# EXPRESSIONS OF PSYCHOLOGICAL STRESS ON TWITTER: DETECTION AND CHARACTERISATION

RESHMI GOPALAKRISHNA PILLAI

A thesis submitted in partial fulfilment of the requirements of the University of Wolverhampton for the degree of Doctor of Philosophy

This work or any part thereof has not previously been presented in any form to the University or to any other body whether for the purposes of assessment, publication or for any other purpose (unless otherwise indicated). Save for any express acknowledgments, references and/or bibliographies cited in the work, I confirm that the intellectual content of the work is the result of my own efforts and of no other person.



Signature

Date 22.11.2021

LIST OF PUBLICATIONS

Reshmi Gopalakrishna Pillai, Mike Thelwall, Constantin Orasan, Temporal Orientation of High-Stress Tweets, 2019 Proceedings of the 20th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2019, La Rochelle, France.

Reshmi Gopalakrishna Pillai, Mike Thelwall, Constantin Orasan, What Makes You Stressed? Finding Reasons From Tweets, 2018 Proceedings of the Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), EMNLP 2018, Brussels, Belgium.

Reshmi Gopalakrishna Pillai, Mike Thelwall, Constantin Orasan, Trouble on the Road: Finding Reasons for Commuter Stress from Tweets, 2018 Workshop on Intelligent Interactive Systems and Language Generation (2IS&NLG), INLG 2018, Tilburg, Netherlands.

Reshmi Gopalakrishna Pillai, Mike Thelwall, Constantin Orasan, Detection of Stress and Relaxation Magnitudes for Tweets, 2018 Proceedings of the SocialNLP Workshop,The Web Conference WWW'18, Lyon, France. The 2018 Web Conference Companion.

# Abstract

Long-term psychological stress is a significant predictive factor for individual mental health and short-term stress is a useful indicator of an immediate problem. Traditional psychology studies have relied on surveys to understand reasons for stress in general and in specific contexts. The popularity and ubiquity of social media make it a potential data source for identifying and characterising aspects of stress. Previous studies of stress in social media have focused on users responding to stressful personal life events. Prior social media research has not explored expressions of stress in other important domains, however, including travel and politics.

This thesis detects and analyses expressions of psychological stress in social media. So far, TensiStrength is the only existing lexicon for stress and relaxation scores in social media. Using a word-vector based word sense disambiguation method, the TensiStrength lexicon was modified to include the stress scores of the different senses of the same word. On a dataset of 1000 tweets containing ambiguous stress-related words, the accuracy of the modified TensiStrength increased by 4.3%.

This thesis also finds and reports characteristics of a multiple-domain stress dataset of 12000 tweets, 3000 each for airlines, personal events, UK politics, and London traffic. A two-step method for identifying stressors in tweets was implemented. The first step used LDA topic modelling and k-means clustering to find a set of types of stressors (e.g., delay, accident). Second, three word-vector based methods - maximum-word similarity, context-vector similarity, and cluster-vector similarity - were used to detect the stressors in each tweet. The cluster vector similarity method was found to identify the stressors in tweets in all four domains better than machine learning classifiers, based on the performance metrics of accuracy, precision, recall, and f-measure.

Swearing and sarcasm were also analysed in high-stress and no-stress datasets from the four domains using a Convolutional Neural Network and Multilayer Perceptron, respectively. The presence of swearing and sarcasm was higher in the high-stress tweets compared to no-stress tweets in all the domains. The stressors in each domain with higher percentages of swearing or sarcasm were identified. Furthermore, the distribution of the temporal classes (past, present, future, and atemporal) in high-stress tweets was found using an ensemble classifier. The distribution depended on the domain and the stressors.

This study contributes a modified and improved lexicon for the identification of stress scores in social media texts. The two-step method to identify stressors follows a general framework that can be used for domains other than those which were studied. The presence of swearing, sarcasm, and the temporal classes of high-stress tweets belonging to different domains are found and compared to the findings from traditional psychology, for the first time. The algorithms and knowledge may be useful for travel, political, and personal life systems that need to identify stressful events in order to take appropriate action.

# Contents

# List of Tables

# List of Figures

# Acknowledgements

I would like to thank many people whose generous support and guidance made this Ph.D. thesis possible. Firstly, my Director of Studies, Professor Mike Thelwall, who immensely helped me throughout the last five years, as a great supervisor and mentor. He was very approachable and always available to discuss even the finest details about the research. I cannot thank him enough for the courage and confidence he instilled in me through the course of my Ph.D. Professor Constantin Orasan was my second supervisor; I have to thank him for the enormous effort and time he spent on guiding my research and the many productive suggestions he gave me. Together they made a great supervisory team and helped me shape up the research into its final form. Many thanks to the research teams of SCRG and RGCL at the University of Wolverhampton, for their valuable inputs and ideas.

The last few years were a time for many moves and changes too and my family and friends were always there to support me in the process. My father whose consistent encouragement ensured I never lost track of the research goals – my doctoral study was his ambition as much as mine; my mother who always dreamed big for me the way only mothers can – it's been 9 years since we lost her, but her memories will never cease to inspire me; my husband Kiran and our little ones- Harith and Adwaith - to whom I could always turn at the end of long days; my brothers Bijuettan, Sajiettan and sister-in-law Anju – thanks for being my strength, my pride; my friends- Asher, who made me rediscover my potential whenever I was lost; Deepa whose love, trust and patience was a great driving force for me all through the years we have known each other; Pai and Shan who were my lanterns many years back in the proverbial tunnel which seemed to go on without a light in sight, but which eventually led me to Wolverhampton and my doctoral study; Preethy, Sampath and Rony, for the wonderful memories of Nuneaton and beyond. Thank you.

In 2019, I moved to the Netherlands just before the pandemic hit us all; the days of masks, quarantines and lockdowns wouldn't have been bearable without the lovely friends I made here – Anjana, Subi, Andrea, Catalin, Priyanka and Dipankar – thank you guys for your love and support. My colleagues and students at the University of Amsterdam – I thank you all; especially Dr. Frank Nack whose constant motivation kept me on track during the last leg of the doctoral journey.

There are still a lot of people I have not mentioned; this note is too small a space to express my gratitude to the many loving people and great colleagues who helped me throughout the last four years. Please know you are in my heart.

This research was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 636160-2, the Optimum project (www.optimumproject.eu).

# CHAPTER 1

# INTRODUCTION

Psychological stress as a response to an external stimulus is often detrimental to mental and physical health. Prolonged exposure to stressful scenarios could be a precursor to various physiological and psychological ailments. Like other mental states and traits, stress can be expressed directly or indirectly in language. Accurately detecting stress from the social web may help to identify stressed people so that remedial action can be taken if relevant, and information about stressful situations may help planners to detect and deal with emerging stressors, such as traffic jams. The popularity and ubiquity of social media make it a suitable source for detecting stress, compared to traditional methods such as questionnaires and surveys. Furthermore, inputs from social media can be analyzed in a timely manner, often resulting in the proactive identification of mental health red flags, whereas responses to questionnaires/surveys can be delayed and be more of a reactive nature. In addition, tweets sent during a stressful event may capture finer-grained information than responses to a survey about the event sent a few days or weeks afterwards.

This thesis focuses on detecting psychological stress from social media posts. In particular, a method to find stressors from collections of tweets is introduced. The presence of swearing and sarcasm and the distribution of temporal classes in high-stress tweets are also studied. A word-sense disambiguation method to improve the performance of an existing lexicon-based method to measure stress in tweets is evaluated.

## 1.1 Background

### 1.1.1 Social Media Language

Psycholinguistics is the study of cognitive processes and mental capacities which help us to acquire and produce language (Schmitt, 2002). Since language is fundamentally a mode of self-expression, it is natural that the way an individual uses linguistic elements reflects their state of mind. Even though the constructs and elements of a language are similar for all users, the pragmatics differ greatly between individuals. Language holds powerful clues about the speaker or writer. Intuitively, this variance might be more visible in media with a high scope for individual freedom in format and content. Freedom of expression is especially possible in social media websites, where there is usually little monitoring of the content. Social websites sometimes have content guidelines for users, but, as shown in the policy description from Twitter, Figure 1.1, they tend to be directive rather than restrictive.

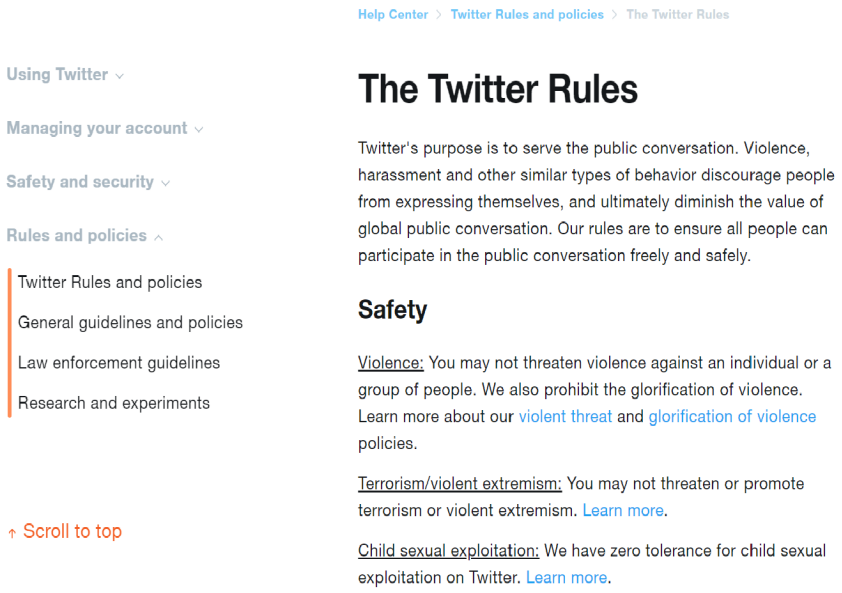


**Figure 1.1 Twitter's rules for the content of tweets (https://help.twitter.com/en/rules-and-policies/twitter-rules)**

The high volume and free nature of posts make it very difficult for social media websites to actively monitor content and it is largely left to the users to report inappropriate tweets. With this lack of screening or proof-reading (Mrva-Montoya, 2012) the language in social media has significant scope for detecting or inferring states of mind.

### 1.1.2 Psychological Stress and Stress-prone Domains

The seminal work, 'The Stress of Life' (Selye, 1956) describes stress as the non-specific response of the body to any demand for change. This definition puts forward two distinct but related aspects of stress: The stimulus (i.e., the demand for a change) and the response to it. The latter can manifest itself in different forms, mainly as variations in physiological parameters. Stress has also been described as a "condition characterized by a perceived discrepancy between information about a monitored variable and criteria for eliciting patterned effector responses" (Goldstein & Kopin, 2007). These studies focus on the physiological manifestations of stress responses such as an increase in heartbeat rate, the release of the stress hormone cortisol into the bloodstream, or pupil dilation. At the same time, the psychological effects of stress have been the focus of other research studies. It has been widely agreed that stress can lead to various psychological disorders such as depression, post-traumatic stress disorder, and postpartum depression (Cohen, Kessler & Gordon, 1995; Cohen, Janicki-Deverts & Miller, 2007). This motivates the assessment of stress from a psychological health perspective.

The concept of stressors is pivotal to the study of physiological or psychological stress. If stress is the response, stressor is the stimulus inducing the response. It is the perceived/actual threat to an organism (Selye, 1956).

Stressors vary greatly between contexts and it is useful to narrow the scope of stress detection to a small number of topics that can be analysed in detail. This thesis focuses on stress related to personal events, travel, and politics, which are briefly introduced in the next sections.

The choice of these domains stems from the existing research in traditional psychology about the psychological stress due to or about them. Travel (specifically road traffic and airlines), politics, and personal events have been found to induce stress, as established by independent researches (Sections 1.1.2.1 , 1.1.2.2 , 1.1.2.3). Further, in the stress measurement experiments conducted in the initial stage of our research also pointed to the relatively high mean score of the stress of tweets belonging to these domains (Table 4.2). Intuitively, work-life could have been another domain with potential expressions of stress. However, the presence of tweets belonging to work-life was very low in the dataset we collected initially (Chapter 3). This could be because of the apprehensions of potential professional damage from sharing work-life induced stress on a public platform like Twitter.

Traditional psychology studies pointed to distinct stressors in surveys conducted in different domains. This motivated us to collect the datasets with keywords specific to these domains. The following domains were chosen on the background of the existing psychology studies on stress induced by activities related to them.

#### 1.1.2.1 Traffic and airlines

Travel can be a stressful activity. Stressors related to travel rarely last long, however. If such stressors occur regularly (traffic congestion on the way to work, fear of flights in frequent travelers, etc.), they may have a cumulative long-term effect as well. For example, a survey of a sample of business travelers found three broad components of airline-related stress (Bricker, 2005).

1) events related to air travel

2) reactions to fellow passengers

3) lack of trust in airlines/airports

Even mundane activities during air travel, like take-off and landing, can be stressful for some air travelers (McIntosh et al., 2006). Multiple studies have reflected on the stress experienced by travelers, especially those who travel for business. Globalization has increased the frequency of travel by business executives (Defrank, Konopaske & Ivancevich, 2000). Travel arrangements, hotel preferences, travel inconveniences, unhealthy lifestyles, concerns about the destination, and work/personal issues are the six main factors in travel stress (Chen, 2017).

Surveys have found luggage issues, lack of quality internet connection, length of journey, and airport delays to be important stressors (Segalla et al., 2012). Everyday commutes to work and related stressors could be detrimental to mental and physical well-being too.

Leisure and recreation travel can also generate stress, though they are often undertaken to alleviate work or personal life stress. Specifically, during the planning stage, the search and preparations could induce stress in travelers. Budgets are also stressful (Crotts & Zehrer, 2012). All these issues might be found in tweets sent by stressed travelers.

#### 1.1.2.2 Politics

Political awareness and engagement by a majority of people are instrumental for a meaningful democracy. In addition to the right candidates assuming office, it could also benefit the participants in terms of the camaraderie of working towards a common goal with people having similar ideologies (Olson, 1965) or the gratification of contributing to civic duties (Riker & Ordeshook, 1968). Nevertheless, in the wider society, political events can exacerbate or reveal polarizations. Non-academic surveys have suggested that the exit of the UK from the European Union following the June 2016 referendum has triggered or worsened stress and mental health issues in the public (Hughes, 2019).

