (1.575)	0.453 (0.011)	0.839 (0.007)	0.779 (0.029)
Modular$_{WNC}$	41.768 (2.816)	0.358 (0.026)	0.788 (0.015)	0.743 (0.070)
Modular$_{HT}$	48.070 (3.104)	0.275 (0.089))	**0.912 (0.006)**	**0.785 (0.068)**
Concurrent$_{HT}$	36.909 (2.596)	0.391 (0.030)	0.752 (0.049)	0.718 (0.026)

Table 5.1: **Paraphrase Experiment Results**

Results from experiments on automated paraphrasing of hyperbole. Results are the mean and standard deviations across multiple random dataset splits.

Model	**Fluency**	**Meaning**	**Hyperbolicity**
Copy	-	-	-
Delete	-0.580(0.835)	-0.240(1.079)	**1.840**(0.581)
RTT	**-0.039(0.631)**	0.549(0.832)	-0.804(0.849)
Modular$_{WNC}$	-0.300(0.543)	**0.600(0.571)**	-0.680(1.039)
Modular$_{HT}$	-0.360(0.663)	0.380(0.901)	0.360(1.467)
Concurrent$_{HT}$	-0.600(0.728)	0.400(0.808)	-0.940(0.586)

Table 5.2: **Paraphrase Experiment Results - Manual Evaluation**

Manual evaluation results from experiments on the interpretation of hyperboles in HyperTwit.

by the FAIRSEQ [156] [9] Python library. German is used as the pivot language and the model was run with standard parameters. For the **Modular$_{WNC}$** model, the code and model checkpoint provided on the GitHub page[10] associated with the publication was used. For the **Concurrent$_{HT}$** model, the Pointer Seq2Seq implementation was used with a learning rate of 0.0003, debias weight of 1.3, BERT as encoder and BERT embeddings. The decoder was pre-trained for 10 epochs on 35k unlabelled tweets on the same keywords as used in HyperTwit and the model with best performance on the development set was retained. The **Concurrent$_{HT}$** model was trained for 20 epochs and the model with best performance as evaluated on the development set was retained. The trained **Concurrent$_{HT}$** version described above was used as checkpoint for the joint training of **Modular$_{HT}$** model, aligning with the instructions from the original paper.

[9] https://ai.facebook.com/tools/fairseq/

[10] https://github.com/rpryzant/neutralizing-bias

5.5 Results

The results of the experiments are shown in Tables 5.1 and 5.2. The naive **Delete** baseline achieves the best *BLEU* and *TER*. This result suggests that the deletion of hyperbolic tokens is a common edit operation performed by annotators when interpreting a hyperbole. The interpretations of $\textbf{Modular}_{HT}$ are the most semantically similar with the ground truth interpretations, SIM_R and SIM_A of 0.912 and 0.785 respectively.

With respect to manual assessment metrics a Kruskal-Wallis test is performed for each metric revealing that there is a statistically significant difference between all models for *Fluency*[11], *Hyperbolicity*[12] and *Meaning*.

For *Fluency* it can be seen that overall the generated interpretations are relatively fluent with the exception of $\textbf{Concurrent}_{HT}$ which is surprisingly worse than **Delete**. The interpretations of **RTT** are adequate generic paraphrases, notably achieving the highest fluency with a mean of -0.039 (i.e., barely less fluent than the ground truth interpretation on average). However, the *Hyperbolicity* of **RTT** is poor with a mean of -0.804 indicating the the hyperbolicity of the original tweet was not removed, on average, in the paraphrases generated by **RTT**. This suggests generating literal interpretations via generic paraphrase is an inadequate strategy with respect to the hyperbolicity of the interpretations.

$\textbf{Modular}_{HT}$ and **Delete** are the only two models to achieve positive *Hyperbolicity* of, 0.360 and 1.840 respectively. This indicates that on average the interpretations of $\textbf{Modular}_{HT}$ and **Delete** are less hyperbolic then the original hyperbolic tweets. However, these two models differ considerably in *Meaning* and *Fluency*. Particularly, **Delete** is the only model with a negative mean *Meaning*, indicating that the hyperbolic tokens are on average important to the intended meaning of the hyperbolic utterance. This is intuitive because some hyperboles must be interpreted to maintain the intended meaning and simply can not be deleted (e.g., *This video gave me* ***eye cancer*** *→ This video gave me*).

The other important result to observe here is the statistically significant ($p < 0.01$)[13] difference in *Hyperbolicity* between $\textbf{Modular}_{WNC}$ and $\textbf{Modular}_{HT}$. This result suggests that automatically neutralizing subjective bias and interpreting hyperbole are considerably different tasks, despite similarities mentioned in Section 5.3.2.

[11] $p < 0.01$
[12] $p < 0.01$
[13] Mann-Whitney U test

Original Tweet(X)	**Ground Truth Interpretation (Y)**	**Generated Interpretation ($\hat{Y}$)**
you are like an electron and i am like a proton and they say that opposites attract	**we should be together**	**you are like opposite**
istg if anyone says that i look like her your obviously blind asf	istg if anyone says that i look like her **you are wrong** asf	istg if anyone says that i look like her your obviously asf
old nollywood shaped an entire generation	old nollywood **was influential**	old nollywood **is a good terrible**
this choreo looks impossible im	this choreo looks **very difficult** im	this choreo looks **terrible**
kamala harris has the exact face as norma bates	kamala harris has **a similar** face as norma bates	kamala harris has a face
malcovichs partner is my ideal personality	malcovichs partner is **a great** personality	malcovichs partner is **a good**
my mama dont like you and she likes everyone	my mama dont like you and she likes **most**	my mama dont like you and she likes
whole 2nd verse on jaded is biblical	whole 2nd verse on jaded is **great**	whole 2nd verse on jaded is **terrible**
bts is garbage whta the fuck is this song	bts **is terrible so is** this song	bts is **terrible** whtahta
going from 8 hour shift directly to 3 hour class is my stupidest idea why do i keep doing this	going from 8 hour shift directly to 3 hour class is my **stupid** idea why do i keep doing this	going from 8 hour shift directly to 3 hour class is a **stupid**
why do i always feel this way	why do i often feel this way	why do i feel this way **this way**
this debate already feels like a benadryl nightmare	this debate already **boring and frustrating**	this debate already feels like a **ben**

Table 5.3: **Paraphrase Errors 1**

Original Tweet (X) is the original tweet. ***Ground Truth Interpretation (Y)*** is the literal interpretation provide by annotator. ***Generated Interpretation ($\hat{Y}$)*** is the tweet generated by **Modular**$_{HT}$. **Emphasis** indicates words not in the original tweet.

Original Tweet(X)	**Ground Truth Interpretation (Y)**	**Generated Interpretation ($\hat{Y}$)**
every family have their own secrets either good or bad	**many** families have their own secrets either good or bad	every family have their own secrets <*either good or bad*>
i feel like an old person even when i walk i hurt from literally everywhere c	i feel like an old person even when i walk i hurt	i feel like an old person even when i walk i hurt <*from literally everywhere c*>
if u love batman more than spiderman you have no taste just saying	spiderman **is better than batman**	if u love batman more than spiderman you have no taste <*just saying*>
my real life hasnt started yet the real me is still asleep so thats why my life is such garbage	my life **is so terrible**	my real life hasnt started yet the real me is still asleep so thats why my life is <*such garbage*>
the damn fly was more personable than the pile of shit he was on	he was **not very personable**	the damn fly was more personable than the pile <*of shit*> he was on on
the fbi has officially declared me handsome sorry	**i am** handsome sorry	the fbi has officially declared me handsome <*sorry*>
we wont have to listen to trump blathering about voting fraud it will feel like america again	we wont have to listen to trump blathering about voting fraud it will feel so good	we wont have to listen to trump blathering about voting fraud it will feel like america
hogan hitting that 180 real quick on his response to covid dude is a bum and always will be	hogan **changing his mind** real quick on his response to covid dude is **hopeless**	hogan hitting that 180 real quick on his response to covid dude is a bum <*and always will be*>
jen rubin decided to dress up like a heartless gremlin 29 days early for some reason	jen rubin decided to dress up 29 days early for some reason	jen rubin decided to dress up like a heartless gremlin 29 days early

Table 5.4: **Paraphrase Errors 2**

Hyperbole Type	**Modular**$_{HT}$	**Modular**$_{WNC}$	**RTT**
ECF	0.448(1.503)	-0.666(0.916)	-0.640(1.113)
Qualitative	0.444(1.423)	-0.666(1.188)	-0.882(0.485)
Quantitative	-0.125(1.553)	-0.750(1.164)	-1.125(0.353)

Table 5.5: **Hyperbolicity, Hyperbole Type and Model**

Hyperbolicity scores across hyperbole types and models

5.6 Error Analysis

An error analysis is undertaken to identify avenues for future research. The paraphrases generated by the **Modular**$_{HT}$ model are the focus of this analysis as that model performs the best with respect to adequately paraphrasing hyperbole without lacking fluency or meaning.

A particular error observed was paraphrases in which the hyperbolicity of a tweet was reduced, but the intended meaning was distorted ('*this choreo looks **impossible** im*' → '*this choreo looks **terrible***'), see Table 5.3. Another error is the reduction in some of the evaluative content from the original tweet but not enough to reduce the overall hyperbolic nature of the tweet, see Table 5.4. This error was often the case in tweets with long hyperbolic phrases or multiple hyperboles (*the damn fly **was more personable than the pile of shit** he was on → the damn fly was more personable than the pile <of shit> he was on*).

During manual assessment it was observed that paraphrases generated by **Modular**$_{HT}$ appeared to be less effective at reducing the hyperbolicity in long and complex hyperbolic expressions, see Tables (5.3 and 5.4). However, an insignificant positive Pearson correlation, 0.086, was observed between the ratio of hyperbolic tokens in a tweet and the hyperbolicity rating provided during manual assessment. This suggests that the length of the hyperbole alone is not indicative of the difficulty in generating an adequate interpretation.

With respect to the different types of hyperbole, Quantitative hyperboles are the most significant source of error. Pairwise correlation analysis between **Modular**$_{HT}$ vs. **Modular**$_{WNC}$ and **Modular**$_{HT}$ vs. **RTT** found a positive correlations between the hyperbolicity scores and the type of hyperbole[14], see Table 5.5. From this table it can be observed that the hyperbolicity is similar between ECFs and Qualitative hyperboles and considerably better than the hyperbolicity for Quantitative hyperboles across the model

[14] $r = 0.94$ and $r = 0.99$ respectively

pairs.

5.7 Conclusion

This chapter poses automatic hyperbole interpretation as a paraphrasing task as well as introduces and evaluates baseline models for this problem. Several contributions towards the research questions and aims of this thesis have resulted from the work in this chapter.

The evaluation of various paraphrasing models for hyperbole interpretation showed that the automatic generation of literal interpretations is a challenging task and fruitful avenue for further research. It was observed that generic paraphrases do not adequately interpret the hyperbolic content present in an expression. Further, models trained for neutralizing of subjective bias do not adequately remove hyperbolic content. These findings contribute to research question ii) and research objective ii), specifically by showing *that various models for style-transfer and generic paraphrases do not adequately interpret the excessive contrast inherit in a hyperbolic expression*.

A detailed error analysis found that models often defaulted to simple heuristic of simply deleting the hyperbolic content and not interpreting the intend meaning of that content. Another error identified was the generation of text that did not correctly interpret the hyperbole or was nonsensical in some cases. Findings from the error analysis addressed research question iii) and research objective iii) by *identifying areas of further research for improving interpretation of hyperbole*.

Part III

Conclusion

CHAPTER 6

DISCUSSION AND CONCLUSION

6.1 Introduction

This chapter concludes the thesis by restating the thesis statement, explicitly states the answers to the research questions and provides evidence of contributions to research objectives. Finally, a discussion on potential directions for future research that have emerged from the findings presented within the thesis will close out this chapter and the document.

6.2 Thesis Statement

Accurate computational detection of figurative language on social media is a complex task that requires modification of existing and creation of new datasets and methodologies.

6.3 Research Questions

This section will restate the key research questions, see Section 1.3, addressed in this thesis as well as summarising the answers to these questions as revealed throughout the body of this thesis.

6.3.1 Research Question i)

How does figurative language occur on social media and how does this differ in comparison to the occurrence of figurative language in traditional forms of communication?

A number of findings from this thesis provide answers to this research question:

i. Figurative language occurs frequently on social media in the context of symptom and disease words, importantly it was observed that some symptom and disease words were more likely to be used in a figurative sense than in a literal sense (see Chapter 2).

ii. A significantly greater prevalence of hyperbole was observed on Twitter compared to that found in corpus studies of hyperbole in different communicative forms (i.e., conversational English), (see Chapter 3).

iii. Hyperbole was commonly used on Twitter to express sentiment on a broad range of topics, often in a complex manner that went beyond simple comprehension on the linguistic contents of the expressions (see Chapter 3).

iv. With respect to the diversity of hyperbole expressions on Twitter, some hyperbolic expression were simply parroted by different Twitter users but also a number of novel, elaborate and specific hyperbole were expressed by Twitter users (see Chapter 3).

These findings indicate that figurative language, particularly hyperbole, is common and complex linguistic phenomena on social media. Computationally understanding figurative language is an important task to fully understand discourse on social media and important research area.

6.3.2 Research Question ii)

How adequate are current resources (i.e., datasets, models) for the accurate detection and interpretation of figurative utterances found on social media?

A number of findings from this thesis provide answers to this research question:

i. The majority of false positive errors observed when evaluating a text classifier trained to classify health mentions on Twitter were a result of figurative expressions of disease and symptom words (see Chapter 2).

ii. Both idiomatic hyperbole and hyperbole expressed on Twitter were difficult to detect using a variety of text classifiers, suggesting that current NLP methodologies were inadequate for the task (see Chapter 4).

iii. Generic paraphrases do not adequately interpret the hyperbolic content present in an expression indicating difficulty in understanding the intentions of hyperbole (see Chapter 5).

iv. Models trained for neutralizing subjectively biased content do not adequately remove hyperbolic content indicating difficulty in understanding the intentions of hyperbole (see Chapter 5).

These findings strengthen the key claim in the thesis statement that figuratively language is a complex phenomena and requires new resources and methodologies for accurate understanding.

6.3.3 Research Question iii)

How can the computational detection and interpretation of figurative utterances be improved?

A number of findings from this thesis provide answers to this research question:

i. Experiments provided evidence that better incorporation of sentiment signals could improve the detection of figurative mentions within the context of text classifiers for symptom and disease words (see Chapter 2).

ii. Analysis showed that hyperbolic expressions of symptom and disease words for the purpose of exaggerating the opinion of an author, not to actually express that existence of a disease or symptom were wrongly classified. Better handling of hyperbolic expression is a key focus for improving figurative language understanding (see Chapter 2).

iii. The incorporation of literal paraphrases into text classification model resulted in considerable improvements over baseline classifiers trained to classify hyperbolic expression (see Chapter 4).

iv. Evaluation revealed that better annotation strategies for complex hyperbole is an important area of future research (see Chapter 4).

These findings strengthen the key claim in the thesis statement, that figuratively language is a complex phenomena, and provides solutions that improve understanding.

6.4 Research Objectives

This section will detail how the thesis satisfies the three research objectives described in Section 1.4.

6.4.1 Research Objective i)

Creation of annotated datasets that allow for the study of figurative language on social media.

A number of contributions are made to this research objective throughout the thesis:

i. The collection, annotation and exploratory analysis of **HMC2019** provides a resource for the study of figurative language in the context of symptom and disease words on Twitter (see Chapter 2).

ii. The data collection, annotation and exploratory data analysis of **HyperTwit** provided a resource the study of hyperbole as expressed on Twitter (see Chapter 3).

iii. The generation and annotation of **HyperProbe** provide a benchmark for behavioural testing of hyperbole detection models (see Chapter 3).

6.4.2 Research Objective ii)

Produce quantitative evidence of the phenomenon of figurative language on social media and how the phenomenon impacts the predictive performance of existing NLP text classification models.

A number of contributions are made to this research objective throughout the thesis:

i. Exploratory data analysis produced quantitative evidence of the presence of figurative language usage of symptom and disease words on Twitter (see Chapter 2).

ii. Experiments provided evidence that traditional text classifiers trained for detecting health events misunderstood figurative expressions of symptom and disease words as actual health events (see Chapter 2).

iii. Manual error analysis provided evidence that hyperbolic expressions of symptom and disease words were particularly challenging expression for text classifiers to understand (see Chapter 2).

iv. Experiments provided evidence that both idiomatic hyperbole and hyperbole expressed on Twitter were difficult to detect using a variety of NLP text classifiers, suggesting that current NLP methodologies were inadequate for the task (see Chapter 4).

v. Manual evaluation provided evidence that generic paraphrases do not adequately interpret the hyperbolic content present in an expression, indicating difficulty in understanding the intentions of hyperbole. This manual evaluation also showed that models trained for neutralizing subjectively biased content do not adequately remove hyperbolic content indicating difficulty in understanding the intentions of hyperbole (see Chapter 5).

6.4.3 Research Objective iii)

Develop and evaluate machine learning algorithms for the task of figurative language understanding on social media.

A number of contributions are made to this research objective throughout the thesis:

i. The proposal, implementation and evaluation of various text classification models for the figurative expression of symptom and disease words satisfied this particular research objective (see Chapter 2).

ii. The proposal, implementation and evaluation of models that incorporate affective signals through a variety of mechanisms and models that incorporate literal paraphrases as a type of privileged information satisfied this particular research objective (see Chapter 4).

iii. The proposal, implementation and evaluation of various mono-lingual machine translation approaches to understand how individuals interpret hyperbole and to automatically generate hyperbole interpretations (see Chapter 5).

6.5 Future Directions

It is clear from this thesis that the computational understanding of figurative language remains a challenging problem for NLP. Several fruitful directions for further research on the computational understanding of figurative language have emerged from this thesis.