The divisive nature of partisan politics has detrimental effects on relationships and mental health. Research has pointed to significant though short-lived upsurges in anxiety, stress, and sleep difficulties in a population of college students around the 2016 US presidential elections, for example. This also negatively affected social functioning, with an increased perception of being marginalized or targeted (Roche & Jacobson, 2019). These effects are not limited to this specific demographic or event. A broader survey, with 800 respondents varying in age, gender, economic status, found political views or discussions disrupting well-being. Here, 38% of the respondents agreed that politics had caused them stress and 26.4% reported feeling depressed when their favorite candidate lost (Smith, Hibbing & Hibbing, 2019). These studies relied primarily on survey responses and the expressions of stress due to political aspects in social media are not studied so far.

#### 1.1.2.3 Personal Events

Critical life events related to finance, relationships, and well-being were identified to elicit psychological stress on humans (Holmes & Rahe, 1967). Even happy and productive changes in life such as joining a new job, pregnancy, and marriage are included in the contributing factors for stress. A change or event, irrespective of its nature, requires the individual to adapt or cope with the modified circumstances. The analysis of the changed circumstances, together with the efforts to deal with it successfully induce an experience of psychological stress in some individuals. The expression of the psychological aftermath of life events on social

media has been studied in the context of mental health disorders such as post-partum depression and post-traumatic stress disorder (a detailed analysis is given in Section 2.1.2). However, the analysis of psychological stress related to life events in social media has been relatively sparse.

One approach (Li et al., 2014) used congratulations and condolences responses as a filter to collect tweets that potentially included mentions of life events. Using LDA topic modelling and manual labelling, the life events were extracted from this tweet collection. A multi-task LSTM using word sequence vector features was used to identify non-events, implicit events, and explicit events from tweets by active Twitter users (Yen, Huang & Chen, 2018). The dataset for personal events was collected based on the concept of a life script, which is the expected course of specific events during a typical life. The perceptions of cultural life script were similar for participants irrespective of age, gender, and educational background (Janssen & Rubin, 2011). This background has a research gap of psychological stress identification related to personal life events using social media content.

### 1.1.3 Linguistic Expressions of Stress

This section considers some key aspects of stress-related language that are investigated in the thesis.

#### 1.1.3.1 Swearing

People sometimes 'give vent' to a high-arousal mental state like stress through language. While the appropriateness of language is highly variable according to the context and speaker-listener/writer-reader relationship, swearing is usually an expression of negative emotions (Jay & Janschewitz, 2008). Though not conforming to the norms of language in formal situations, swearing performs important pragmatic roles and is a normal part of informal speech in many contexts. It is estimated to occur in 0.5-0.7% of conversational speech (Cachola et al., 2018; Mehl & Pennebaker, 2003).

There are different uses for swearing:

1) To express aggressiveness
2) To express the intensified form of an emotion
3) To emphasize an opinion
4) To accentuate informality
5) To signal group membership or friendship

In addition to these, profanity can also be cathartic, as a way of releasing or reducing pain (Pinker, 2007). This pain-relieving effect is more pronounced in people that are not habituated to offensive language. This was found in a study focused on participants' responses to a physical stress condition (a cold pressor), evaluated using the Perceived Pain Scale. The prevalence of swearing in response to everyday psychological stress (both

long-term and short-term) remains to be investigated. Similarly, the different facets of offensive language such as swearing, hate speech, and cyber-bullying in social media posts have been studied extensively as discussed in Section 2.2.4 and Section 5.1.1. However, the relation between socially offensive language and psychological stress has not been studied so far in the context of social media posts.

The offensiveness of words is highly dependent on the context of usage, without which it lacks meaning (Wajnryb, 2008). In our research, we distinguish between the offensive and non-offensive usage of swear words. However, swear words, even in the socially offensive sense, could serve one of many distinct pragmatic functions like expressing or intensifying positive or negative sentiments. This exact pragmatic role is not explored in this research.

#### 1.1.3.2 Sarcasm

Humour has long been considered to be an effective alleviator of psychological stress (Savage et al., 2017). Sarcasm, however, is a distinct genre of humour; one in which the intended meaning is often opposite to the apparent meaning and multiple emotions are expressed, typically conflicting with each other. Though it can be used as both hostile humour (Bowes & Katz, 2011) or praise (Huang, Gino & Galinsky, 2015) sarcasm in workplaces is usually studied in the context of interpersonal conflicts. Workplace anger and sarcasm have been shown to stem from personality traits, work culture, management styles, and, most importantly to our context, psychological stress (Calabrese, 2000).

Similar to swearing, sarcasm in tweets is a well-researched topic; however, no research has examined stress-related sarcasm in tweets or how it varies between domains.

#### 1.1.3.3 Temporal Understanding

Temporal understanding is a key aspect of our cognitive perception of events. Intuitively, human beings understand events with respect to a reference point in time, usually as before, during, or after the current instant. This segregates the temporal continuum into three categories: past, present, or future. The prevalence of mentions or indicators of each of these categories in someone's speech or written text correlates with their psychological state. This has been studied using two related constructs: temporal perspective, which determines the extent (distant/immediate past or future); and focus (on any of the three-time zones) (Fraisse, 1984)

The relationship between perceived temporal orientation and the ability to cope with psychological distress is a well-researched area. A study on three samples of people having experienced various traumatic incidents (incestuous abuse, Vietnam war, fire) found that a past temporal orientation is associated with prolonged

trauma-induced psychological distress (Holman & Silver, 1998). Also, time splitting (separating the past from present and future) has been found to help with refugee mental health (Beiser & Wickrama, 2004).

These studies focus on the temporal orientation of participants in the aftermath of stressful events. The temporal nature of the events could vary in high-stress content in social media. The possibility of updating events in near real-time means that the stressful events reported in social media need not necessarily belong to the past; they might still be happening or might be anticipated for the future. This relates to an additional aspect of the high-stress tweets – what is the temporal intent of responses to stressful events in different domains?

## 1.2 Motivation

Our research draws motivation from the paradigm switch in traditional psychology studies from people's explicit self-evaluations to implicit behavioural styles in order to gauge and understand their mental states. That people reveal their fundamental characteristics through their postures, gestures, and mannerisms was acknowledged early on (Allport, 1961). Usage of language constructs also bears this stamp of psychological state and characteristics as established in early psychological researches. Words people use in everyday life has been established as a third and prominent source of insight into psychological states, after self-reports and physiological markers (Boyd & Pennebaker, 2017). Specific language factors such as word length, usage of negations, articles, emotion words, etc. mark personality and psychological states, comparable to traditional self-reports and behavioural measures (Pennebaker & King, 1991). Compared to the traditional questionnaires and self-reports, a language-based psychological analysis doesn't rely on explicit self-evaluations and makes inferences from implicit cues. Our research can be viewed as an extension of this well-established school of psychological studies connecting linguistic features with mental states and characteristics.

Vast amounts of data are generated and collected online every day. In 2017, The Economist[1] named data, and specifically data collected by companies like Google, Amazon, Apple, Facebook, and Microsoft as the 'oil of the digital era', the new most valuable resource. Like oil, data also is much more useful and valuable when it is processed to extract patterns and structures. When the data is text, Natural Language Processing (NLP) plays a key role in converting it to relevant information for various purposes.

Text data from social media websites can vary in format, length, and linguistic styles. Product and movie reviews can run into several paragraphs, mostly conforming to standard grammar. On the other hand, typical posts in microblogging sites are very short sentences, including non-standard constructs like hashtags,

[1] https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

usernames, and emojis. Amidst this wide variety lies information about the collective or individual mental states in everyday life or even during a pandemic or the aftermath of a natural disaster.

Several studies have captured linguistic cues about psychological disorders from social media texts. Collecting and analysing social media posts (either anonymized or with the permission of the volunteering participants) has emerged as a viable alternative to traditional survey methods for some tasks.

Previous research has focused on specific mental disorders (post-partum stress disorder, post-traumatic stress disorder, depression) and how they manifest in social media language. The few studies of stress identification from social media have analysed stressful events in personal life, but other domains could also incite psychological stress. There is, therefore, a need to identify and analyse stress expressions from different domains and compare them to each other. This thesis examines the stressors in tweets collected from four domains.

The relationship between psychological stress and swearing, sarcasm, and temporal categories used in stressful language has been studied in traditional psycholinguistics. However, these aspects have not been analysed in social media texts, and the prevalence of swearing, sarcasm and temporal categories in stressful posts in social media is not clear. It would, therefore, be useful to analyse the occurrence of swearing, sarcasm, and temporal intent in stress-related tweets, which is an additional goal of this thesis.

## 1.3 Research Questions

The above discussion underpins the focus of this research. Since word sense disambiguation has been shown to improve the accuracy of many text classification tasks (as illustrated in Section 2.2.2), it needs to be evaluated for its potential to improve TensiStrength, the only existing lexical software for stress identification. TensiStrength's lexicon consists of terms related to stress, negative emotions, or situations that could potentially cause them. Based on the presence of these terms, the given text is assigned a stress score on a scale of -1 (no stress) to -5 (highest stress). Furthermore, given the popularity of social media, and the findings in traditional psychology about stress related to personal events, travel (airlines and traffic), and politics, it is important to examine expressions of stress in tweets belonging to these domains. In particular, stressors, sarcastic or offensive usage of language, and temporal intent in high-stress tweets are explored.

The key research questions of this study are as follows.

1. Can word sense disambiguation improve the accuracy of the lexicon-based stress strength software, TensiStrength?
2. Can stressors for different domains be identified from social media posts?
3. How prevalent is swearing and sarcastic language in social media posts with expressions of stress?
4. What is the distribution of the four categories of temporal intent (past, present, future, and atemporal) in social media posts with expressions of stress?

## 1.4 Thesis Outline

The rest of this thesis is divided into seven chapters.

Chapter 2 introduces background concepts and the core methods of the thesis, together with a brief review of literature related to the topics of the thesis. Different aspects of psychological stress are discussed together with the tools used to detect it. Existing literature on using social media posts to infer mental states and personality traits is discussed. Natural language processing (NLP) concepts such as word sense disambiguation, offensive language identification, sarcasm detection, and topic modelling, which are used in the analysis in the further chapters are briefly discussed.

Chapter 3 reports the modifications to improve the accuracy of the existing, lexicon-based method, TensiStrength, for identifying the strength of stress in tweets. A word-vector based method is employed to disambiguate the contextual meaning of ambiguous words. The existing lexicon of TensiStrength is modified to incorporate different stress strengths for different senses of ambiguous words.

Chapter 4 introduces a framework for finding stressors. It details the collection and pre-processing of the high-stress corpus. For each domain under consideration, tweets were collected based on hashtags. The tweets were pre-processed and manually annotated for the stress strength scores. Combining Latent Dirichlet Allocation topic modelling and k-means clustering, a potential stressors list is constructed for each domain. Three separate word vector methods form the next part of the framework to choose the most appropriate stressor in each tweet from the potential stressors list. The performance of the new method is evaluated in comparison to baselines from machine learning classification algorithms.

Chapter 5 introduces a Convolutional Neural Network (CNN) based classifier for identifying swearing and a Multilayer Perceptron (MLP) for identifying sarcasm. Using these classifiers, the presence of sarcasm and swearing is detected in the high-stress corpus. Also, the unigrams predominantly present in the high-stress corpus are identified and analysed based on the SAGE (Sparse Additive GEnerative) method.

Chapter 6 introduces an ensemble classifier combining SVM, Naïve Bayes, and Decision Tree classifiers to identify the temporal intent of the high-stress tweets. The distribution of four temporal classes over the corpus is presented.

Chapter 7 concludes the thesis with a summary of the experimental results, discussion, and possible directions for future work.

# Chapter 2

# BACKGROUND

This chapter sets the context for this thesis by giving relevant background information and reporting research related to the topic of the thesis. This chapter summarises research using social media posts as an indicator of mental state/health and the natural language processing (NLP) concepts employed in this research. The chapter first reviews conventional stress detection tools based on physiological manifestations of stress and individual responses to questionnaires. The benefits and drawbacks of these methods will be discussed as well as the potential for social media-based alternatives to at least partially address some of the limitations. This thesis employs natural language processing concepts such as word sense disambiguation (WSD), vector representations of words, topic modelling of tweets, clustering, and identification of swearing, sarcasm, and temporal intent of text to understand and evaluate expressions of psychological stress. This forms the premise of this research and for this reason, a brief literature review of each of these concepts is also included in this chapter. Detailed literature reviews of these concepts are provided in subsequent, corresponding chapters rather than here.

This chapter is organized as follows: In Section 2.1 the features of psychological stress are described along with evaluation using traditional tools and social media analysis. Section 2.2 gives a brief background on each of the natural language processing concepts used in this thesis. The chapter ends with a summary in Section 2.3.

## 2.1 Understanding and Detecting Psychological Stress

### 2.1.1 Detecting Stress – Traditional Tools

Stress has been traditionally measured by monitoring physiological parameters. One approach demonstrated the detection of mental stress from an analysis of heart-rate variability (Choi, 2009). With the help of spectral features and principal dynamic modes of heart rates, this system predicts the activation levels of stress-inducing sympathetic and relaxation-related para-sympathetic nervous systems. Galvanic Skin Response (GSR) has also been used as a measure of stress and arousal levels in response to cognitive load during the usage of single-modal and multi-modal versions of the same interface (Shi et al., 2007). In this study, the users were evaluated based on 36 distinct tasks, all in the domain of traffic control management. GSR levels were shown to increase with the cognitive load but tended to be lower while using a multimodal interface as opposed to a unimodal interface. The study verified that GSR can be used as a reliable indicator of user stress. Pupil Diameter (PD) is yet another physiological indicator of stress. A more recent study showed that the PD signal

is more effective and robust for identifying stress scores in computer users than the GSR signal, based on prediction accuracies of a Multilayer Perceptron and Naïve Bayes classifier (Peng et al., 2011).

Questionnaires can be used to assess the probability of stress-related disorders in individuals. There are different self-assessing questionnaires and scales widely used in stress measurement which are accepted as standards. The Perceived Stress Scale (PSS), for example, lists 10 questions that require the respondents to rate themselves on a scale of 0 to 4 (Cohen, Kamarck & Mermelstein, 1983). The questions do not probe specific incidents but prompt the user to evaluate their responses and feelings over the previous month (e.g. in the last month, how often have you felt nervous and stressed?, in the last month, how often have you found that you could not cope with all the things that you had to do?). Another widely accepted system, the Social Readjustment Scale, listed 43 life events that were found from clinical experience to require considerable life adjustments (Holmes & Rahe, 1967). Each was assigned a life change score based on how traumatic it was perceived to be by the participants. The primary usage of this scale would be to calculate stress probabilities. For each individual, the scores experienced over a year were totalled. If the sum is less than 150, he/she has a 30% chance of being stressed. If it was 150-299, this chance was 50%, and for scores above 300, an 80% chance. However, such an evaluation of stress ignores the subjectivity of perceived stress. The personality of individuals could also be a contributing factor to stress (Kittel, Kornitzer & Dramaix, 1986). The same event could elicit different responses in different individuals depending on their ability to cope with the associated change. Also, this study only considers events in personal life, including relationships (death of a spouse, divorce, pregnancy), work (troubles with the boss, being fired, retirement) finances (new mortgage, foreclosure of the mortgage), and social life (detention in jail, change in church activity, change in social activity). In our research, we consider three more domains, traffic, politics, and airlines in addition to personal events, and draw comparisons between the stressors and expressions of stress in each domain.

The above stress detection methods are mostly reactive in nature, requiring the researchers to continuously monitor sensors or rely on users who might be reluctant to share direct observations on their mental states or manipulate their responses to meet self-imposed psychological images. These limitations provide a motivation to develop novel, unobtrusive sources of information to analyse stress and relaxation expressions, and mental health states, in general.

### 2.1.2 Evaluating Mental States from Social Media Content

The ubiquity of social media makes it a potential tool for behavioural and mental health evaluation. In recent decades, social media has been increasingly used for sharing events, connections, and memories. This has created a repository of information that could potentially be mined to gain insights into personality types and mental disorders. Research about the psychology of social media often focuses on one of the following:

1) Personality types and traits
2) Psychological disorders
3) The influence of social media on mental states

The relevant works in these three categories can be classified according to four parameters (Table 2.1):

- Focus – the trait/disorder analysed in the study.
- Social media – the social media platform from which the data was collected.
- Factors – Features extracted from social media posts.
- Analysis tool – Algorithms and techniques used to analyse the chosen features.

**Table 2.1 Analysis of research on mental health and personality traits using social media**

| Topic | Authors | Focus | Social Media | Factors | Analysis Tool |
|---|---|---|---|---|---|
| Personality And Traits | Karanatsiou, 2019 | Psychological profiling of organization leaders vs random users | Twitter | Emotion features Linguistic features(tf-idf, ngrams, POS tags) Behaviour features (twitter activity, statistics) | Random forest regression model |
| | Tutaysalgir et al., 2019 | Big 5 personality | Twitter | | clustering |
| | Hoover & Portillo, 2019 | Morality | Tweets | Word frequencies | SVM/ LSTM |
| | Giachanou et al., 2019 | User behaviour in spreading fake messages or checking facts | Twitter | Word embeddings combined with personality traits and linguistic features | CNN |
| | Baali & Ghneim, 2019 | Emotions | Twitter | Word, sentence embeddings | CNN |
| | Bayrak & Alper, 2021 | Morality | Twitter | Linguistic features | Differential analysis |
| | Kulkarni et al., 2017 | Latent Traits | Facebook | Linguistic features | Differential analysis & correlation coefficients |
| | Buechel et al., 2018 | Empathy and distress | Online news articles | Linguistic features | CNN-based predictive model |
| | Liu et al., 2016 | Big-five personality traits | Twitter | Profile picture (colour, composition, type, demographics, expressions) | Pearson correlation between facial features and traits |
| | Souri, Rahmani & Hosseinpour, 2018 | Big-five Personality traits | Facebook | Likes, profile, networking, and posts information | Classification algorithms |
| | Wang, 2015 | MBTI personality traits | Twitter | Bag of n-grams, POS | Logistic Regression classification model |

| | | | | | |
|---|---|---|---|---|---|
| | | | | tags, and word vectors | |
| | Zamani, Buffone & Schwartz, 2018 | Human trustfulness | Facebook | Ngrams (1 to 3) and LDA topics of status updates | Ridge regression |
| Psychological Disorders | Kotikalapudi & Chellappan, 2012 | Depression | Email/chatting | Internet usage | Statistical analysis using correlation |
| | Moreno et al., 2011 | Depression | Facebook | Status updates | Negative binomial regression analysis |
| | Gamon et al., 2013 | Depression | Twitter | Emotional and linguistic features | SVM classifiers |
| | Gruda & Hasan, 2019 | Anxiety | Twitter | Vector embeddings, unigrams, and bigrams of words and emojis | Bayesian Ridge Regression |
| | Leis, Ronzano & Mayer, 2019 | Depression | Twitter | Behavioural and linguistic features | Differential feature analysis |
| | Coppersmith et al., 2014b | PTSD | Twitter | Ngrams and LIWC data of Tweets | Logistic regression classifier |
| | Eichstaedt et al., 2018 | Depression | Facebook | Linguistic, emotional, interpersonal, and cognitive | Logistic regression |
| | Choudhury, Counts & Horvitz, 2013 | Post-partum depression | Twitter | Social engagement, emotional and linguistic styles | Predictive model |
| | Coppersmith et al., 2014a | PTSD, Depression Bipolar and Seasonal Affective disorders | Twitter | LIWC, language models | Correlation coefficients and classification models |
| | Lin et al., 2016a | Stress | Sina Weibo | Tweet level and user level attributes | CNN based mobile App (Moodee) |
| | Lin et al., 2016b | Stress | Sina Weibo | Word embeddings | Hybrid model combining multitask learning with CNN |
| | Lin et al., 2017 | Stress | Sina Weibo | Social network interactions | Factor graph model and CNN |
| | Lin et al., 2014 | Stress | Sina Weibo Tencent Weibo and Twitter | Tweet level and user level | CNN to generate user attributes and deep neural networks to detect stress |
| | Jia et al., 2014 | Stress | Sina Weibo | Tweet level attributes (linguistic, | Deep sparse neural networks to detect stress |

| | | | | | |
|---|---|---|---|---|---|
| | | | | visual, and social) | |
| | Zhou, Hu & Wang, 2019 | Depression, Anxiety, Panic, OCD, and Bipolar | Twitter | 11-dimensional sentiment categories based on linguistic features | Sentiment distribution similarity calculation |
| | Jia, 2018 | Stress and depression | Sina Weibo | Content and posting info, social interactions | Cross-domain DNN |
| Addiction to social media | Kircaburun & Griffiths, 2018 | Instagram addiction and personalities | Instagram | Instagram addiction scales, self-liking, and Big five traits | Questionnaires and statistical analysis |
| | Kuss & Griffiths, 2011 | Addiction to social networking sites | Various social networking sites | Personality traits and addiction | Literature review |
| | Kuss & Griffiths, 2012 | Addiction to online gaming | Online gaming communities | Addiction scales | Literature review |

**Focus** An individual's personality traits govern their inner feelings and this, in turn, is reflected in their interactions with the world. In the context of social media, this is expressed by the language of posts, images/videos shared, and the volume of connections/interactions with the other members. Data from Twitter and Facebook has been used to categorise and predict the personality types of the user. The collective information of demographic sections has been even used to manipulate elections and for the benefit of business organizations, as exposed in the Cambridge Analytica scandal (Boldyreva, 2018).

Personality research often focuses on the personality traits of Extraversion, Openness, Conscientiousness, Neuroticism, and Agreeableness, which are collectively called the Big Five (Digman, 1990; McCrae & John, 1992). In contrast, the Myers-Briggs Type Indicators (Myers, 1962) assign one of sixteen types to each individual based on the paired characteristics of Introversion/Extraversion, Sensing/Intuition, Thinking/Feeling, and Judging/Perceiving. Expressions in social media text have been analysed in the context of personality profiles based on the Big Five (Liu et al., 2016; Souri, Rahmani & Hosseinpour, 2018)and MBTI (Wang, 2015) classes. Apart from these widely studied traits, granular mental characteristics like empathy, distress (Buechel et al., 2018), trustfulness (Zamani, Buffone & Schwartz, 2018), latent traits (Kulkarni et al., 2017), morality (Bayrak & Alper, 2021; Hoover et al., 2019) have been analysed from social media inputs.

Social media data has been mined for indicators of psychological disorders, including depression (Gamon et al., 2013) Post Traumatic Stress Disorder (PTSD) (Coppersmith et al., 2014b), anxiety (Gruda & Hasan, 2019), and post-partum behavioural changes (Choudhury, Counts & Horvitz, 2013). The linguistic features of content and networking information have been analysed through statistics and machine learning classifiers to predict the onset of such disorders. Some studies (Coppersmith et al., 2014a; Zhou, Hu & Wang, 2019) analyse multiple depression-related mental disorders, like panic attacks, anxiety disorder, and seasonal affective disorder.

Internet usage patterns (Kotikalapudi & Chellappan, 2012) and Facebook status updates (Moreno et al., 2011) can help to identify people at risk of depression. The potential for mining social media postings for indicators of mental illness is known, including multimedia opportunities and ethical challenges (Gamon et al., 2013). In a related work, tweets have been used as inputs to predictive models about the postpartum emotional and behavioural changes in new mothers (Choudhury, Counts & Horvitz, 2013). Using observations on prenatal behaviours, specific attributes such as engagement, emotion, ego network, and linguistic styles, these models could classify mothers who would exhibit significant postpartum changes with an accuracy of 71%. While additionally considering the initial postnatal data, this accuracy increased to 83%. Similarly, in crowd-sourced Twitter user data was used to build a classifier that estimates risk before the onset of a condition. This study further analyses the dataset characteristics to reach observations on diurnal patterns, social connectedness, and volume of postings for both depression and non-depression classes. The model was able to predict the onset of depression with an accuracy of 72.4% with all features and reduced dimensions (Gamon et al., 2013). (Coppersmith et al. (2014)) use an automated analysis followed by manual refinement to find positive and negative samples for PTSD. The tweets posted by these users were collected during a time window to create a corpus of positive and negative PTSD data, to train three different classifiers. These were used to identify and evaluate PTSD trends in tweets across military and civilian populations. This method was further enhanced (Coppersmith et al., 2014a) for indicators of bipolar disorder, major depressive disorder, and seasonal affective disorder.

Social media can itself contribute to mental health issues. Addiction to various social media can impede the normal functioning of individuals. There have been studies about the excessive usage of social networking sites (Kuss & Griffiths, 2011; Kuss & Griffiths, 2012). Addiction to Instagram has also been shown to relate to Big-five personality traits, using a self-report survey (Kircaburun & Griffiths, 2018).

Since stress is a known underlying reason for many psychological disorders, there have been efforts to identify the expressions of stress from social media content. One hybrid system (Lin et al., 2017) combines a factor graph model and convolutional neural network (CNN) and estimates the relationship between users' psychological stress levels and their social network interactions. This method investigates the social interactions of stressed and non-stressed users and improves stress detection by 9% in terms of F-measure

over an earlier study (Lin et al., 2014) which considered only user-level posting attributes. Moodee (Lin et al., 2016a), a practical mobile application, addresses the challenges of stress detection from social media content, including missing data, time-series modelling and data sparsity problems. It extracts tweet-level linguistic, visual, and social attributes defined in a prior study (Lin et al., 2014). These attributes are fed into cross auto-encoders which are embedded in a CNN, integrating them to user-level content attributes. This system recommends links to users to dissipate stress. Another system (Lin et al., 2016b) identifies stress in social media texts. This uses stressor event categories and stressor subjects and a collection of words related to each category and subject based on word embeddings. The stressor event and subjects are identified using a novel hybrid model that combines multitask learning with CNN. Tweets are assigned a stress value based on the Social Readjustment Rating Scale (SRRS) (Holmes & Rahe, 1967). The stress scores obtained this way give results comparable to some state-of-the-art machine learning models. Though these studies address psychological stress expressed in social media, their work uses data from Sina Weibo, the Chinese equivalent of Twitter. The applicability of the results in the English content of Twitter is thus yet to be verified. Furthermore, the research focuses on user attributes and personal life and the analysis of stressors in other domains is an open research question.

TensiStrength (Thelwall, 2017) is the first published system to detect the strength of both stress and relaxation expressed in tweets. It primarily uses a lexical approach. The lexicon is partly derived from LIWC (Tausczik & Pennebaker, 2010; Chung & Pennebaker, 2007), General Inquirer (Stone et al., 1966), and emotion terms from SentiStrength (Thelwall et al., 2010) a similar lexicon-based sentiment analysis program. For the original evaluation, a corpus of 3066 English tweets was human-coded for stress and relaxation on two parallel five-point scales: from -1 denoting no stress to -5 denoting the highest stress; from +1 for no relaxation and +5 for the highest relaxation. Its performance (with an accuracy of 49.3%) was similar to several machine learning algorithms, including Support Vector Machines. More information about TensiStrength is provided in Chapter 3, where an experiment on improving its accuracy using word sense disambiguation is discussed.

**Social media** Most social media research uses Twitter or its Chinese equivalent Sina Weibo for data (Table 2.1). Facebook status messages and network statistics have been analysed for cues about personality or mental disorders too. Other studies use discussion forums and email/internet usage statistics. In Twitter, messages are limited to 280 characters. Despite this brevity, Tweets have been found to communicate enough information about some users' mental states to make inferences. Twitter's free APIs, which facilitate real-time collection and analysis, could be a reason for its popularity as a data source for mental health investigations.

**Factors considered** Multiple factors from social media data have been shown to be cues for mental states. Typically, linguistic, visual, and social features of the users' activities have been extracted and analysed. N-grams, LIWC data, Part-of-Speech tags, and word vectors are among the linguistic features used. Images are a

significant part of social media messages, including profile pictures. They contain visual clues like colour, composition, and expressions which are useful predictors of a mental state or trait. A third category of features is the activity of users and the relative strength and density of their networks.