6.5.1 Hyperbole

Hyperbolic expressions were a key focus in this thesis with several findings suggesting that continued focus on these expressions is required to achieve adequate understanding. The difficulty in detecting hyperbolic usage of symptom and disease words by various NLP models for text classification (see Chapter 2). Similar difficulties were observed when attempting to detect idiomatic hyperbole and hyperbole expressed on Twitter (see Chapter 3).

6.5.1.1 Compound Hyperbole

A key observation made throughout the thesis was the inability to deal with complex hyperbolic expressions, (see Sections 4.4 and 5.6). Particularly, when using literal paraphrases to help ground hyperbolic expressions to literal intentions. The literal paraphrasing of compound hyperbole was too destructive which was detrimental to approaches that used the paraphrase to teach a model the intentions of hyperbole.

Consider the following hyperbolic expression '*this policy will plunge the country into a chaos*'. This expression can be considered a compound hyperbole because the hyperbolicity of the expression is contained in multiple phrases in the expression. The use of the verb '***plunge***' in the verb phrase '*policy will* ***plunge*** *the country into a chaos*', and the noun '***chaos***' in the preposition '*into a* ***chaos***'.

Consider an example literal paraphrase of this hyperbole, ('*this policy will plunge the country into a chaos*' → '*this policy is a bad idea*'). This literal paraphrase is quite destructive, an example of a better paraphrase would be to replace the verb '*plunge*' with '*put*', and the noun '*chaos*' with the noun phrase '*bad situation*' (i.e. '*this policy will put the country into a bad situation*'). This paraphrase maintains the syntactic structure of the original expression and removes the hyperbolicity of both '*plunge*' and '*chaos*' in the sentence, whilst maintaining the syntactic structure and the immediate contexts in which the two hyperbolic tokens occur.

A hypothesis for the success of hyperbole detection on simple hyperbolic expressions was due to the minimal differences between the original and the paraphrase (e.g. '*what an absolute idiot*' → '*what an idiot*'). Careful and methodical annotation of compound hyperbole that focuses on syntax preservation whilst neutralizing hyperbole is likely to be a fruitful research direction. Several research questions are of interest:

- What is the relationship between the edit operations needed to interpret a hyperbole and the difficulty of classification?
- Can syntax-preserving literal paraphrases of compound hyperboles improve the hyperbole detection on these expressions?

6.5.1.2 Quantitative Hyperbole

Quantitative hyperbole is an important type of hyperbole that posed challenges to hyperbole detection and interpretation models (see Chapters 4 and 5). Findings indicated that NLP models did not understand excessive contrasts along various quantitative dimensions (e.g, time, size, currency, etc.). Research questions of interest for this topic:

- How to encode temporal knowledge into NLP models to better understand excessive contrasts along temporal scales?
- How to encode knowledge of intrinsic quantities of objects (e.g., height, weight, length, etc) into NLP models to better understand excessive contrasts along these scales?

6.5.1.3 Automated Interpretation of Hyperbole

Initial work on the automated interpretation of hyperbole was conducted in this thesis (see Chapter 5). Findings revealed a challenging task for existing approaches to similar tasks in mono-lingual machine translation.

Automatic generation of hyperbole interpretations would be beneficial for hyperbole detection models that rely on interpretations (see Section 4.3. Manual generation of these interpretation is a labour intensive annotation task that can introduce unwanted biases from human annotators. The automated generation of interpretations would also be helpful for the comprehension of these expressions and downstream tasks (i.e. aspect-based sentiment analysis).

6.5.1.4 Hyperbole and Sentiment Analysis

Hyperbolic expressions and intentions are laden with sentiment, however the sentiment being expressed is not always stated obviously (see Chapter 3. A deeper exploration of the relationship between hyperbolic expressions and sentiment analysis is a fruitful direction of research.

Consider the sentiment conveyed in the hyperbolic expression '*my bedroom is the size of a postage stamp*'. A *negative* opinion regarding the *size* of the *bedroom* is being expressed by comparison to the size of postage stamp, a ridiculous and excessive comparison. This expression is an example of hyperbole used to convey sentiment without the use of strong sentiment bearing words that requires world knowledge and reasoning to understand the intended sentiment of the expression. The task of identifying the polarity of sentiment expressed, the target of that sentiment and the aspect in such hyperbolic expressions is a fruitful avenue for further investigation.

Potential research questions that are fruitful for further research:

- Can NLP models for sentiment analysis accurately predict the sentiment expressed in hyperbolic expressions?
- Can NLP models for aspect-based sentiment analysis accurately identify the sentiment, target and the aspect in hyperbolic expressions?

6.5.2 Metaphor and Sarcasm

Hyperbolic expressions are an important figure of speech in social media and received significant attention throughout the thesis (see Chapter 3). Hyperbole has a high rate of co-occurrence with other types of figurative language, see Section 3.2. Therefore, more accurate understanding of hyperbolic expressions will likely result in a 'trickle-down' effect where benefits could be seen for the understanding of several other types of figurative language.

Probing this 'trickle-down' effect is a potential future direction of research. Research questions that follow up on this topic could be:

- Does the inclusion of literal interpretations help in the detection of other figurative devices (e.g. metaphor, sarcasm)?
- Are figurative expressions that employ multiple figurative devices (e.g. hyperbole and metaphor, hyperbole and sarcasm) more complicated than expressions that employ single figurative devices?

BIBLIOGRAPHY

[1] G. ABERCROMBIE AND D. HOVY, *Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of twitter conversations*, in Proceedings of the ACL 2016 Student Research Workshop, 2016, pp. 107–113.

[2] M. ABULAISH AND A. KAMAL, *Self-deprecating sarcasm detection: An amalgamation of rule-based and machine learning approach*, in 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE, 2018, pp. 574–579.

[3] M. ABULAISH, A. KAMAL, AND M. J. ZAKI, *A survey of figurative language and its computational detection in online social networks*, ACM Transactions on the Web (TWEB), 14 (2020), pp. 1–52.

[4] C. C. AGGARWAL AND C. ZHAI, *A survey of text classification algorithms*, in Mining text data, Springer, 2012, pp. 163–222.

[5] S. ALDRICH AND C. ECCLESTON, *Making sense of everyday pain*, Social science & medicine, 50 (2000), pp. 1631–1641.

[6] B. ALTINEL AND M. C. GANIZ, *Semantic text classification: A survey of past and recent advances*, Information Processing & Management, 54 (2018), pp. 1129–1153.

[7] V. ANDRYUSHECHKIN, I. WOOD, AND J. O' NEILL, *NUIG at EmoInt-2017: BiLSTM and SVR ensemble to detect emotion intensity*, in Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Copenhagen, Denmark, Sept. 2017, Association for Computational Linguistics, pp. 175–179.

[8] R. ARTSTEIN AND M. POESIO, *Inter-coder agreement for computational linguistics*, Computational Linguistics, 34 (2008), pp. 555–596.

[9] S. BACCIANELLA, A. ESULI, AND F. SEBASTIANI, *Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.*, in LREC, vol. 10, 2010, pp. 2200–2204.

[10] D. BAMMAN AND N. A. SMITH, *Contextualized sarcasm detection on twitter*, in Ninth international AAAI conference on web and social media, 2015.

[11] F. BARBIERI AND H. SAGGION, *Automatic detection of irony and humour in twitter.*, in ICCC, 2014, pp. 155–162.

[12] F. BARBIERI, H. SAGGION, AND F. RONZANO, *Modelling sarcasm in twitter, a novel approach*, in Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2014, pp. 50–58.

[13] A. BARUAH, K. DAS, F. BARBHUIYA, AND K. DEY, *Context-aware sarcasm detection using bert*, in Proceedings of the Second Workshop on Figurative Language Processing, 2020, pp. 83–87.

[14] J. BENNETT, *Omg! the hyperbole of internet-speak*, 2015.

[15] J. BERGER AND K. L. MILKMAN, *What makes online content viral?*, Journal of marketing research, 49 (2012), pp. 192–205.

[16] S. K. BHARTI, K. S. BABU, AND S. K. JENA, *Parsing-based sarcasm sentiment recognition in twitter data*, in Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, ACM, 2015, pp. 1373–1380.

[17] Y. BISK, R. ZELLERS, J. GAO, Y. CHOI, ET AL., *Piqa: Reasoning about physical commonsense in natural language*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 7432–7439.

[18] Y. BIZZONI AND S. LAPPIN, *Predicting human metaphor paraphrase judgments with deep neural networks*, in Proceedings of the Workshop on Figurative Language Processing, 2018, pp. 45–55.

[19] A. BOSSELUT, H. RASHKIN, M. SAP, C. MALAVIYA, A. CELIKYILMAZ, AND Y. CHOI, *COMET: Commonsense transformers for automatic knowledge graph construction*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 2019, Association for Computational Linguistics, pp. 4762–4779.

[20] D. A. BRONIATOWSKI, M. J. PAUL, AND M. DREDZE, *National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic*, PloS one, 8 (2013), p. e83672.