**Analysis tools** Statistical analysis (Kotikalapudi & Chellappan, 2012), traditional machine learning (Eichstaedt et al., 2018; Coppersmith et al., 2014a; Choudhury, Counts & Horvitz, 2013; Wang, 2015) and deep learning models (Lin et al., 2014; Lin et al., 2016a; Lin et al., 2016b; Lin et al., 2017) have all been used to investigate social media data. Machine learning algorithms like Support Vector Machines (SVM) and Logistic Regression are widely used in classification and prediction tasks of mental health categories from social media, with significant results. Deep learning techniques like deep sparse neural networks and convolutional neural networks are also used in some studies. The performances of these different methods are not directly comparable since these studies investigate different aspects of mental health in different datasets. As early as 2012, social media was identified as an influential element in clinical practices and training (Kolmes, 2012). With the proliferation of spontaneous sharing of personal experiences and states, it was a natural source of information for understanding willing patients/participants in a better way. As established by the corpus of previous literature (Table 2.1), presently, social media language is a useful tool for identifying mental health markers. Together with traditional research methods in psychology, it is valuable as a non-obtrusive instrument to help understand the social ties and psychological states of the participants. The analysis of unprompted usage of language on social media platforms may reveal psychological traits and states.

Particularly, NLP methods are applied to textual data from social media to extract various psychological constructs such as personality traits, behavioural patterns, and mental states. The focus of our research is to employ NLP techniques to explore the markers of psychological stress present in tweets.

We can clearly demarcate the gaps in the existing research which our research is addressing:

1. Identification of stressors in English tweets in multiple domains
2. Domain-wise analysis of expressions of psychological stress in English tweets, not limited to the measure but also considering swearing/socially offensive language, sarcasm.
3. Distribution of temporal classes in English tweets with expressions of stress belonging to different domains.

The next section investigates the natural language processing concepts which form the background of the methods and analysis of the research questions in this thesis.

## 2.2 Natural Language Processing Concepts in Social Media Analysis

### 2.2.1 Sentiment Analysis of Text

The extraction of sentiments or opinions present explicitly or implicitly in social media content is a challenging and useful research problem, which is comparable to the task of identifying expressions of stress.

Sentiment analysis can be defined as the extraction of sentiments or opinions towards a specific entity or its attributes (Liu, 2012). Data from social media texts often contain non-standard language and a key step in sentiment analysis of social media data is pre-processing the text to handle features like repeated punctuation, emoticons, irregular casing, and slang (Balahur & Jacquet, 2015) Another aspect of social media content is its multimodality- often, social media posts combine text with videos or images. There have been efforts to use visual features in the prediction of sentiments. One such study classified Flickr images into 5 emotion categories (happiness, sadness, fear, violence, and love) using various machine learning and neural network classification methods (SVM, RESNET, and VGGImageNet) (Gajarla & Gupta, 2015)

Multilinguality and code-mixing are challenging aspects of social media content. Non-native English speakers often combine words in English and their native language, which makes it difficult for sentiment analysis algorithms to perform accurately. A SemEval task, SentiMix, recently addressed this challenge of predicting sentiments of tweets in code-mixed languages in English-Hindi and English-Spanish (Patwa et al., 2020).

Sentiment analysis of social media posts has been used for a multitude of applications. The collection and analysis of customers' opinions about a product or service is a common example. The ability of social media to distribute emergency information has been used in emergency responses to several natural disasters. An analysis of sentiment in such posts could help authorities make better-informed decisions (Beigi et al., 2015). Predicting election results has been an interesting application of sentiment analysis of social media, especially that of Twitter. In the past, analysis of Twitter sentiment has been applied in predicting congress and senate elections in the US and Irish general elections. More recently, word co-occurrences and semantic relations captured using the BiTerm Topic Model were employed in the analysis of elections in the Indian state of Uttar Pradesh (Bansal & Srivastava, 2018) Despite these results, electoral predictions using sentiment analysis are also deemed invalid by some researchers (Gayo-Avello, 2012) because the research is often post-hoc, uses the raw volume of tweets and lacks a reliable gold standard.

Sentiment analysis has been applied to tweets related to the financial domain for various purposes. Using SVM-based sentiment classification of tweets, a study predicted stock price movements (Smailovic et al., 2013). Another research work (Ao, 2018) built a sentiment analysis model for financial news tweets and brought out the relations between new sentiment scores and stock market prices. Also, a lexicon for sentiment

analysis of the financial domain demonstrated that including context and domain information in the lexicon significantly improved the accuracy of sentiment score prediction of financial tweets (Tabari & Hadzikadic, 2019).

There have been several different approaches to classify sentiment in social media texts using lexicon-based or supervised machine learning approaches. Recently, deep learning approaches seem to have been working particularly well. Lexicons consolidate words (typically adjectives) with their sentiment scores. For a given text, the sentiment scores of the adjectives present in it are aggregated in a predetermined method to generate a single sentiment score. There are several sentiment lexicons widely used for social media text; such as SentiWordNet (Baccianella, Esuli & Sebastiani, 2010), VADER-Valence Aware Dictionary for sentiment Reasoning (Hutto & Gilbert, 2015), and SentiStrength (Thelwall et al., 2010). Some of these use sentiment strengths from previously and independently created lexicons like LIWC (Tausczik & Pennebaker, 2010) and Harvard General Inquirer (Stone et al., 1966) and add further dictionary entries to create novel lexicons for social media texts.

Supervised machine learning algorithms like Support Vector Machines (SVM), Naïve Bayes (NB), Logistic Regression (LogReg), and Deep learning algorithms like Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Bidirectional Encoder Representations from Transformers (BERT) have also been used extensively to solve the task of sentiment analysis in tweets (Table 2.2).

**Table 2.2 Selected sentiment analysis studies of Twitter**

| Authors | Method | Domain | Features |
|---|---|---|---|
| Li & Fleyeh, 2018 | Elastic Net (outperforms Logistic Regression, Naïve Bayes, Neural Network, Random Forest and SVM) | Swedish tweets about IKEA stores | Word vector representations |
| Salas-Zarate, 2017 | Lexicon-based using SentiWordNet | Diabetes-related tweets | n-grams |
| Alfina et al., 2017 | Naïve-Bayes | Political tweets | Sentiment polarity of hashtags |
| Goel, Gautam & Kumar, 2016 | Naïve Bayes | Generic | Sentiment140 lexicon features |
| Badeaa, Gomaa & Haggag, 2017 | Majority voting between sentistrength lexicon scores, Sequential Minimal Optimization, and CNN | University of Michigan Sentiment Analysis competition on Kaggle | Lexicon scores and word vector representations |
| Monika, Deivalakshmi & Janet, 2019 | RNNs and LSTMs | US airline tweets | Word vector representations |
| Chen et al., 2018 | LSTM | Twitter dataset crawled by REST APIs (350000 users) | Bi-sense emoji embeddings |
| Gamallo & Garcia, 2014 | Naïve-Bayes | Generic (SemEval-2014 task 9 dataset) | lemmas, polarity lexicons, and valence shifters |
| Assefa, 2014 | Logistic Regression | Generic (SemEval-2014 Task 9 Twitter dataset) | Bag of words, lexicon features (NRC Hashtag, Seniment140), word vector representations |
| Kenyon-Dean et al., 2018 | Logistic Regression | McGill Dataset 7026 tweets on sports, food, media, general, and communication technology | n-grams, word embeddings, SentiWordNet scores |
| Oscar et al., 2017 | Keyword lexicon and machine learning classifiers | Alzheimer's disease-related tweets | n-grams |
| Azzouza, Astouti & Ibrahim, 2020 | BERT | SemEval 2017 dataset | Word embeddings as input to BERT |
| Yuan & Zhou, 2015 | Recursive Neural Networks | SemEval-2013 dataset 6092 tweets | Binary dependence tree built from pre-processed tweets |

This is not a comprehensive survey of Twitter sentiment analysis research, which is a large research field. It represents the variety of methods used in the analysis of sentiments in Twitter and the applications for which they have been used. In addition to lexicons, classification models like Support Vector Machines, Naïve Bayes, and Logistic Regression, deep learning models like Convolutional Neural Networks and Recursive Neural Networks are also viable, high-performing methods. These methods have been used to detect sentiments in tweets belonging to politics, health, sports, and airlines (Table 2.2).

This review points to the usefulness of different methods and features in the task of sentiment analysis which varies according to the specific domain and dataset being analysed. This motivates us to consider a number of

well-performing methods based on the literature study and choose the one which is most suitable for our dataset. This approach is followed in the tasks of identifying stressors (Chapter 4), swearing/socially offensive language(Chapter 5), sarcasm (Chapter 5), and temporal intent (Chapter 6) in this thesis.

### 2.2.2 Word Sense Disambiguation (WSD)

Since word meanings can differ between contexts, it is important to decipher the contextual meaning of the constituent words to interpret the meaning of a text. WSD is the process of finding the correct sense of an ambiguous word in a given context (Pal & Saha, 2015). Lexical databases like WordNet (Miller et al., 1991) include hierarchically organized word senses. Traditional machine learning and deep learning methods have been used to disambiguate word senses too.

An important challenge for detecting expressions of stress or any affective state is that social media texts use non-standard grammar and informal language (Navigli, 2009). Affect words can be ambiguous, with their sense changing according to the context. For example, in "He gave me a cool stare", the affect word "cool" indicates stress whereas in "She wore a pretty cool dress", it does not. A system to resolve the meaning of this word is essential to accurately identify stress. The natural language processing task of WSD is the traditional response to this issue.

WSD has been demonstrated useful in sentiment analysis systems. A WSD system (Sumanth & Inkpen, 2015) using Babelfy and SentiWordNet illustrates how it can improve the accuracy of sentiment detection in Twitter and SMS test data. Babelfy is a multilingual graph-based method for disambiguating word senses. It is based on the semantic network BabelNet 3.0 (Navigli & Ponzetto, 2012). For an analysis with this tool, the tweets were pre-processed using the tokenizer of the Carnegie Mellon University (CMU) Twitter NLP tool (Gimpel et al., 2011) and NLTK[2]. The tweets were represented as vectors with three features – sums of corresponding positive, negative, and neutral scores of words in SentiWordNet. A supervised Random Forest Decision Trees classifier was trained based on these representations. This system was shown to have better accuracy (58.55%) than the baseline method (45.26%). Though this study ignores features such as negations, punctuation, and emoticons which are critical in identifying sentiments, it essentially establishes that WSD improves sentiment analysis accuracy for tweets, which motivates our study on incorporating WSD in the TensiStrength lexicon.

A study of sentiment analysis for figurative language (Rentoumi et al., 2009) applied WSD and assigned a polarity to a word sense through a graph-based method. Sentence-level polarity was detected based on two Hidden Markov Models, each trained for positive and negative cases. This system has a significantly better recall and precision compared to a polarity detection method without WSD and a baseline WSD method (which

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

assigns the first sense entry in WordNet to all senses). On a dataset of 1000 newspaper headlines containing metaphors, this system was found to perform with precision and recall of 76.5% and 72.6%, a substantial improvement over systems with no WSD (precision- 48.5% and recall- 46.75%) and baseline WSD (precision- 57.0% and 54.2%).

Incorporating WSD based on the context of word-of-mouth documents to modify SentiWordNet lexicons has been found to improve sentiment analysis performance. Similarly, a WSD system (Hulpus, Prangnawarat & Hayes, 2015) uses a path-based semantic relatedness to find the most appropriate sense and it is found to give a better classification F-measure for sentiment analysis task. While varying in methodology and lexical resources, these studies illustrate that sentiment analysis systems with WSD substantially outperform those without WSD. However, the implications of incorporating WSD in stress/relaxation analysis have never been studied before and our research treats it as a potential way to improve the accuracy of an existing lexical method.

### 2.2.3 Topic Modelling of Tweets

Topic modelling is the extraction of latent topics from documents. Different topic modelling algorithms work on the common assumption that each document is a collection of topics and each topic is a collection of words. Their common objective is to discover the semantics of documents. We use this extraction of topics to find stressors from a collection of texts. Two common topic modelling methods for documents are Latent Semantic Analysis (LSA) (Deerwester, 1988) and Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) LSA is a frequently used topic modelling method that forms a term-document matrix M and then factorizes it by Singular Value Decomposition into a product of three different matrices $M = U*S*V^T$ where S is a diagonal matrix of singular values, U is the document-topic matrix and V is the term-topic matrix. LDA is a Bayesian version of probabilistic LSA. For the LDA model, a document is generated as follows: A topic distribution is chosen from a Dirichlet distribution φ, and from this, a specific topic T is selected. From another Dirichlet distribution, the word distribution ϕ is chosen and a specific word w is selected from it.

The applicability of these methods to tweets is hindered by unconventional constructs, slang, and emoticons. Models like LDA perform well on longer documents. Unconventional constructs like abbreviations, URLs, hashtags, mentions, and emoticons need to be handled to preprocess the tweets for topic modeling. Emoticons consist of punctuation marks to indicate the expression of emotions. They can be useful features in NLP tasks such as sarcasm detection and sentiment analysis. Such features are retained while analyzing tweets for opinions or emotions. However, in topic modeling, such constructs are typically removed as a part of preprocessing and tweets are represented simply as a bag of words (Sha et al., 2020; Yu et al.,2019; Vayansky, Kumar & Li, 2019; Xue et al., 2019).

In addition, traditional topic modelling techniques are designed to work on documents that are a few hundred words long and hence are not ideal for tweets which are typically much shorter (Jonsson & Stolee, 2016). To overcome these issues, aggregating related tweets into individual documents has been proposed as a potential solution, called pooling. Hashtag pooling, in which tweets containing the same hashtag are aggregated together and treated as a single document, is one of the most widely accepted pooling methods to overcome the limited coherence of LDA on Twitter data. Tweets can be aggregated based on other factors, such as the author, hour of publication, and burst (trending topic). Hashtag pooling performs better than the other pooling schemes (author-wise, hourly, and burst-wise) based on Point-wise Mutual Information (PMI), NMI scores, and purity scores (Mehrotra et al., 2013). This comparison was performed on a generic dataset, a named-entity-specific dataset, and an event dataset. Conversation-pooling is another scheme for tweet pooling in which tweets and their replies are aggregated into a single document (Alvarez-Melis & Saveski, 2016). The users who participate in the conversation are considered to be co-authors of this pooled document. Conversation pooling is shown to have better performance over hashtag pooling, however, the dataset was collected using the public tweets of the most influential users in each topic under consideration. No information is given about the distribution of hashtags over the collected tweets. With a lack of tweets sharing the hashtags, pooling would be ineffective and the relatively worse performance of hashtag pooling could be attributed to that. In the absence of a randomly collected dataset, the performance comparison of different pooling methods could be inconclusive. In our research, topic modelling methods have been used in extracting stressors in Tweets and it is discussed in more detail in Chapter 4.