[21] C. BROOKER, *This awesome dissection of internet hyperbole will make you cry and change your life | charlie brooker*, Oct 2014.

[22] J. S. BROWNSTEIN AND C. FREIFELD, *Healthmap: the development of automated real-time internet surveillance for epidemic intelligence*, Weekly releases (1997–2007), 12 (2007), p. 3322.

[23] S. BULLO AND J. H. HEARN, *Parallel worlds and personified pain: A mixed-methods analysis of pain metaphor use by women with endometriosis*, British Journal of Health Psychology, 26 (2021), pp. 271–288.

[24] C. BURGERS, B. C. BRUGMAN, K. Y. RENARDEL DE LAVALETTE, AND G. J. STEEN, *Hip: A method for linguistic hyperbole identification in discourse*, Metaphor and Symbol, 31 (2016), pp. 163–178.

[25] K. BUSCHMEIER, P. CIMIANO, AND R. KLINGER, *An impact analysis of features in a classification approach to irony detection in product reviews*, in Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2014, pp. 42–49.

[26] R. CARSTON AND C. WEARING, *Hyperbolic language and its relation to metaphor and irony*, Journal of Pragmatics, 79 (2015), pp. 79–92.

[27] T. CHAKRABARTY, D. GHOSH, S. MURESAN, AND N. PENG, *Rˆ3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 7976–7986.

[28] T. CHAKRABARTY, S. MURESAN, AND N. PENG, *Generating similes like a pro: A style transfer approach for simile generation*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6455–6469.

[29] L. CHEN, K. T. HOSSAIN, P. BUTLER, N. RAMAKRISHNAN, AND B. A. PRAKASH, *Syndromic surveillance of flu on twitter using weakly supervised temporal topic models*, Data mining and knowledge discovery, 30 (2016), pp. 681–710.

[30] S. F. CHEN AND J. GOODMAN, *An empirical study of smoothing techniques for language modeling*, Computer Speech & Language, 13 (1999), pp. 359–394.

[31] J. H. CHO AND B. HARIHARAN, *On the efficacy of knowledge distillation*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4794–4802.

[32] C. CHU AND R. WANG, *A survey of domain adaptation for machine translation*, Journal of information processing, 28 (2020), pp. 413–426.

[33] C. CLARIDGE, *Hyperbole in English: A corpus-based study of exaggeration*, Cambridge University Press, 2010.

[34] N. COLLIER, A. KAWAZOE, L. JIN, M. SHIGEMATSU, D. DIEN, ET AL., *The biocaster ontology: A multilingual ontology for infectious disease outbreak surveillance: Rationale, design and challenges. j lang resources eval 40: 405–413*, 2007.

[35] H. L. COLSTON AND A. N. KATZ, *Figurative language comprehension: Social and cultural influences*, Routledge, 2004.

[36] M. COLTHEART, *The mrc psycholinguistic database*, The Quarterly Journal of Experimental Psychology Section A, 33 (1981), pp. 497–505.

[37] M. CONWAY, J. DOWLING, AND W. CHAPMAN, *Developing an application ontology for mining free text clinical reports: the extended syndromic surveillance ontology*, in 3rd international workshop on health document text mining and information analysis (LOUHI 2011), Citeseer, 2011, pp. 75–82.

[38] G. COPPERSMITH, M. DREDZE, AND C. HARMAN, *Quantifying mental health signals in twitter*, in Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality, 2014, pp. 51–60.

[39] P. CRAMER, *A study of homographs*, in Norms of word association, Elsevier, 1970, pp. 361–382.

[40] R. DALE, *Nlp commercialisation in the last 25 years*, Natural Language Engineering, 25 (2019), pp. 419–426.

[41] M. DE CHOUDHURY, M. GAMON, S. COUNTS, AND E. HORVITZ, *Predicting depression via social media*, in Seventh international AAAI conference on weblogs and social media, 2013.

[42] Z. DEMJÉN AND E. SEMINO, *Using metaphor in healthcare*, The Routledge handbook of metaphor and language, (2016), p. 385.

[43] J. DEVLIN, M.-W. CHANG, K. LEE, AND K. TOUTANOVA, *BERT: Pre-training of deep bidirectional transformers for language understanding*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, Association for Computational Linguistics, pp. 4171–4186.

[44] A. DHAR, H. MUKHERJEE, N. S. DASH, AND K. ROY, *Text categorization: past and present*, Artificial Intelligence Review, 54 (2021), pp. 3007–3054.

[45] J. E. DIAZ-VERA, *Metaphor and metonymy across time and cultures: Perspectives on the sociohistorical linguistics of figurative language*, vol. 52, Walter de Gruyter GmbH & Co KG, 2014.

[46] O. E. DICTIONARY, *Oxford english dictionary*, Retrieved March, 10 (2019).

[47] X. DONG AND J. SHEN, *Triplet loss in siamese network for object tracking*, in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 459–474.

[48] J. DUNN, *Measuring metaphoricity*, in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2014, pp. 745–751.

[49] D. EDWARDS, *Extreme case formulations: Softeners, investment, and doing nonliteral*, Research on language and social interaction, 33 (2000), pp. 347–373.

[50] A. EL ABADDI, L. BACKSTROM, S. CHAKRABARTI, A. JAIMES, J. LESKOVEC, AND A. TOMKINS, *Social media: source of information or bunch of noise*, in Proceedings of the 20th International conference companion on World Wide Web, 2011, pp. 327–328.

[51] Y. ELAZAR, A. MAHABAL, D. RAMACHANDRAN, T. BEDRAX-WEISS, AND D. ROTH, *How large are lions? inducing distributions over quantitative attributes*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 2019, Association for Computational Linguistics, pp. 3973–3983.

[52] A. ERMILOV, N. MURASHKINA, V. GORYACHEVA, AND P. BRASLAVSKI, *Stierlitz meets svm: Humor detection in russian*, in Conference on Artificial Intelligence and Natural Language, Springer, 2018, pp. 178–184.

[53] V. W. FENG AND G. HIRST, *Text-level discourse parsing with rich linguistic features*, in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2012, pp. 60–68.

[54] E. FERSINI, F. A. POZZI, AND E. MESSINA, *Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers*, in 2015 IEEE international conference on data science and advanced analytics (DSAA), IEEE, 2015, pp. 1–8.

[55] M. FORBES, A. HOLTZMAN, AND Y. CHOI, *Do neural language representations learn physical commonsense?*, Proceedings of the 41st Annual Conference of the Cognitive Science Society, (2019).

[56] S. R. FUSSELL AND M. M. MOSS, *Figurative language in emotional communication*, Social and cognitive approaches to interpersonal communication, (1998), pp. 113–141.

[57] M. GAUR, U. KURSUNCU, A. ALAMBO, A. SHETH, R. DANIULAITYTE, K. THIRUNARAYAN, AND J. PATHAK, *Let me tell you about your mental health!: Contextualized classification of reddit posts to dsm-5 for web-based intervention*, in Proceedings of the 27th ACM International Conference on Information and Knowledge Management, ACM, 2018, pp. 753–762.

[58] A. GHOSH, G. LI, T. VEALE, P. ROSSO, E. SHUTOVA, J. BARNDEN, AND A. REYES, *SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter*, in Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, June 2015, Association for Computational Linguistics, pp. 470–478.

[59] A. GHOSH AND T. VEALE, *Fracking sarcasm using neural network*, in Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis, 2016, pp. 161–169.

[60] R. W. GIBBS, *Irony in talk among friends*, Metaphor and symbol, 15 (2000), pp. 5–27.

[61] R. W. GIBBS JR AND H. FRANKS, *Embodied metaphor in women's narratives about their experiences with cancer*, Health communication, 14 (2002), pp. 139–165.

[62] R. GINN, P. PIMPALKHUTE, A. NIKFARJAM, A. PATKI, K. O?CONNOR, A. SARKER, K. SMITH, AND G. GONZALEZ, *Mining twitter for adverse drug reaction mentions: a corpus and classification benchmark*, in Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing, Citeseer, 2014, pp. 1–8.

[63] S. GLUCKSBERG, M. S. MCGLONE, Y. GRODZINSKY, AND K. AMUNTS, *Understanding figurative language: From metaphor to idioms*, Oxford University Press on Demand, 2001.

[64] A. GO, R. BHAYANI, AND L. HUANG, *Twitter sentiment classification using distant supervision*, CS224N Project Report, Stanford, 1 (2009), p. 2009.

[65] J. GOU, B. YU, S. J. MAYBANK, AND D. TAO, *Knowledge distillation: A survey*, International Journal of Computer Vision, 129 (2021), pp. 1789–1819.

[66] A. GRAVES, A.-R. MOHAMED, AND G. HINTON, *Speech recognition with deep recurrent neural networks*, in 2013 IEEE international conference on acoustics, speech and signal processing, IEEE, 2013, pp. 6645–6649.

[67] P. GROUP, *Mip: A method for identifying metaphorically used words in discourse*, Metaphor and symbol, 22 (2007), pp. 1–39.

[68] S. HAN, *Web 2.0*, Routledge, 2012.