### 2.2.4 Identifying Swearing/Socially Offensive Language in Social Media

Abusive behaviour is a pervasive issue on online social media networks. It can take the form of inappropriate swearing or hate speech targeted at an individual or a specific group of people. Swearing is also commonly used in non-offensive contexts. An obvious solution to detect swearing is to scour the messages by list-based or manual social moderation systems for occurrences of inappropriate words/phrases. However, it is very easy to circumvent list-based systems using partial disguises (b***h, @$$, etc.) or misspelt words (hellll, dammn). Screening each message using manual moderators is impractical in active social networks with many users and messages. Moreover, the context determines the implied meaning of many words. Non-offensive words like pests, rodents, and mice have been used in racist tweets, targeting specific communities, which would easily escape list-based systems.

Machine learning algorithms and classification systems have been used to identify offensive language and swearing in social networks. A Lexical Syntactic Feature-based architecture (Chen et al., 2012) detects offensive language at the user level by hand-authoring syntactic rules. There have been efforts to translate offensive sentences into non-offensive ones (Bahdanau, Cho & Bengio, 2014) describes a simple encoder-

decoder with attention that functions as a reasonably well-performing translator. A detailed analysis of the detection of offensive language from social media texts and related shared tasks is presented in Chapter 5, which motivates the choice of model in our study of swearing/socially offensive language in stressful tweets.

### 2.2.5 Identifying Sarcasm in Social Media

Sarcasm is a form of speech in which the implied meaning contradicts the literal meaning. It is a non-trivial task for even humans to correctly identify sarcastic speech in many cases, which makes it a hard problem for machines. There have been attempts to identify sarcastic messages in social media using machine learning algorithms and deep learning.

González-Ibáñez (2011) used SVM-based classifiers with unigrams, dictionary-based features, and user references for sarcastic and non-sarcastic classification. Reyes, Rosso, & Veale (2013) identifies irony using features such as ambiguity, polarity, unexpectedness, and emoticons. Juxtaposing a negative situation with an expression of positive sentiment is a key feature of sarcastic tweets which has also been used (Riloff et al., 2013). When used in conjunction with an SVM classifier this contrast method gave an F-measure of 0.51.

The subtle ways in which sarcasm can be expressed make it difficult to identify with n-grams and other text features. There have been efforts to address the problem with deep learning frameworks. Poria, et al. (2016) developed a pre-trained CNN for extracting sentiment, emotion, and personality features for sarcasm detection. It uses word embedding (Word2Vec) representations of sentences. Pre-trained CNN models have been used to extract sentiment (positive, negative, and neutral) emotion (anger, disgust, surprise, sadness joy, or fear), and personality features (Big Five personality traits). Two different predictive models using only CNN and using CNN-SVM (features extracted by CNN and then fed to SVM) were evaluated on datasets of Tweets and the latter gave the better performance. Another study (Zhang, Zhang & Fu, 2016) exploits deep neural networks (DNN) for sarcasm detection. It uses a two-component structure to extract features- a gated RNN and gated pooling function for tweet content features and a pooling function for contextual features. This neural model achieves significantly better accuracy compared to a baseline system. We explore the different models and features for the identification of sarcasm in Chapter 5 and construct a model for detecting sarcastic language in our corpus based on the learnings from these existing research studies.

### 2.2.6 Temporal Intent of Social Media Posts

The temporal intent of a post is relevant to the detection of stress in it (e.g., “I am in a rush” vs. “I was in a rush but am finished now”). The TempEval (Verhagen et al., 2007; UzZaman, 2012) and NTCIR Temporalia tasks (Joho, Jatowt & Blanco, 2014) established the identification of temporal class as a significant Natural Language Processing task.

The different approaches to distinguish a unit of text as belonging to past, present or future temporal classes can be broadly divided into rule-based systems and machine learning classifiers. Rule-based systems HeidelTime (Strötgen & Gertz, 2010; Strotgen et al., 2013, Strötgen et al., 2014), SUTime (Chang & Manning, 2012) focus on understanding linguistic patterns and regular expressions to extract temporal information. On the other hand, classifier-based systems often use neural networks (Hasanuzzaman, et al., 2017) or machine learning algorithms (Kamila 2018) to learn temporal classes from social media text, based on features like vector representation of text, n-grams, lexicon-based tags, and time expressions. This temporal information is often further aggregated at the user level to predict the socio-economic and psychological characteristics of individuals, inspired by traditional psychology studies. For example, one study has investigated the correlation between users' temporal orientation and factors like openness, life-satisfaction, depression, IQ, and conscientiousness, based on language in Facebook posts (Schwartz et al., 2015). Present orientations decreased with age and past orientations increased with age. Other interesting results were a positive correlation between future orientation and satisfaction with life, a negative correlation between openness to experiences and future orientation, and a negative correlation between extraversion and past orientation. Hasanuzzaman, et al. (2017) found a positive correlation between future temporal orientation and higher income. Kamila (2018) explored the psycho-demographic factors of users based on the temporal orientation of their sentiment.

These studies establish the significance of temporal orientation being related to socio-demographic factors and mental health states. This prompts us to further examine the existing research methods in the identification of the temporal classes in social media text and design a method most suitable for our application (Chapter 6).

## 2.3 Summary

This chapter reviewed research detecting mental health issues, and specifically stress, from social media using natural language processing methods. Section 2.1 described the concept of stress, the traditional methods used to assess it with the shortcomings, and the broader area of mental health assessment using social media content. Research on mental health and traits assessment using social media was categorized depending on the focal area: personality traits, mental disorders, addiction to social media. The review showed that though mental health disorders such as depression or post-partum disorder have been studied using social media, research on the expressions of psychological stress is rather sparse. The few existing studies focused on stress as a response to prolonged traumatizing events in personal life and did not consider domains such as politics or travel which have been found to be stress-inducive in traditional psychology studies. Section 2.2 examined natural language processing research for processing social media content. Sentiment analysis, the extraction of sentiments or opinions from natural language text is closely related to the task of evaluating stress. Section

2.2.1 presented a review of the challenges, applications, and methods of the task of sentiment analysis. Disambiguating the sense of a word based on the context of its usage has improved the performance of sentiment analysis tools (Section 2.2.2). Specifically, incorporating disambiguated word senses was shown to improve the performance of SentiWordNet, the sentiment strength lexicon, which motivated our experiment on improving the TensiStrength lexicon with WSD. The various topic modelling techniques were discussed in Section 2.2.3 together with its applicability in social media data. Pooling of tweets based on criteria like hashtags, author, or conversation was found to improve the performance of topic modelling. Conversation pooling outperformed the other pooling techniques, but as described in Chapter 5, we chose hashtag pooling for the identification of stressors.

The identification of swearing(Section 2.2.4) sarcasm(Section 2.2.5) and temporal intent(Section 2.2.6) from social media was briefly discussed in the following sections. A few key existing works were discussed for each of these tasks. A more detailed analysis of the state of the art in each of these tasks is presented in the respective chapters (swearing, sarcasm – Section 5.1.1; temporal intent –Section 6.1), which formed the basis of our choice of methods to identify these aspects in the high-stress corpus.

# Chapter 3

# IMPROVING THE STRESS LEXICON WITH WORD SENSE DISAMBIGUATION

This chapter describes the improvement of the performance of the lexical algorithm TensiStrength for identifying the magnitude of stress and relaxation expressed in Tweets. We chose an existing word-vector based method for word sense disambiguation (Chen, Liu & Sun, 2014) and updated the TensiStrength lexicon to include multiple senses of ambiguous words.

The chapter is organized as follows.

Section 3.1 describes the related research we use in this chapter: TensiStrength, WordNet, and the word sense disambiguation method using word vectors. Section 3.2 describes the modifications to TensiStrength, dataset and annotation, and the experimental setup. Section 3.3 presents and analyses the results. The error analysis and limitations are presented in Section 3.4. The chapter concludes with a summary of the results and findings in Section 3.5.

## 3.1 Background

Identification of stress from social media texts is a relatively new research task. In a study (Lin et al., 2016b) of Sina Weibo, Weibo posts' stress score is determined by the SRRS (Social Readjustment Rating Scale (Holmes & Rahe, 1967) entry for the stressor events and subjects in the tweet found using a CNN-based hybrid model. The SRRS is a collection of 43 events in personal life with a corresponding score indicating the level of stress induced by the events. These stress-inducing events, called Life Change Units (LCUs) were collected from empirical evidence of a clinical history of 5000 patients. Though the SRRS is widely used in psychological stress research, this approach does not leverage the linguistic cues from words other than those indicating the stressor events. Further, the events themselves are limited to those only from personal life (marriage, death of a spouse, childbirth, etc.). Another approach (Lin et al., 2016a) uses a DNN with content attributes learned by a CNN with Cross Auto-Encoders (CAE) and manually defined statistical attributes. Though this system performs better than an array of machine learning models, the disadvantage is that the dataset considers only explicit declarations with the sentence pattern "I feel stressed". With this criterion, implicit expressions of stress are missed, and this might affect the feature learning and performance evaluation. Additionally, the linguistic attributes are limited to emoticons, punctuation, and emotion words in LIWC. However, stress and negative emotions are not the same.

As observed in psychological studies, stress and negative emotions are different though they co-occur often. Stress was found to positively predict anger in cancer patients (Lee et al., 2005). Stress is influenced by certain negative emotions such as fear or anger but not found to be related to certain others e.g. disgust (Lazarus,

1993). . Hence the occurrence of negative emotions cannot be equated to that of stress. Furthermore, there is no consensus on what comes under the discussion on negative emotions. There are different classifications of negative emotions presented in psychological studies.

Based on facial expressions, happiness, surprise, fear, anger, sadness, and disgust were identified as basic emotions, the first two being positive and the rest negative (Ekman et al., 1972). Another widely accepted psycho-evolutionary theory presents a circumplex in which basic emotions - anger, joy, sadness, trust, anticipation, fear, surprise, and disgust - interact with each other to generate more complex ones (Plutchik et al., 1980). More recently, 27 distinct emotions were identified and analysed as responses to short videos (Cowen & Keltner, 2017). Thus, it is necessary to study psychological stress separately from negative emotions because of two reasons. Firstly, there is no single scheme for emotion classification in psychological research. Secondly, psychological stress, while influenced by multiple negative emotions, cannot be simply equated to one or some of them. This points to the requirement for a lexicon specifically developed for stress scores of affect terms, which can be applied in more general datasets and domains.

TensiStrength (Thelwall, 2017) is the only existing lexicon for stress and relaxation measurement. It uses a list of terms annotated with strengths of stress and relaxation. The TensiStrength lexicon is a combination of terms from LIWC (Tausczik & Pennebaker, 2010), General Inquirer (Stone, et al., 1966), and the lexicon of the sentiment detection program, SentiStrength (Thelwall et al., 2010) with a few manual additions. Its basic approach is to assign each sentence/ tweet two scores: one to indicate the stress strength and one for the relaxation strength. This is based on the observation that the same sentence can contain expressions of both stress and relaxation. The scores of the highest stress term and highest relaxation term are taken as the sentence score, after being modified by rules considering issues like spelling, negation, and booster words. In the case of multi-sentence tweets, the highest stress/relaxation scores for any of the constituent sentences are assigned to the tweet. The sentences are identified by the presence of punctuation marks '.', '?' or '!'.

TensiStrength uses a scale of scores from -1 to -5 for stress and +1 to +5 for relaxation. -1 indicates an absence of stress and -5 indicates extreme stress. Similarly, +1 is absence of relaxation and +5 highest relaxation. This comes from the understanding that text can contain different degrees of psychological stress/relaxation expressed in it. Separate scales exist for stress or relaxation because the tasks of detecting stress and relaxation are not equivalent though related (Thelwall, 2017). The absence of stress does not equate to relaxation and vice versa. Hence separate scales were devised to indicate them. An alternative would have been to use binary scales for the presence or absence of stress and relaxation. However, this approach does not acknowledge the different degrees in which stress and relaxation can be possibly expressed in a given text.

The rules considered while assigning stress/relaxation values in TensiStrength are the following:

1. Two or more repeated letters increase the stress/relaxation values by 1. (scarryy has a higher stress value than scary.)
2. Idioms are treated as a single unit with assigned stress/relaxation value. The scores of individual words are ignored.
3. The negation of stress words neutralises them.
4. The negation of relaxing words turns them into stress words.
5. Emoticons are assigned appropriate stress/relaxation values.
6. Spelling correction to delete repeated letters to form words.

This lexical method was evaluated on a corpus of 3066 tweets which consists of 6 sub-corpora with emotion terms, insults, stress terms, opinions, common terms, and travel terms. Its performance is evaluated with respect to seven machine learning classification models with n-grams as features (Table 3.1).

**Table 3.1 Accuracy percentage of TensiStrength in detecting the stress scores, compared to the best performing machine learning classifiers (Logistic Regression and Support Vector Machines) on three different datasets**

| Corpus | TensiStrength | Logistic Regression | Support Vector Machine |
|---|---|---|---|
| Stress terms (n=655) | 36.5 | 48.8 | 49.1 |
| Travel (n=528) | 51.8 | 48.2 | 50.8 |
| Combined(n=3066) | 49.3 | 48.9 | 49.3 |

In the combined corpus of 3066 tweets, TensiStrength's accuracy for stress score prediction is the same as the SVM model and better than the Logistic Regression model. In the travel corpus, TensiStrength performs better than both the machine learning models. It performs worse than both the machine learning models in the stress terms corpus.

A key disadvantage of TensiStrength is that it depends solely on the occurrence of the listed stress or relaxation terms and linguistic cues such as negation, punctuation, repeated spelling, etc. However, contextual information about any stress/relaxation in the tweet is not utilized. The machine learning classifiers were trained on relatively small datasets, so they might outperform TensiStrength on other datasets. The study does not explore whether the better performance of TensiStrength on the travel subcorpus can be replicated in other topic-specific datasets.

Despite these limitations, the TensiStrength lexicon is still a significant research contribution as it is the pioneer and the only existing lexicon of stress/relaxation scores of tweets. It has been used in hybrid systems which considered the lexicon-assigned stress scores as features for a classifier, instead of taking it as the absolute stress score of the text and used it for the analysis of language of stressful posts in Facebook and Twitter (Guntuku et al., 2018).

Hence, attempts to improve the TensiStrength lexicon are relevant and could contribute to improving the task of identifying stress/relaxation scores of tweets. Since the lexicon contains ambiguous words which could have potentially different stress scores according to the context, sense disambiguation was explored as a potential way to improve the lexicon performance.

We used WordNet (Miller et al., 1991) as the source of the different senses of the ambiguous word because of its good performance in NLP research involving similar tasks. Incorporating word similarity information from WordNet has been shown to improve the performance of sentiment analysis measured by F-measure, by 2% in SVM-based classification model and 4% in Naïve Bayes-based model (Bellot, Hamdan & Béchet, 2013) Annotating WordNet synsets with positive or negative polarity has created a new lexical resource SentiWordNet (Baccianella, Esuli & Sebastiani, 2010) which has been used to improve the accuracy of sentiment analysis algorithms (Goel, Gautam & Kumar, 2016). WordNet follows standard English and excludes certain kinds of words which are characteristics of social media text such as forms of laughter (e.g. lol, rofl), repeated letters (e.g. haapppyyy, worrriiieed), acronyms, joint words, shortened words, and technology-related terms (Domingo, Gonzalez-Ferrero & Gonzalez-Dios, 2021). However, in our research, the usage of WordNet is limited to finding multiple senses of the stress terms and not for the identification of the term itself. Repeated letters and spelling corrections are handled by TensiStrength program as explained earlier.

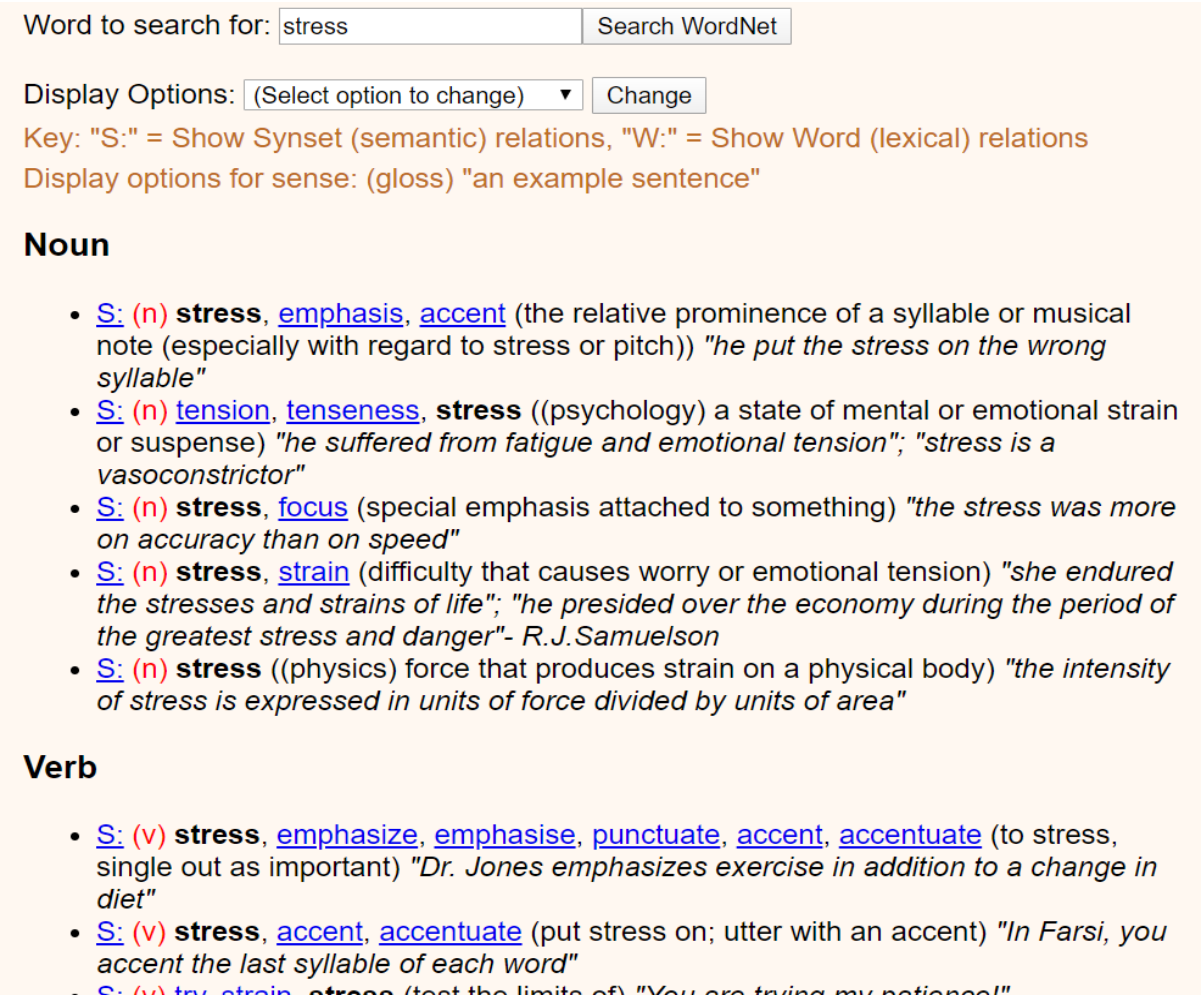


**Figure 3.1 Synset of the word 'stress' from WordNet**

WordNet is a lexical database of English that organizes words based on their semantic relations like hyponyms, meronyms, and synonyms. Words in WordNet belong to one of four categories: nouns, verbs, adjectives, and adverbs. Information about each word is organized as lemmas, gloss, and synsets. A synset of a word is the collection of words that could be used interchangeably. Each synset consists of one or more lemmas, each representing one sense of the given word. A gloss is the definition for a specific sense of the word, typically followed by one or a few short sentences illustrating the usage.

In Figure 3.1, the synset of the word ‘stress’ is given. The synset consists of several different senses of the word, each described by a gloss as follows:

S: (n) stress, emphasis, accent (the relative prominence of a syllable or musical note (especially about stress or pitch)) "he put the stress on the wrong syllable"

At the time of the experiments, vector embeddings was a relatively novel and promising method to represent words. The pivotal advantage of vector representations was its ability to embed both words and their inter-

relationships. Compared to traditional methods treating each word as a discrete unit, word embeddings created dense, distributed representations and hence were useful in modelling the semantic relationships between words. The word vectors could be mathematically operated (added or subtracted) to reveal analogies and similarities. A comparative study provided experimental evidence that employing word embeddings improved the performance of word sense disambiguation. Skip-gram embeddings consistently performed better in comparison to continuous bag of words representation. A disadvantage, though, of word vector representation was its context-independence. Specifically, a one-to-one representation of words and vectors does not sufficiently represent polysemous words (words with multiple meanings). Lately, there have been enhanced sense embeddings using the BERT representations of constituent words in WordNet glosses. The resultant representation ARES provided state of the results in the word sense disambiguation tasks. Before the invention of context-dependent vectors (e.g. BERT, ELMO) there were already attempts to add contextual information to word vectors.

A sense vector scheme obtained from skip-gram based word vectors and WordNet glosses has been proposed for WSD (Chen, Liu & Sun, 2014). The algorithm in this research works in two steps: initialization of word and sense vectors and sense disambiguation. In the first step, the word vectors are learned through the skip-gram model, and from this, sense vectors for each sense are constructed. For each word in the gloss, the cosine similarity with the original word is calculated. Words with cosine similarity higher than a threshold are added to a candidate set. The average of the vectors of the candidate words is taken as the sense vector.

In the second step of word disambiguation, given a sentence, an initial context vector is calculated by finding the average of content words' vectors. The sense which has the highest cosine similarity to the context vector is taken as the disambiguation result. We implemented this method of sense disambiguation with word vectors in our current experiments.

We chose this unified vector model or word sense representation and disambiguation for the following reasons:

1. It uses a predefined inventory of senses. In our implementation, we pre-computed these vector representations, before actual tweets' processing, thus minimizing the computational overhead.

2. This system comprehensively covers all senses in the standard lexical resource, WordNet, representing each by a corresponding vector. It makes use of the high-quality glosses in WordNet, thus forming the basis for word senses with reliable semantic contexts.

In this study, we focus on whether disambiguating the word senses improves the detection of stress/relaxation strengths. The comparison is with the existing methods of stress/relaxation strength detection and not with

those of word sense disambiguation. More recent methods for disambiguation or context representation could arguably provide better results, however, the relevance of the current experiments is in establishing that addressing context-based polysemy improves the performance of the TensiStrength method in comparison to the existing version.

## 3.2 Methods

### 3.2.1 Modified TensiStrength

A new Word Sense Disambiguation step was added to pre-process the ambiguous words in tweets before assigning stress/relaxation scores (Figure 3.2). For each ambiguous word in the given tweet, the correct sense was chosen from a list of possible senses from WordNet[3], using a word-vector based disambiguation method.

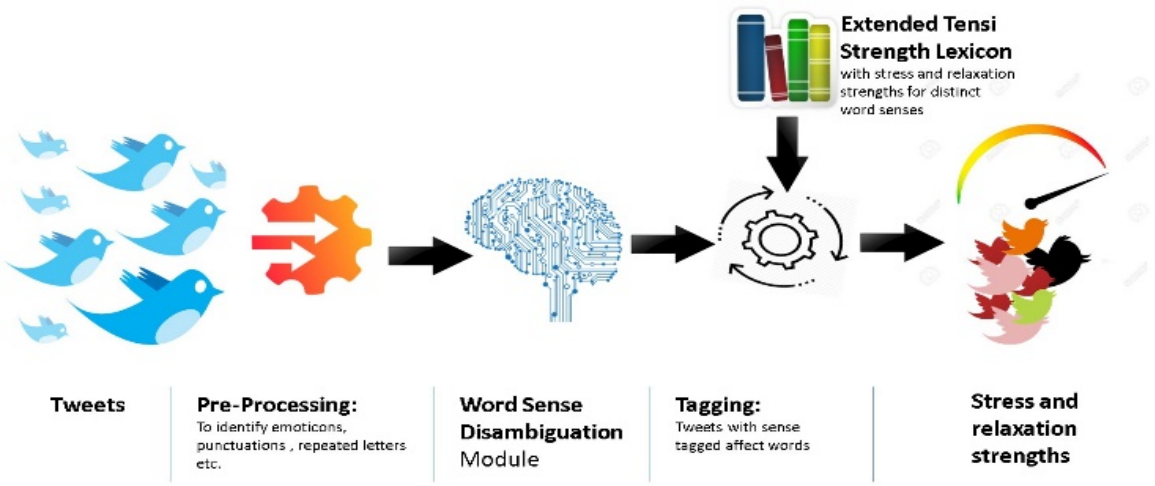


**Figure 3.2 Tweet processing Workflow in the modified TensiStrength algorithm (with WSD)**

We identified 40 ambiguous affect words in the existing TensiStrength lexicon and updated it to accommodate the stress/relaxation values for different senses of each of them. This experiment thus considered only ambiguous words in the TensiStrength lexicon as affect words (Table 3.2).

**Table 3.2 Ambiguous affect words with stress/relaxation scores updated in the modified TensiStrength lexicon.**

| | | | | |
|---|---|---|---|---|
| Fine | Crude | Fair | Block | Nervous |
| Concern | Dark | Tight | Fire | Breakdown |
| Chill | Wreck | Heavy | Strain | Fume |
| Late | Turbulence | Fraught | Radical | Noise |

[3] https://wordnet.princeton.edu/

| | | | | |
|---|---|---|---|---|
| Scene | Mad | Hard | Scheme | Accident |
| Artificial | Brazen | Callous | Desert | Coarse |
| Waste | Critical | Trick | Protest | Stress |
| Tax | Cold | Cool | Choke | Stick |

For example, the lexicon for TensiStrength without WSD had only one entry for the affect word 'cool' as +2, indicating moderate relaxation. But the modified lexicon acknowledges the various senses of the same word, obtained from the resource WordNet. For example, the ninth sense of the word 'cool' in WordNet is 'unfriendly or unresponsive or showing dislike' which has an indication for stress, hence it is assigned the value of -2. Similarly, different senses of each of these ambiguous words are assigned appropriate stress/relaxation values too. The ordinal number of the sense is represented by a suffix to the original word. For example, below are the new senses of the affect word 'fine', added to the lexicon. The gloss of the sense is given in brackets.

*fine_1 -3 (**fine**, mulct, amercement (money extracted as a penalty))*

*fine_2 -3 (ticket, **fine** (issue a ticket or a fine to as a penalty) "I was fined for parking on the wrong side of the street"; "Move your car or else you will be ticketed!")*

*fine_3 2 (all right, **fine**, o.k., ok, okay, hunky-dory, cool (being satisfactory or in satisfactory condition) "an all-right movie"; "the passengers were shaken up but are all right"; "is everything all right?"; "everything's fine"; "things are okay"; "dinner and the movies had been fine"; "another minute I'd have been fine")*

*fine_4 2 (**fine** (minutely precise especially in differences in meaning) "a fine distinction")*

Following the approaches adopted in the creation of the TensiStrength (Thelwall, 2017) and SentiStrength (Thelwall et al., 2010) lexicons, the stress and relaxation strengths to these word senses were assigned first intuitively and then modified using a hill-climbing approach. This assessed whether an increase or decrease of the stress/relaxation strength by 1 changed the overall accuracy of the classifications over the annotated dataset of tweets. If the change improved the overall accuracy by 2, then the change was kept. This process was repeated for each of the new addition to the TensiStrength lexicon until there were no more changes to the stress/relaxation strengths.

Thus, with WordNet as the source for the different senses of ambiguous words, the lexicon of TensiStrength was extended to include different scores for these senses rather than a single score for all senses.

**Table 3.3 Examples of tweets with stress/relaxation values assigned by TensiStrength with and without WSD (with affect (stress/relaxation) terms highlighted)**

| Tweet | TensiStrength without WSD | TensiStrength with WSD |
|---|---|---|
| Everyone in the valley has a **heavy** heart but the best thing to do was **worship** | (-3,+2) | (-3,+2) |
| Don't move **heavy** furniture by urself | (-3, +1) | (-2, +1) |
| She was wearing a **cool** pink dress | (-1,+2) | (-1,+2) |
| The lady gave me a **cool** stare | (-1,+2) | (-2,+1) |

In the first example (Table 3.3), there is one stress term ('heavy') and one relaxation term ('worship'). The corresponding stress and relaxation scores of these terms were assigned to the tweet. In the second tweet, there was one stress term, 'heavy', and no relaxation term. TensiStrength without WSD had the stress score for -3 for 'heavy', but TensiStrength with WSD identified the sense in which this ambiguous word ('heavy') was used in the specific context of the given tweet and assigned the corresponding stress score for that sense of the affect word 'heavy'. Since there is no other affect word, it was assigned as the stress score of the tweet too. Thus, TensiStrength with and without WSD differed in the stress score assignment to the tweet.