[69] Y. HAO AND T. VEALE, *An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes*, Minds and Machines, 20 (2010), pp. 635–650.

[70] K. J. HARRINGTON, *The use of metaphor in discourse about cancer: a review of the literature.*, Clinical journal of oncology nursing, 16 (2012).

[71] D. M. HARTLEY, C. M. GIANNINI, S. WILSON, O. FRIEDER, P. A. MARGOLIS, U. R. KOTAGAL, D. L. WHITE, B. L. CONNELLY, D. S. WHEELER, D. G. TADESSE, ET AL., *Coughing, sneezing, and aching online: Twitter and the volume of influenza-like illness in a pediatric hospital*, PLoS One, 12 (2017), p. e0182008.

[72] A. HERMANS, L. BEYER, AND B. LEIBE, *In defense of the triplet loss for person re-identification*, arXiv preprint arXiv:1703.07737, (2017).

[73] D. HILLS, *Metaphor*, Sep 2016.

[74] L. HJORTH AND S. HINTON, *Understanding social media*, Sage, 2019.

[75] S. HOCHREITER AND J. SCHMIDHUBER, *Long short-term memory*, Neural computation, 9 (1997), pp. 1735–1780.

[76] C. HOMMERBERG, A. W. GUSTAFSSON, AND A. SANDGREN, *Battle, journey, imprisonment and burden: patterns of metaphor use in blogs about living with advanced cancer*, BMC palliative care, 19 (2020), pp. 1–10.

[77] C. HOOTON, *How hyperbole 'won the internet'*, Jan 2015.

[78] N. HOSSAIN, J. KRUMM, AND M. GAMON, *"president vows to cut <taxes> hair": Dataset and analysis of creative text editing for humorous headlines*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, Association for Computational Linguistics, pp. 133–142.

[79] J. HOWARD AND S. RUDER, *Universal language model fine-tuning for text classification*, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, July 2018, Association for Computational Linguistics, pp. 328–339.

[80] Z. HU, Z. YANG, X. LIANG, R. SALAKHUTDINOV, AND E. P. XING, *Toward controlled generation of text*, in International Conference on Machine Learning, PMLR, 2017, pp. 1587–1596.

[81] Y.-H. HUANG, H.-H. HUANG, AND H.-H. CHEN, *Irony detection with attentive recurrent neural networks*, in European Conference on Information Retrieval, Springer, 2017, pp. 534–540.

[82] G. D. IANNETTI, T. V. SALOMONS, M. MOAYEDI, A. MOURAUX, AND K. D. DAVIS, *Beyond metaphor: contrasting mechanisms of social and physical pain*, Trends in cognitive sciences, 17 (2013), pp. 371–378.

[83] A. IYER, A. JOSHI, S. KARIMI, R. SPARKS, AND C. PARIS, *Figurative usage detection of symptom words to improve personal health mention detection*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 2019, Association for Computational Linguistics, pp. 1142–1147.

[84] H. JANG, S. MOON, Y. JO, AND C. ROSE, *Metaphor detection in discourse*, in Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2015, pp. 384–392.

[85] K. JIANG, R. CALIX, AND M. GUPTA, *Construction of a personal experience tweet corpus for health surveillance*, in Proceedings of the 15th workshop on biomedical natural language processing, 2016, pp. 128–135.

[86] K. JIANG, S. FENG, Q. SONG, R. A. CALIX, M. GUPTA, AND G. R. BERNARD, *Identifying tweets of personal health experience through word embedding and lstm neural network*, BMC bioinformatics, 19 (2018), p. 210.

[87] B. JIN, A. JOSHI, R. SPARKS, S. WAN, C. PARIS, AND C. R. MACINTYRE, *'watch the flu': A tweet monitoring tool for epidemic intelligence of influenza in australia*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 13616–13617.

[88] S. JORDAN, S. HOVET, I. FUNG, H. LIANG, K.-W. FU, AND Z. TSE, *Using twitter for public health surveillance from monitoring and prediction to public response*, Data, 4 (2019), p. 6.

[89] A. JOSHI, P. BHATTACHARYYA, AND M. J. CARMAN, *Understanding the phenomenon of sarcasm*, in Investigations in Computational Sarcasm, Springer, 2018, pp. 33–57.

[90] A. JOSHI, S. KARIMI, R. SPARKS, C. PARIS, AND C. R. MACINTYRE, *Survey of text-based epidemic intelligence: A computational linguistics perspective*, ACM Comput. Surv., 52 (2019).

[91] A. JOSHI, V. SHARMA, AND P. BHATTACHARYYA, *Harnessing context incongruity for sarcasm detection*, in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), vol. 2, 2015, pp. 757–762.

[92] A. JOSHI, V. TRIPATHI, P. BHATTACHARYYA, AND M. CARMAN, *Harnessing sequence labeling for sarcasm detection in dialogue from tv series 'friends'*, in Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, 2016, pp. 146–155.

[93] A. JOSHI, V. TRIPATHI, K. PATEL, P. BHATTACHARYYA, AND M. CARMAN, *Are word embedding-based features useful for sarcasm detection?*, in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, Nov. 2016, Association for Computational Linguistics, pp. 1006–1011.

[94] A. JOULIN, E. GRAVE, P. BOJANOWSKI, AND T. MIKOLOV, *Bag of tricks for efficient text classification*, arXiv preprint arXiv:1607.01759, (2016).

[95] D. JURAFSKY AND J. H. MARTIN, *Speech and language processing*, vol. 3, Pearson London, 2014.

[96] S. KANDULA AND J. SHAMAN, *Reappraising the utility of google flu trends*, PLoS computational biology, 15 (2019), p. e1007258.

[97] S. KANOUCHI, M. KOMACHI, N. OKAZAKI, E. ARAMAKI, AND H. ISHIKAWA, *Who caught a cold?-identifying the subject of a symptom*, 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), vol. 1, 2015, pp. 1660–1670.

[98] P. KARISANI AND E. AGICHTEIN, *Did you really just have a heart attack?: Towards robust detection of personal health mentions in social media*, in Proc. WWW '18, 2018, pp. 137–146.

[99] J. KAROUI, F. BENAMARA, V. MORICEAU, N. AUSSENAC-GILLES, AND L. H. BELGUITH, *Towards a contextual pragmatic model to detect irony in tweets*, in 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015), 2015, pp. PP–644.

[100] M. KHOKHLOVA, V. PATTI, AND P. ROSSO, *Distinguishing between irony and sarcasm in social media texts: Linguistic observations*, in 2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT), IEEE, 2016, pp. 1–6.

[101] L. KONG, C. LI, J. GE, B. LUO, AND V. NG, *An empirical study of hyperbole*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7024–7034.

[102] L. KONG, N. SCHNEIDER, S. SWAYAMDIPTA, A. BHATIA, C. DYER, AND N. A. SMITH, *A dependency parser for tweets*, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1001–1012.

[103] R. R. KOPP AND M. J. CRAW, *Metaphoric language, metaphoric cognition, and cognitive therapy.*, Psychotherapy: theory, research, practice, training, 35 (1998), p. 306.

[104] O. KOVALEVA, A. ROMANOV, A. ROGERS, AND A. RUMSHISKY, *Revealing the dark secrets of BERT*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Nov. 2019, Association for Computational Linguistics, pp. 4365–4374.

[105] K. KOWSARI, K. JAFARI MEIMANDI, M. HEIDARYSAFA, S. MENDU, L. BARNES, AND D. BROWN, *Text classification algorithms: A survey*, Information, 10 (2019), p. 150.

[106] R. J. KREUZ AND R. M. ROBERTS, *Figurative language occurrence and co-occurrence in contemporary literature*, Empirical approaches to literature and aesthetics, (1996), pp. 83–97.

[107] K. KRIPPENDORFF, *Computing krippendorff's alpha-reliability*, (2011).

[108] G. LAKOFF AND M. JOHNSON, *Conceptual metaphor in everyday language*, The journal of Philosophy, 77 (1980), pp. 453–486.

[109] A. LAMB, M. J. PAUL, AND M. DREDZE, *Separating fact from fear: Tracking flu infections on twitter*, in Proceedings of the 2013 Conference of the North

American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 789–795.

[110] J. LAMBERT, O. SENER, AND S. SAVARESE, *Deep learning under privileged information using heteroscedastic dropout*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8886–8895.

[111] D. LAZER, R. KENNEDY, G. KING, AND A. VESPIGNANI, *The parable of google flu: traps in big data analysis*, Science, 343 (2014), pp. 1203–1205.

[112] Q. LE AND T. MIKOLOV, *Distributed representations of sentences and documents*, in International conference on machine learning, PMLR, 2014, pp. 1188–1196.

[113] J. LEE, R. TANG, AND J. LIN, *What would elsa do? freezing layers during transformer fine-tuning*, 2019.

[114] J. S. LEGGITT AND R. W. GIBBS, *Emotional reactions to verbal irony*, Discourse processes, 29 (2000), pp. 1–24.

[115] O. LEVY AND Y. GOLDBERG, *Dependency-based word embeddings*, in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2014, pp. 302–308.

[116] M. LEWIS, Y. LIU, N. GOYAL, M. GHAZVININEJAD, A. MOHAMED, O. LEVY, V. STOYANOV, AND L. ZETTLEMOYER, *Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*, arXiv preprint arXiv:1910.13461, (2019).

[117] J. LI, R. JIA, H. HE, AND P. LIANG, *Delete, retrieve, generate: a simple approach to sentiment and style transfer*, in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana, June 2018, Association for Computational Linguistics, pp. 1865–1874.

[118] Q. LI, *Literature survey: domain adaptation algorithms for natural language processing*, Department of Computer Science The Graduate Center, The City University of New York, (2012), pp. 8–10.

[119] C. LIEBRECHT, F. KUNNEMAN, AND A. VAN DEN BOSCH, *The perfect solution for detecting sarcasm in tweets# not*, (2013).

[120] C. LIU AND R. HWA, *Heuristically informed unsupervised idiom usage recognition*, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 1723–1731.

[121] L. LIU, D. ZHANG, AND W. SONG, *Exploiting syntactic structures for humor recognition*, in Proceedings of the 27th international conference on computational linguistics, 2018, pp. 1875–1883.

[122] ——, *Modeling sentiment association in discourse for humor recognition*, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2018, pp. 586–591.

[123] N. F. LIU, M. GARDNER, Y. BELINKOV, M. E. PETERS, AND N. A. SMITH, *Linguistic knowledge and transferability of contextual representations*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, Association for Computational Linguistics, pp. 1073–1094.

[124] Y. LIU, M. OTT, N. GOYAL, J. DU, M. JOSHI, D. CHEN, O. LEVY, M. LEWIS, L. ZETTLEMOYER, AND V. STOYANOV, *Roberta: A robustly optimized bert pretraining approach*, arXiv preprint arXiv:1907.11692, (2019).

[125] Z. LIU, Y. WANG, J. KASAI, H. HAJISHIRZI, AND N. A. SMITH, *Probing across time: What does RoBERTa know and when?*, in Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, Nov. 2021, Association for Computational Linguistics, pp. 820–842.

[126] C.-K. LO, *Yisi-a unified semantic mt quality evaluation and estimation metric for languages with different levels of available resources*, in Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), 2019, pp. 507–513.

[127] J. LUCARIELLO, *Situational irony: A concept of events gone away*, Irony in language and thought, (2007), pp. 467–498.

[128] F. LUO, D. DAI, P. YANG, T. LIU, B. CHANG, Z. SUI, AND X. SUN, *Learning to control the fine-grained sentiment for story ending generation*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 6020–6026.

[129] D. MAGAÑA AND T. MATLOCK, *How spanish speakers use metaphor to describe their experiences with cancer*, Discourse & Communication, 12 (2018), pp. 627–644.

[130] C. MANNING, P. RAGHAVAN, AND H. SCHÜTZE, *Introduction to information retrieval*, Natural Language Engineering, 16 (2010), pp. 100–103.

[131] C. MANNING AND H. SCHUTZE, *Foundations of statistical natural language processing*, MIT press, 1999.

[132] N. MATHUR, T. BALDWIN, AND T. COHN, *Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 4984–4997.

[133] M. MCCARTHY AND R. CARTER, *"there's millions of them": hyperbole in everyday conversation*, Journal of pragmatics, 36 (2004), pp. 149–184.

[134] W. MEDHAT, A. HASSAN, AND H. KORASHY, *Sentiment analysis algorithms and applications: A survey*, Ain Shams engineering journal, 5 (2014), pp. 1093–1113.

[135] O. MELAMUD, J. GOLDBERGER, AND I. DAGAN, *context2vec: Learning generic context embedding with bidirectional lstm*, in Proceedings of the 20th SIGNLL conference on computational natural language learning, 2016, pp. 51–61.

[136] L. MENG, L. CHEN, X. YANG, D. TAO, H. ZHANG, C. MIAO, AND T.-S. CHUA, *Learning using privileged information for food recognition*, in Proceedings of the 27th ACM International Conference on Multimedia, MM '19, New York, NY, USA, 2019, Association for Computing Machinery, pp. 557–565.

[137] P. MICHEL, O. LEVY, AND G. NEUBIG, *Are sixteen heads really better than one?*, in Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds., vol. 32, Curran Associates, Inc., 2019.

[138] T. MIKOLOV, I. SUTSKEVER, K. CHEN, G. S. CORRADO, AND J. DEAN, *Distributed representations of words and phrases and their compositionality*, in Advances in neural information processing systems, 2013, pp. 3111–3119.

[139] G. A. MILLER, *Wordnet: a lexical database for english*, Communications of the ACM, 38 (1995), pp. 39–41.

[140] J. S. MIO AND A. N. KATZ, *Metaphor: Implications and applications*, Psychology Press, 2018.

[141] A. MISHRA, D. KANOJIA, S. NAGAR, K. DEY, AND P. BHATTACHARYYA, *Harnessing cognitive features for sarcasm detection*, in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, Aug. 2016, Association for Computational Linguistics, pp. 1095–1104.

[142] A. MISHRA, T. TATER, AND K. SANKARANARAYANAN, *A modular architecture for unsupervised sarcasm generation*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Nov. 2019, Association for Computational Linguistics, pp. 6144–6154.

[143] S. MOHAMMAD, *Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words*, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, July 2018, Association for Computational Linguistics, pp. 174–184.

[144] S. M. MOHAMMAD, *Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words*, in Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL), Melbourne, Australia, 2018.

[145] L. C. MORA, *All or nothing: A semantic analysis of hyperbole*, Revista de Lingüística y lenguas Aplicadas, 4 (2009), pp. 25–35.

[146] I. MUNDAY, T. NEWTON-JOHN, AND I. KNEEBONE, *'barbed wire wrapped around my feet': Metaphor use in chronic pain*, British journal of health psychology, 25 (2020), pp. 814–830.

[147] S. NEILSON, *Pain as metaphor: metaphor and medicine*, Medical Humanities, 42 (2016), pp. 3–10.

[148] N. NG, K. YEE, A. BAEVSKI, M. OTT, M. AULI, AND S. EDUNOV, *Facebook fair's wmt19 news translation task submission*, in Proceedings of the Fourth

Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), 2019, pp. 314–319.

[149] D. Q. NGUYEN, T. VU, AND A. TUAN NGUYEN, *BERTweet: A pre-trained language model for English tweets*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, Oct. 2020, Association for Computational Linguistics, pp. 9–14.

[150] N. R. NORRICK, *Hyperbole, extreme case formulation*, Journal of Pragmatics, 36 (2004), pp. 1727–1739.

[151] K. O'CONNOR, P. PIMPALKHUTE, A. NIKFARJAM, R. GINN, K. L. SMITH, AND G. GONZALEZ, *Pharmacovigilance on twitter? mining tweets for adverse drug reactions*, in AMIA annual symposium proceedings, vol. 2014, American Medical Informatics Association, 2014, p. 924.

[152] B. OFOGHI, M. MANN, AND K. VERSPOOR, *Towards early discovery of salient health threats: A social media emotion classification technique*, in Biocomputing 2016: Proceedings of the Pacific Symposium, World Scientific, 2016, pp. 504–515.

[153] S. OPREA, S. WILSON, AND W. MAGDY, *Chandler: An explainable sarcastic response generator*, in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2021, pp. 339–349.

[154] T. O'REILLY, *What is web 2.0*, " O'Reilly Media, Inc.", 2009.

[155] W. H. ORGANIZATION, *Public health surveillance*, Sep 2017.

[156] M. OTT, S. EDUNOV, A. BAEVSKI, A. FAN, S. GROSS, N. NG, D. GRANGIER, AND M. AULI, *fairseq: A fast, extensible toolkit for sequence modeling*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Minneapolis, Minnesota, June 2019, Association for Computational Linguistics, pp. 48–53.

[157] K. PAPINENI, S. ROUKOS, T. WARD, AND W.-J. ZHU, *Bleu: a method for automatic evaluation of machine translation*, in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[158] C. PAQUET, D. COULOMBIER, R. KAISER, AND M. CIOTTI, *Epidemic intelligence: a new framework for strengthening disease surveillance in europe*, Eurosurveillance, 11 (2006), pp. 5–6.

[159] N. PARDE AND R. NIELSEN, *Detecting sarcasm is extremely easy;-*, in Proceedings of the workshop on computational semantics beyond events and roles, 2018, pp. 21–26.

[160] M. J. PAUL AND M. DREDZE, *A model for mining public health topics from twitter*, Health, 11 (2012), p. 1.

[161] M. J. PAUL, A. SARKER, J. S. BROWNSTEIN, A. NIKFARJAM, M. SCOTCH, K. L. SMITH, AND G. GONZALEZ, *Social media mining for public health monitoring and surveillance*, in Biocomputing 2016: Proceedings of the Pacific symposium, World Scientific, 2016, pp. 468–479.

[162] D. PECHYONY AND V. VAPNIK, *On the theory of learnining with privileged information*, Advances in neural information processing systems, 23 (2010), pp. 1894–1902.