### 3.2.2 Dataset and Annotation

A set of one thousand tweets with the 40 ambiguous affect words (Table 3.2) was collected over a month (1st February 2017 to 1st March 2017) as the dataset for the experiments. The tweets were collected using the Tweepy[4] library, with ambiguous words as search keywords. From the tweets obtained from the search, duplicates, retweets, tweets with less than three words, tweets that consist of only hashtags and/or URLs were removed. Furthermore, 1000 tweets were randomly selected from this collection, 25 tweets for each of the 40 ambiguous words.

The tweets were annotated individually and independently by a set of three human coders on a five-point scale for stress and relaxation.

The coders had participated in an identical annotation procedure as part of the TensiStrength experiment (Thelwall, 2017). Hence, they had prior experience with the stress/relaxation strength annotation task. The coding instructions were similar to those provided in the preparation of the gold standard for the detection of stress and relaxation strength using TensiStrength.

[4] https://www.tweepy.org/

The coders were instructed to mark the perceived measure of stress in each tweet. The measures ranged from -1 (no stress perceived) to -5 (very high level of stress perceived).

**Guidelines:**

- Assign an appropriate stress score from -1 to -5 to the given tweet based on the level of stress present in it.
- Stress can be expressed as fear, worry, or anger about a specific incident/person or the perception about the general circumstance.
- Assign -1 if there is no expression of stress in the tweet.
- Assign -5 if there is a very high level of stress in the tweet.
- Assign a score between -2 and -4 if there is some expression of stress but not very high.
- Use individual judgement while assigning the stress score

The coders also had to mark the perceived measure of relaxation for each tweet. Relaxation was defined as either an absence of stress or the expression of the ability to overcome stress. Similar to stress measures, the relaxation scores ranged from +1 to +5. +1 was to be given for tweets with no relaxation expressed, +5 for tweets with a very high level of relaxation, and scored from +2 to +4 for tweets with intermediate states of relaxation between these two extremes.

**Guidelines:**

- Assign an appropriate relaxation score from +1 to +5 to the given tweet based on the level of relaxation expressed in it.
- Relaxation could be the mention of a relaxing activity/context, expression of general well-being, or indication for the absence of stress.
- Assign +1 if there is no expression of relaxation in the tweet.
- Assign +5 if there is an expression of a very high level of relaxation in the tweet.
- Assign a score between +2 and +4 if there is some expression of relaxation but it is not very high.
- Use individual judgement while assigning the exact relaxation score.

The arithmetic mean of the three coders' annotation scores was calculated for each tweet and rounded off to the nearest integer value. This was taken as the gold standard of stress and relaxation values. Krippendorff's Alpha (α) (Krippendorff, 2004) was used as a measure of the agreement between annotators. It is a reliability coefficient for assessing the representational value of data, when multiple values of the same object have been generated using different methods. α is defined as:

$$\alpha = 1 - (D_0/D_e)$$

where $D_0$ is the observed disagreement

$D_e$ is the disagreement expected by chance.

The value of α lies between 0 and 1. A Krippendorff's α value of 0 indicates unreliability of value assignments and α value of 1 indicates perfect reliability. A value greater than 0 but less than 1 indicates better agreement than what can be expected by chance- disagreements are present but they are systematic.

Krippendorff suggests the lowest conceivable limit of $\alpha > 0.667$. The overall agreement between the three coders for stress (0.730) and relaxation (0.754) was found to be high enough.

**Table 3.4 Inter-coder agreement (Krippendorff's α) for stress annotation for the current experiment and the identical task performed for earlier version of TensiStrength**

| Research study | A and B | B and C | A and C |
|---|---|---|---|
| TensiStrength with WSD | 0.774 | 0.781 | 0.771 |
| TensiStrength without WSD (Thelwall, 2017) | 0.730 | 0.714 | 0.705 |

**Table 3.5 Inter-coder agreement (Krippendorff's α) for relaxation score annotation for the current experiment and the identical task performed for earlier version of TensiStrength**

| Research Study | A and B | B and C | A and C |
|---|---|---|---|
| TensiStrength with WSD | 0.787 | 0.802 | 0.781 |
| TensiStrength without WSD | 0.767 | 0.725 | 0.727 |

Higher agreement values in the current experiment, compared to the identical task done for an earlier research study (Thelwall, 2017) suggest that the coders are consistently improving with the annotation task (Table 3.4, Table 3.5). It also indicates that the problem statement and annotation instructions are well-defined.

### 3.2.3 Experimental Setup

A Twitter Word2Vec (Mikolov et al., 2013) model trained on 400 million tweets, released as part of an ACL W-NUT task (Godin et al., 2015) was used for training the word and sense vectors for this experiment. ACL W-NUT focussed on NLP tasks on noisy, user-generated text in social media. Specifically, two shared tasks hosted by the workshop were Named Entity Recognition in Twitter and Normalization of Noisy Text. The former task

required the participants to detect named entities belonging to one of the ten categories (company, person, movie, geolocation, etc.) from tweets. Though created originally to address this NER task, the Word2Vec model trained on the Twitter corpus of 400 million tweets has been shown to perform well in various other tasks such as the identification of cyber-bullying (Zhao, Zhou & Mao, 2016), election-related tweets (Yang, Macdonald, & Ounis, 2018) and traffic events. For each word in the ambiguous words list mentioned in the previous section, we calculated vectors of dimension 200, corresponding to each different sense, treating WordNet as the sense inventory. A WSD module (Chen, Liu & Sun, 2014) was implemented using Anaconda Python (versions, Anaconda-4.4.0 and Python-3.6.1) packages Scipy, Numpy, and nltk, to evaluate its stress and relaxation strength detection performance.

## 3.3 Results

### 3.3.1 Performance of TensiStrength with WSD

The accuracy of stress and relaxation detection in tweets using TensiStrength with WSD was compared with TensiStrength without the WSD pre-processing phase and a range of standard machine learning algorithms: Adaptive Boosting algorithm (AdaBoost), Naïve Bayes classifier (Bayes), Decision tree (J48 Tree), Logistic Regression (Logistic) and Support Vector Machines (SVM). These were chosen to represent a range of different machine learning algorithms, as used in the original TensiStrength paper. Another approach was to formulate the problem as regression. The original TensiStrength program was modelled with distinct classes as stress/relaxation levels. As the focus of our experiments is to compare the improvement with WSD over the original TensiStrength program, we continue the classification approach. However, it is also interesting to see if a regression approach improves the stress/relaxation score detection performance. A regression model can also take advantage of the fact that 1 is closer to 2 than to 3, for example. So, we also implemented a linear regression model for comparison with the modified TensiStrength. The predicted value was rounded off to the nearest integer to evaluate the performance metrics. For all the models, unigrams, bigrams, and trigrams were used as features, again as used in the comparison with TensiStrength without WSD (Thelwall, 2017). Each classifier was implemented using its configuration in Weka 3.6. The performance of the machine learning algorithms was evaluated using 10-fold cross-validation 30 times, with the average scores across the 30 iterations recorded.

Based on Pearson correlations and exact match percentages with the human-annotated data, TensiStrength with a pre-processing phase to disambiguate the word senses performs considerably better than the selected machine learning algorithms and TensiStrength without WSD (Table 3.6, Table 3.7). The exact match % indicates the percentage of cases in which the stress score from the model exactly matched the human-annotated score. Match % (within 1) indicates the percentage of near-misses, in which the model-assigned stress score was +/- 1 (differed by 1) with the human-annotated score.

**Table 3.6 Performance of TensiStrength with WSD with respect to the other methods in detecting stress score for the human annotated set of 1000 tweets containing one of the 40 ambiguous terms**

| Method | Pearson correlation | Exact match % | Match % (within 1) |
|---|---|---|---|
| AdaBoost | 0.35 | 30.2 | 83.2 |
| Bayes | 0.32 | 35.3 | 90.2 |
| Logistic | 0.49 | 49.3 | 89.2 |
| SVM | 0.51 | 51.3 | 91.6 |
| Linear Regression | 0.52 | 52.1 | 91.8 |
| TensiStrength without WSD | 0.47 | 48.8 | 83.2 |
| TensiStrength with WSD | 0.54 | 53.10 | 92.41 |

**Table 3.7 Performance of TensiStrength with WSD with respect to the other methods in detecting relaxation score for the human annotated set of 1000 tweets containing one of the 40 ambiguous terms**

| Method | Pearson correlation | Exact Match % | Match % (within 1) |
|---|---|---|---|
| AdaBoost | 0.32 | 34.13 | 83.54 |
| Bayes | 0.35 | 38.92 | 84.67 |
| Logistic | 0.49 | 54.32 | 89.17 |
| SVM | 0.55 | 58.73 | 91.65 |
| Linear Regression | 0.55 | 59.43 | 93.01 |
| TensiStrength without WSD | 0.53 | 56.38 | 85.83 |
| TensiStrength with WSD | 0.56 | 60.69 | 93.12 |

The WSD phase substantially improves the stress/relaxation detection accuracy in terms of Pearson's correlation, exact match percentage, and match percentage within 1. TensiStrength with WSD outperforms the machine learning methods as well. But the regression model performs better than the classification models. The small size of the annotated dataset (and the training data) could have been a reason for the relatively poor performance of the machine learning methods. Further results with more tweets could better evaluate the effectiveness of TensiStrength with WSD over standard machine learning methods as a stress and relaxation strength detection method. Also, tweets in the current dataset contain the ambiguous words added to the TensiStrength lexicon. The performance on a randomly chosen set of tweets could be different and the performance difference would be much smaller.

### 3.3.2 Domain Analysis of Stress for Tweets

We manually classified the tweets into the following major constituent domains to obtain extra context about the results: Climate, Entertainment, Food, Health, Personal Events, Politics, Sports, and Travel (

Table 3.8).

**Table 3.8 Stress detection performance of TensiStrength with and without WSD in different domains**

| Domain | % of tweets | Mean Stress value | Pearson correlation for TensiStrength (Without WSD) | Pearson correlation for TensiStrength (With WSD) |
|---|---|---|---|---|
| Climate | 11% | -1.46 | 0.485 | 0.512 |
| Entertainment | 8% | -1.19 | 0.487 | 0.523 |
| Food | 6% | -1.02 | 0.365 | 0.376 |
| Health | 9% | -1.31 | 0.418 | 0.452 |
| Personal | 18% | -4.23 | 0.501 | 0.532 |
| Politics | 17% | -4.01 | 0.567 | 0.603 |
| Sports | 12% | -1.92 | 0.452 | 0.597 |
| Travel | 19% | -3.95 | 0.482 | 0.523 |

The highest percentage of tweets belong to personal events, politics, and travel categories. The rest of the tweets are almost uniformly distributed across the remaining 8 domains. WSD improves the performance of the TensiStrength algorithm in tweets belonging to all domains.

Politics, personal events, and travel have higher stress values for the tweets, and food and entertainment are domains with the lowest stress values. This agrees with the traditional psychology studies discussed in Section 1.1.2 which put forward politics, personal events, and travel as the stress-prone domains.

## 3.4 Discussion

The experiments illustrate the improvement of the TensiStrength lexicon by incorporating different stress/relaxation scores for different senses of the same word. The modified lexicon can improve the stress detection accuracy of TensiStrength in direct and hybrid applications (e.g., (Guntuku et al., 2018)). The tweets have higher values of stress compared to relaxation. The mean value of stress expressed across all domains is -2.81 whereas the mean value of relaxation expressed is +1.72. Since the results are matched to human judgements of the strength of these affective dimensions (for the evaluation of tweets), there are multiple possible explanations. The queries used to generate the tweets might have matched high-stress tweets better than high relaxation tweets, which seems likely. Alternatively, TensiStrength might be more effective at identifying high-stress tweets than at identifying high relaxation tweets. The inherent nature of stress as a high arousal state and relaxation as a low arousal state (Thelwall, 2017), could also be a possible explanation of higher mean stress score compared to the mean relaxation score.

### 3.4.1 Error Analysis

The misinterpreted stress/relaxation expressions occur in tweets with no direct affect words or affect words with context words misleading the WSD module (Table 3.9). In the first category, we can find tweets such as 'I am a train wreck' or 'I had a heavy heart to carry'. The correct (human-annotated) stress score for 'I am a train wreck' is -3 and for 'I had a heavy heart to carry' it is -4. However, the WSD module incorrectly disambiguates the sense of the ambiguous affect word (wreck, heavy) because of the context words train and carry, and assigns a lower stress score of -2 to both. TensiStrength without WSD gives the scores solely dependent on the lexicon scores of the affect words present.

**Table 3.9 Examples of errors by TensiStrength with WSD in identifying stress and relaxation strengths correctly in the human annotated dataset of 1000 tweets**

| Tweet | Humans | TensiStrength (without WSD) | TensiStrength (with WSD) | Affect Word | Context Word |
|---|---|---|---|---|---|
| I am a human train wreck | (-3, +1) | (-3, +1) | (-2, +1) | Wreck | Train |
| I had a heavy heart to carry | (-4,+1) | (-2, +1) | (-2, +1) | Heavy | Carry |
| Great night mad crowd, thanks everyone | (-1, +3) | (-1, +2) | (-3, +1) | Mad | Crowd |
| Antioxidants eliminate destructive potential of free radical | (-1, +2) | (-2, +1) | (-2, +1) | Radical | Destructive |

Another source of error is indirect expressions of stress and relaxation. Examples in our dataset are 'Sat by the fire; little things in life' (correct stress relaxation score is (-1, +4), our system score is (-2, +1)). 'Straining ears for a laughter clip; but hearing nothing' (correct score (-2, +1) our system score (-1, +3)). Such indirect expressions are challenging for all automated systems in finding the sentiment or stress strengths.

### 3.4.2 Limitations

The results showed that word sense disambiguation improved the performance of the TensiStrength lexicon for identifying stress and relaxation in the dataset of tweets with ambiguous words.