[163] P. PEDINOTTI, E. DI PALMA, L. CERINI, AND A. LENCI, *A howling success or a working sea? testing what bert knows about metaphors*, in Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, 2021, pp. 192–204.

[164] L. PELED AND R. REICHART, *Sarcasm SIGN: Interpreting sarcasm with sentiment based monolingual machine translation*, in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, July 2017, Association for Computational Linguistics, pp. 1690–1700.

[165] M. S. PENA, F. RUIZ DE MENDOZA, AND A. ATHANASIADOU, *Construing and constructing hyperbole*, Studies in Figurative Thought and Language, 56 (2017), p. 41.

[166] J. PENNINGTON, R. SOCHER, AND C. MANNING, *Glove: Global vectors for word representation*, in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[167] P. PÉREZ-SOBRINO AND J. LITTLEMORE, *What makes an advert go viral?: The role of figurative operations in the success of internet videos*, in Performing Metaphorical Creativity across Modes and Contexts, John Benjamins Publishing Company, 2020, pp. 119–152.

[168] M. E. PETERS, M. NEUMANN, M. IYYER, M. GARDNER, C. CLARK, K. LEE, AND L. ZETTLEMOYER, *Deep contextualized word representations*, in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana, June 2018, Association for Computational Linguistics, pp. 2227–2237.

[169] M. E. PETERS, M. NEUMANN, L. ZETTLEMOYER, AND W.-T. YIH, *Dissecting contextual word embeddings: Architecture and representation*, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, Oct.-Nov. 2018, Association for Computational Linguistics, pp. 1499–1509.

[170] M. E. PETERS, S. RUDER, AND N. A. SMITH, *To tune or not to tune? adapting pretrained representations to diverse tasks*, in Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), Florence, Italy, Aug. 2019, Association for Computational Linguistics, pp. 7–14.

[171] P. PEXMAN, H. COLSTON, AND A. KATZ, *Figurative language comprehension: Social and cultural influences*, (2005).

[172] M. PHUONG AND C. LAMPERT, *Towards understanding knowledge distillation*, in International Conference on Machine Learning, PMLR, 2019, pp. 5142–5151.

[173] H. R. POLLIO AND J. M. BARLOW, *A behavioural analysis of figurative language in psychotherapy: One session in a single case-study*, Language and Speech, 18 (1975), pp. 236–254.

[174] A. POMERANTZ, *Extreme case formulations: A way of legitimizing claims*, Human studies, 9 (1986), pp. 219–229.

[175] S. PORIA, E. CAMBRIA, D. HAZARIKA, AND P. VIJ, *A deeper look into sarcastic tweets using deep convolutional neural networks*, arXiv preprint arXiv:1610.08815, (2016).

[176] M. POST, *A call for clarity in reporting BLEU scores*, in Proceedings of the Third Conference on Machine Translation: Research Papers, Brussels, Belgium, Oct. 2018, Association for Computational Linguistics, pp. 186–191.

[177] M. PRAMANICK, A. GUPTA, AND P. MITRA, *An lstm-crf based approach to token-level metaphor detection*, in Proceedings of the Workshop on Figurative Language Processing, 2018, pp. 67–75.

[178] R. PRYZANT, R. D. MARTINEZ, N. DASS, S. KUROHASHI, D. JURAFSKY, AND D. YANG, *Automatically neutralizing subjective bias in text*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 480–489.

[179] A. QADIR, E. RILOFF, AND M. WALKER, *Automatically inferring implicit properties in similes*, in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1223–1232.

[180] S. RAI AND S. CHAKRAVERTY, *A survey on computational metaphor processing*, ACM Computing Surveys (CSUR), 53 (2020), pp. 1–37.

[181] N. F. RAJANI, B. MCCANN, C. XIONG, AND R. SOCHER, *Explain yourself! leveraging language models for commonsense reasoning*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 2019, Association for Computational Linguistics, pp. 4932–4942.

[182] A. RAMPONI AND B. PLANK, *Neural unsupervised domain adaptation in nlp—a survey*, arXiv preprint arXiv:2006.00632, (2020).

[183] S. RAO AND J. TETREAULT, *Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer*, in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana, June 2018, Association for Computational Linguistics, pp. 129–140.

[184] K. RAVI AND V. RAVI, *Irony detection using neural network language model, psycholinguistic features and text mining*, in 2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), IEEE, 2018, pp. 254–260.

[185] H. E. RECCHIA, N. HOWE, H. S. ROSS, AND S. ALEXANDER, *Children's understanding and production of verbal irony in family conversations*, British Journal of Developmental Psychology, 28 (2010), pp. 255–274.

[186] P. RESNIK, *Selectional preference and sense disambiguation*, in Tagging Text with Lexical Semantics: Why, What, and How?, 1997.

[187] A. REYES AND P. ROSSO, *On the difficulty of automatically detecting irony: beyond a simple case of negation*, Knowledge and Information Systems, 40 (2014), pp. 595–614.

[188] A. REYES, P. ROSSO, AND D. BUSCALDI, *From humor recognition to irony detection: The figurative language of social media*, Data & Knowledge Engineering, 74 (2012), pp. 1–12.

[189] A. REYES, P. ROSSO, AND T. VEALE, *A multidimensional approach for detecting irony in twitter*, Language resources and evaluation, 47 (2013), pp. 239–268.

[190] M. T. RIBEIRO, S. SINGH, AND C. 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, 2016, pp. 1135–1144.

[191] M. T. RIBEIRO, T. WU, C. GUESTRIN, AND S. SINGH, *Beyond accuracy: Behavioral testing of NLP models with CheckList*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020, Association for Computational Linguistics, pp. 4902–4912.

[192] E. RILOFF, A. QADIR, P. SURVE, L. DE SILVA, N. GILBERT, AND R. HUANG, *Sarcasm as contrast between a positive sentiment and negative situation*, in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 704–714.

[193] J. R. RITTER, *Recovering hyperbole: Rethinking the limits of rhetoric for an age of excess*, Philosophy & rhetoric, 45 (2012), pp. 406–428.

[194] A. ROGERS, O. KOVALEVA, AND A. RUMSHISKY, *A primer in bertology: What we know about how bert works*, Transactions of the Association for Computational Linguistics, 8 (2020), pp. 842–866.

[195] A. SARKER, R. GINN, A. NIKFARJAM, K. O'CONNOR, K. SMITH, S. JAYARAMAN, T. UPADHAYA, AND G. GONZALEZ, *Utilizing social media data for pharmacovigilance: a review*, Journal of biomedical informatics, 54 (2015), pp. 202–212.

[196] F. SCHROFF, D. KALENICHENKO, AND J. PHILBIN, *Facenet: A unified embedding for face recognition and clustering*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

[197] E. SEMINO, *Descriptions of pain, metaphor, and embodied simulation*, Metaphor and Symbol, 25 (2010), pp. 205–226.

[198] E. SEMINO AND Z. DEMJÉN, *The Routledge handbook of metaphor and language*, Taylor & Francis, 2016.

[199] E. SEMINO, Z. DEMJÉN, AND J. DEMMEN, *An integrated approach to metaphor and framing in cognition, discourse, and practice, with an application to metaphors for cancer*, Applied linguistics, 39 (2018), pp. 625–645.

[200] E. SEMINO, Z. DEMJÉN, J. DEMMEN, V. KOLLER, S. PAYNE, A. HARDIE, AND P. RAYSON, *The online use of violence and journey metaphors by patients with cancer, as compared with health professionals: a mixed methods study*, BMJ supportive & palliative care, 7 (2017), pp. 60–66.

[201] E. SEMINO, Z. DEMJÉN, A. HARDIE, S. PAYNE, AND P. RAYSON, *Metaphor, cancer and the end of life: A corpus-based study*, Routledge, 2017.

[202] Y. SHU, Q. LI, S. LIU, AND G. XU, *Learning with privileged information for photo aesthetic assessment*, Neurocomputing, 404 (2020), pp. 304–316.

[203] E. SHUTOVA, *Automatic metaphor interpretation as a paraphrasing task*, in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2010, pp. 1029–1037.

[204] ——, *Design and evaluation of metaphor processing systems*, Computational Linguistics, 41 (2015), pp. 579–623.

[205] E. SHUTOVA, L. SUN, AND A. KORHONEN, *Metaphor identification using verb and noun clustering*, in Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), 2010, pp. 1002–1010.

[206] E. SHUTOVA, T. VAN DE CRUYS, AND A. KORHONEN, *Unsupervised metaphor paraphrasing using a vector space model*, in Proceedings of COLING 2012: Posters, 2012, pp. 1121–1130.

[207] C. SILVERMAN, *Lies, damn lies and viral content*, (2015).

[208] P. A. SIMS, *Working with metaphor*, American journal of psychotherapy, 57 (2003), pp. 528–536.

[209] L. SINNENBERG, A. M. BUTTENHEIM, K. PADREZ, C. MANCHENO, L. UNGAR, AND R. M. MERCHANT, *Twitter as a tool for health research: a systematic review*, American journal of public health, 107 (2017), pp. e1–e8.

[210] M. SNOVER, B. DORR, R. SCHWARTZ, L. MICCIULLA, AND J. MAKHOUL, *A study of translation edit rate with targeted human annotation*, in Proceedings of association for machine translation in the Americas, vol. 200, Cambridge, MA, 2006.

[211] W. SONG, S. ZHOU, R. FU, T. LIU, AND L. LIU, *Verb metaphor detection via contextual relation learning*, in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 4240–4251.

[212] D. SPERBER AND D. WILSON, *A deflationary account of metaphors*, The Cambridge handbook of metaphor and thought, 84 (2008), p. 105.

[213] H. SRIVASTAVA, V. VARSHNEY, S. KUMARI, AND S. SRIVASTAVA, *A novel hierarchical bert architecture for sarcasm detection*, in Proceedings of the Second Workshop on Figurative Language Processing, 2020, pp. 93–97.

[214] C. STRAPPARAVA, A. VALITUTTI, ET AL., *Wordnet affect: an affective extension of wordnet.*, in LREC, vol. 4, Lisbon, 2004, p. 40.

[215] C. SUN, L. HUANG, AND X. QIU, *Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, Association for Computational Linguistics, pp. 380–385.

[216] M. TABOADA, J. BROOKE, M. TOFILOSKI, K. VOLL, AND M. STEDE, *Lexicon-based methods for sentiment analysis*, Computational linguistics, 37 (2011), pp. 267–307.

[217] A. TALMOR, J. HERZIG, N. LOURIE, AND J. BERANT, *CommonsenseQA: A question answering challenge targeting commonsense knowledge*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, Association for Computational Linguistics, pp. 4149–4158.

[218] D. TANG, B. QIN, AND T. LIU, *Document modeling with gated recurrent neural network for sentiment classification*, in Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 1422–1432.

[219] Y. R. TAUSCZIK AND J. W. PENNEBAKER, *The psychological meaning of words: Liwc and computerized text analysis methods*, Journal of language and social psychology, 29 (2010), pp. 24–54.

[220] D. TAY, *Using metaphor in healthcare: Mental health*, in The Routledge Handbook of Metaphor and Language, Routledge, 2016, pp. 389–402.

[221] A. TERAI AND M. NAKAGAWA, *A computational system of metaphor generation with evaluation mechanism*, in International Conference on Artificial Neural Networks, Springer, 2010, pp. 142–147.

[222] E. TROIANO, C. STRAPPARAVA, G. ÖZBAL, AND S. S. TEKIROĞLU, *A computational exploration of exaggeration*, in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3296–3304.

[223] S. TROTT AND B. BERGEN, *Why do human languages have homophones?*, Cognition, 205 (2020), p. 104449.

[224] P. TUNGTHAMTHITI, K. SHIRAI, AND M. MOHD, *Recognition of sarcasms in tweets based on concept level sentiment analysis and supervised learning approaches*, in Proceedings of the 28th Pacific Asia conference on language, information and computing, 2014, pp. 404–413.

[225] D. VALCARCE, J. PARAPAR, AND Á. BARREIRO, *Additive smoothing for relevance-based language modelling of recommender systems*, in Proceedings of the 4th Spanish Conference on Information Retrieval, 2016, pp. 1–8.

[226] S. VAN DEN BEUKEL AND L. AROYO, *Homonym detection for humor recognition in short text*, in Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2018, pp. 286–291.

[227] C. VAN HEE, E. LEFEVER, AND V. HOSTE, *Semeval-2018 task 3: Irony detection in english tweets*, in Proceedings of The 12th International Workshop on Semantic Evaluation, 2018, pp. 39–50.

[228] A. VASWANI, N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, Ł. KAISER, AND I. POLOSUKHIN, *Attention is all you need*, in Advances in neural information processing systems, 2017, pp. 5998–6008.

[229] T. VEALE, *Round up the usual suspects: Knowledge-based metaphor generation*, in Proceedings of the Fourth Workshop on Metaphor in NLP, 2016, pp. 34–41.

[230] E. VELASCO, T. AGHENEZA, K. DENECKE, G. KIRCHNER, AND T. ECKMANNS, *Social media and internet-based data in global systems for public health surveillance: A systematic review*, The Milbank Quarterly, 92 (2014), pp. 7–33.

[231] S. WAKAMIYA, Y. KAWAI, AND E. ARAMAKI, *Twitter-based influenza detection after flu peak via tweets with indirect information: text mining study*, JMIR public health and surveillance, 4 (2018), p. e65.

[232] M. A. WALKER, J. E. F. TREE, P. ANAND, R. ABBOTT, AND J. KING, *A corpus for research on deliberation and debate.*, in LREC, vol. 12, Istanbul, 2012, pp. 812–817.

[233] B. C. WALLACE, L. KERTZ, E. CHARNIAK, ET AL., *Humans require context to infer ironic intent (so computers probably do, too)*, in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), vol. 2, 2014, pp. 512–516.

[234] A. B. WARRINER, V. KUPERMAN, AND M. BRYSBAERT, *Norms of valence, arousal, and dominance for 13,915 english lemmas*, Behavior research methods, 45 (2013), pp. 1191–1207.

[235] L. WEITZEL, R. C. PRATI, AND R. F. AGUIAR, *The comprehension of figurative language: what is the influence of irony and sarcasm on nlp techniques?*, in Sentiment Analysis and Ontology Engineering, Springer, 2016, pp. 49–74.

[236] K. A. WHITEHEAD, *Extreme-case formulations*, The international encyclopedia of language and social interaction, (2015), pp. 1–5.

[237] J. WIETING, T. BERG-KIRKPATRICK, K. GIMPEL, AND G. NEUBIG, *Beyond BLEU:training neural machine translation with semantic similarity*, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 2019, Association for Computational Linguistics, pp. 4344–4355.

[238] D. WILSON AND R. CARSTON, *Metaphor, relevance and the 'emergent property' issue*, Mind & Language, 21 (2006), pp. 404–433.

[239] M. WILSON, *Mrc psycholinguistic database: Machine-usable dictionary, version 2.00*, Behavior research methods, instruments, & computers, 20 (1988), pp. 6–10.

[240] I. WOOD, J. P. MCCRAE, V. ANDRYUSHECHKIN, AND P. BUITELAAR, *A comparison of emotion annotation schemes and a new annotated data set*, in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 2018, European Language Resources Association (ELRA).

[241] C. WU, F. WU, Y. CHEN, S. WU, Z. YUAN, AND Y. HUANG, *Neural metaphor detecting with cnn-lstm model*, in Proceedings of the Workshop on Figurative Language Processing, 2018, pp. 110–114.

[242] C. WU, F. WU, S. WU, J. LIU, Z. YUAN, AND Y. HUANG, *Thu_ngn at semeval-2018 task 3: Tweet irony detection with densely connected lstm and multi-task learning*, in Proceedings of The 12th International Workshop on Semantic Evaluation, 2018, pp. 51–56.

[243] C.-Y. WU, R. MANMATHA, A. J. SMOLA, AND P. KRAHENBUHL, *Sampling matters in deep embedding learning*, in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2840–2848.

[244] W. XU, A. RITTER, B. DOLAN, R. GRISHMAN, AND C. CHERRY, *Paraphrasing for style*, in Proceedings of COLING 2012, Mumbai, India, Dec. 2012, The COLING 2012 Organizing Committee, pp. 2899–2914.

[245] A. YADOLLAHI, A. G. SHAHRAKI, AND O. R. ZAIANE, *Current state of text sentiment analysis from opinion to emotion mining*, ACM Computing Surveys (CSUR), 50 (2017), pp. 1–33.

[246] D. YANG, A. LAVIE, C. DYER, AND E. HOVY, *Humor recognition and humor anchor extraction*, in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 2367–2376.

[247] B. YU, T. LIU, M. GONG, C. DING, AND D. TAO, *Correcting the triplet selection bias for triplet loss*, in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 71–87.

[248] S. YU, J. SU, AND D. LUO, *Improving bert-based text classification with auxiliary sentence and domain knowledge*, IEEE Access, 7 (2019), pp. 176600–176612.

[249] M. ZAPPAVIGNA, *Discourse of Twitter and social media: How we use language to create affiliation on the web*, vol. 6, A&C Black, 2012.

[250] J. ZHANG, Z. CUI, X. XIA, Y. GUO, Y. LI, C. WEI, AND J. CUI, *Writing polishment with simile: Task, dataset and a neural approach*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 14383–14392.

[251] L. ZHANG, S. WANG, AND B. LIU, *Deep learning for sentiment analysis: A survey*, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8 (2018), p. e1253.

[252] R. ZHANG, J. GUO, Y. FAN, Y. LAN, J. XU, AND X. CHENG, *Learning to control the specificity in neural response generation*, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1108–1117.

[253] X. ZHOU, R. GURURAJAN, Y. LI, R. VENKATARAMAN, X. TAO, G. BARGSHADY, P. D. BARUA, AND S. KONDALSAMY-CHENNAKESAVAN, *A survey on text classification and its applications.*, Web Intelligence (2405-6456), (2020), pp. 1 – 12.