However, the study has limitations. Firstly, the dataset was limited to 1000 tweets and a bigger dataset could have improved the performance of machine learning models, in comparison. Secondly, the presence of ambiguous words was the search criteria for the tweets and the relative performance improvement of the modified TensiStrength would be much smaller on a random dataset because most tweets would not contain ambiguous words. In a dataset of 22314 tweets, collected using domain-specific hashtags (described in Chapter 4), 4.48% had any one of the 40 ambiguous stress/relaxation terms. This corpus was collected from four domains – politics, airlines, traffic, and personal events. The frequency of these ambiguous terms in a random dataset is yet to be verified.

A third limitation is the relatively simplistic design of the machine learning models. Advanced features in addition to the n-grams might have resulted in their better performance. In addition, the improved

TensiStrength is much slower than the original version. Here the improved accuracy is bought by a substantial increase in the amount of processing needed to disambiguate terms in tweets.

Since the key objective was an improvement in the original TensiStrength lexicon, the baseline machine learning methods were chosen the same as those in the original TensiStrength study. These methods had word n-grams as features. Extended features based on character n-grams and punctuation could have resulted in better performance by the baseline methods. Also, since the time of these experiments, models like FastText and BERT-based solutions have presented a better performance in similar tasks. In the light of these advancements, a comprehensive analysis of the improvement of modified TensiStrength as a stress-detection method would require updated features and methods for a fair baseline. However, in comparison to the existing TensiStrength lexicon, the modified one still offers an improvement for stress/relaxation magnitude in tweets with ambiguous affect words.

The stress and relaxation strengths of senses of the ambiguous words were manually assigned initially and then modified based on a hill-climbing approach as described in Section 3.2. However, with more ambiguous words, this initial score assignment will require further manual effort, which might hamper the scalability of the approach.

Furthermore, the improvement approach only considers ambiguous stress/relaxation words which are part of the TensiStrength lexicon. Words which are not polysemous can still indicate various degrees of stress/relaxation depending on the context. However, this contextual disambiguation is not performed while assigning the stress/relaxation score using TensiStrength, in both the original and the modified versions. Words other than those included in the stress/relaxation lexicon can still contain potential cues about the context of the tweet and the stress/relaxation experienced. This is a limitation of the current approach.

## 3.5 Conclusions

This chapter introduced a pre-processing step for the existing TensiStrength program using word sense disambiguation and demonstrated that it improves accuracy on tweets containing one of the 40 ambiguous words examined.

TensiStrength is the only existing lexicon for assigning stress and relaxation scores for social media text. Even though it gives a comparable performance with respect to machine learning classifiers, its approach is rather simplistic with identifying affect words in the given text. Since it does not consider the contextual information of affect words, word sense disambiguation seemed to be an intuitive step to improve its performance.

Using WordNet as the sense repository, we modified the TensiStrength by assigning separate stress and relaxation scores to different senses of 40 ambiguous affect words in the current lexicon. The new stress and

relaxation scores were first assigned intuitively and were further improved using a hill-climbing approach. A vector-based system for word sense disambiguation was used to disambiguate the ambiguous words in the tweets. Once the correct sense was identified, the corresponding stress/relaxation score was assigned to it using the modified TensiStrength lexicon.

This modified approach was tested on a dataset of 1000 tweets which were collected using the Twitter Search API with the ambiguous words as keywords. The dataset consisted of 25 tweets for each of the 40 ambiguous affect words in the modified TensiStrength lexicon. The accuracy of TensiStrength with WSD was compared to TensiStrength without WSD and four machine learning classifiers. The modified approach was shown to give superior performance in comparison with the other methods (an increase of 4.30% of accuracy over TensiStrength without WSD and 1.80% over SVM which was the best performing machine learning classifier). The dataset was manually classified into 11 broad categories. The mean stress scores of politics, travel and personal events were relatively higher than the other categories. This is in agreement with the traditional psychology observations of these domains as stress-prone. These findings motivate us to choose these domains as the scope of our further studies of high-stress tweets, which are described in Chapters 4, 5, and 6.

The method, though illustrated to be successful, had a few limitations. The dataset was limited in number and the search criteria; in a different dataset, the exact performance of the modified TensiStrength might vary. Also, with advanced feature selection, the compared machine learning classifiers could have given a better performance. Even with these limitations, the improvement to TensiStrength is still significant as it is the only existing stress score lexicon and has been a source in other studies of psychological stress in social media.

# Chapter 4

# FINDING STRESSORS FROM TWEETS

Social media can be harnessed to discover trends in group or individual emotions and moods. Although previous studies have developed methods to detect stress in social media, the stressor also needs to be known so that remedial actions can be targeted more effectively. For example, while analysing Tweets about traffic in a city, it would be beneficial for the traffic authorities to know whether congestion or accidents are the stressors in most tweets. Similarly, for businesses, it is critical to find stressors in the social media responses of customers to take appropriate remedial action. The existing literature focuses on long-term reasons for psychological stress, arising from life events such as divorce, illness, and the death of loved ones. Reasons for short-term stress (e.g. for traffic, political unrest, customer services) are largely unexplored.

In response, this thesis implements a novel framework for finding the stressors expressed in tweets. This chapter introduces a method to classify stressors from tweets belonging to four domains: politics, life events, traffic, and airlines. Stressors differ according to the context - for customers of airlines, it could be abrupt cancellations, long delays, and bad service from the crew, whereas, for political tweets, it could be policy changes, events, economy, and elections. This prompted this research to separate tweets according to their domains for studying the reasons.

We collected tweets in four domains: traffic in London, airlines, UK politics, and personal events. As a pre-processing step, a list of potential stressors in each domain was found by topic modelling and k-means clustering. Three different word-vector based methods were then applied to tweets to select a stressor from the stressors list. The accuracy of these methods was evaluated by comparing them with stressors selected by human annotators.

The rest of the chapter is organized as follows: Section 4.1 presents the background of related research. Section 4.2 describes how the dataset was collected and annotated. The framework to find the stressors from tweets is presented in Section 4.3. The results and discussion are in Sections 4.4 and 4.5 respectively. The chapter concludes with a summary in Section 4.6.

## 4.1 Related Work

Though there have been a few studies on psychological stress identification from social media, very little focus has been given to the stressors, the factors which induce stress. As discussed in Section 2.1.2, stressor events in Sina Weibo, widely considered as the Chinese equivalent of Twitter, have been identified on the basis of the personal event categories listed in the life events stress scale (Holmes & Rahe, 1967). However, in this study (Lin et al., 2016b), only the events in the course of life are considered as stressors. Domains like travel (traffic

and airlines), personal events, and politics are found to often induce stress, as established by traditional psychology studies using questionnaires and surveys. However, the identification of stressors from social media belonging to these domains is an unexplored research direction. The research goal of this chapter is to construct a generic framework which can be employed to identify stressors from multiple domains.

As discussed in chapter 1, the stressor is the stimulus inducing stress. The definition for stressor in psychological parlance is that it is the actual or perceived threat to the organism (Selye, 1956). For the purpose of this study, we clarify this definition as follows: "Stressor is an event, an experience, an interaction or an element in the environment which induces stress expressed in the text." Since we collected tweets based on domains, rather than simply extracting the targets of opinions in each tweet, the objective was to cluster stressors to create a collective understanding of what factors induce stress in each examined domain. This approach is useful in identifying stress-prone situations in each domain and taking remedial actions.

The tweets in the dataset were collected based on domain-related hashtags and not on stress declarations ("I am stressed because", "it is stressful" etc.). As the tweets collected in our datasets do not have explicit causality markers, the inherent topics were relied upon to extract the stressors. The challenges of applying traditional topic modelling techniques to social media text were described in Section 2.2.3. Pooling of tweets based on hashtags or authors has been shown to improve the performance of the topic modelling methods (Mehrotra et al., 2013). Based on the search criteria of dataset collection, we chose hashtag pooling to improve topic modelling.

Topic modelling has been used in tweets for a variety of applications. Topics in Twitter conversations together with the sentiment polarities have been studied to identify communities in the social network (Naskar et al., 2016). The framework Sent_LDA shows that the topic distribution collects users together in sub-communities which display similarities in sentiments. However, the sentiment analysis considers broad categories of positive and negative; finer sentiment categories or the addition of the category of 'neutral' could have resulted in a different assortment of communities. Sentiment analysis of topics extracted from Yahoo! Finance message boards using an LDA-based system called TSLDA was used to predict stock price movements (Nguyen & Shirai, 2015), though the system was a parametric model which required the number of topics beforehand. LDA topic modelling has also been used to identify events in Kenya as part of the Umati project to detect hate speech (Sokolova, 2016).

Thus, the existing works demonstrate the usability of topic modelling, especially LDA in understanding sentiments and social connectivity in tweets. In the present study, LDA topic modelling is used together with clustering to identify stressors from tweets.

## 4.2 Dataset

### 4.2.1 Twitter as a Dataset Source

The corpus used in the experiments carried out in this research was collected from Twitter.com. As of 2020, Twitter had 187 million monetizable daily active users[5] (the number of unique users logged in per day to whom Twitter could show advertisements) sending 500 million tweets per day. Moreover, 80% of Twitter users access it from mobile devices[6] so it can be used for real-time updates of daily events or opinions. Tweeters can post messages of up to 280 characters and this concise format promotes brevity. The brevity of the messages might limit the availability of contextual information, however, Twitter users get around it by multiple, related tweets. Thus, the brief nature of messages encourages more updates per day than for traditional blogs (McCormick et al., 2017). This makes Twitter a suitable option for studying mental reactions to transient situations like traffic congestions, airport hassles, and competition updates.

**Table 4.1 Methods used to collect tweets in previous studies on mental health and personality traits using Twitter**

| Author | Method | Tweets |
|---|---|---|
| Wang, 2015 | GNIP API from Twitter and Stanford Data Science Initiative | 120k |
| Liu et al., 2016 | Twitter REST APIs | 1 billion |
| Gamon et al., 2013 | Twitter Firehose | 2 million |
| Coppersmith et al., 2014b | Twitter APIs with diagnosis regular expressions | 21.8k |
| Choudhury, Counts & Horvitz, 2013 | Twitter Firehose with phrases occurring in newspaper birth announcements as search criteria | 77.3k |
| Coppersmith et al., 2014a | Twitter APIs with diagnosis regular expressions | 2 million |

Twitter data is typically obtained from one of three sources (Table 4.1).

1. Twitter Search API
2. Twitter Streaming API
3. Twitter Firehose

The Twitter search API allows users to 'pull' tweets based on search criteria. These criteria can be specific keyword(s), hashtag(s), or a time period, location, or usernames. Search API results are limited to the tweets from the previous week. In contrast, the Streaming API provides near real-time access to tweets, 'pushing' them to the users as they are created without the need for search criteria. The user can still specify search criteria to filter the tweets the API returns. However, the API only provides a sample of the tweets produced in real-time. The Twitter Firehose is the third method, which is similar to the streaming API but provides 100%

[5] https://blog.hootsuite.com/twitter-statistics/
[6] https://www.omnicoreagency.com/twitter-statistics/

of the tweets instead of a chosen sample. It is a paid service provided by GNIP[7] and Datasift[8]. For this research, Twitter Search API was used because it enabled the collection of tweets based on specific search criteria from the domains under study.

### 4.2.2 Dataset Collection

We performed an analysis of stress scores of tweets, and how word sense disambiguation improved the accuracy of stress score predictions (Chapter 3). This gave average stress scores for a range of domains (Table 4.2).

Based on these results, and the previous literature, discussed in Chapters 1 and 2, showing travel, politics, and personal events to be domains often causing psychological stress, we chose these as the focus of this thesis.

**Table 4.2 Mean stress scores by domain for the tweets used in word sense disambiguation experiments**

| Domain | Mean Stress |
|---|---|
| Personal | -4.23 |
| Politics | -4.01 |
| Travel | -3.95 |
| Sports | -1.92 |
| Climate | -1.46 |
| Health | -1.31 |
| Entertainment | -1.19 |
| Food | -1.02 |

Words or #hashtags can be used to identify and collect messages with the same central topic. The strategy for dataset collection was to search for tweets containing selected keywords/hashtags using Tweepy[9]. Tweepy is an open-source python library which enables Python to use Twitter searching APIs. It interacts with Twitter API using an authentication method called OAuth. The 'cursor' object in Tweepy facilitates searching for tweets matching the specified search criteria of keyword(s)/hashtag(s) and time interval.

To collect political tweets, we used hashtags specific to political parties and leaders in the UK. Similarly, hashtags with traffic-related terms and those typically used to denote any one of 6 major motorways passing through London were used to collect tweets about London traffic. Though we limited the scope of the study geographically, for the sake of brevity, the domains are referred to as politics and traffic in the rest of the thesis (in the place of UK politics and London traffic). Similar to the approach by Dickinson, et al., (2016), we considered five life events (starting school, falling in love, getting married, having a child, and death) for constructing our dataset for personal events. We used hashtags related to these events as search criteria. For

[7] https://support.gnip.com/

[8] https://datasift.com/

[9] https://www.tweepy.org/

airlines, a list of 20 hashtags based on multiple ranking lists[10] of best and worst airlines of 2017 was constructed and those which returned very few tweets were removed.

Out of the 30171 tweets collected with the selected hashtags (Table 4.3), the following were discarded:

1. Retweets
2. Duplicate tweets
3. Tweets consisting of less than 3 words
4. Tweets containing only hashtags and/or URLs

After removing these and manually curating the dataset of tweets which did not belong to the domains under discussion, the dataset consisted of 22,314 tweets.

**Table 4.3 Dataset collection: Hashtags used as queries and number of tweets collected**

| Domain | Hashtags | Dates | Tweets |
|---|---|---|---|
| Politics(UK) | #ukpolitics, #brexit, #conservatives, #tories #theresamay, #uklabour, #jeremycorbin, #NHS, #tories, #corbyn, #ukip #borisjohnson | 01-07-2018 to 31-07-2018 | 7154 |
| Traffic (London) | #londontraffic #traffic AND #London #m25 #m40 #m1 #m3 #m4 #m11 #m25 #m40 | 01-07-2018 to 31-07-2018 | 6527 |
| Airlines | #virginatlantic, #united, #turkishairlines, #swissairlines, #singaporeairlines, #ryanair, #qatarairways, #lionair, #klmairlines, #emirates #americanairlines | 01-04-2018 to 01-05-2018 | 8217 |
| Personal events | #wedding, #marriage #relationships #children #death #deathofaparent #grief #family #childbirth | 01-07-2018 to 31-07-2018 | 8273 |

Examples:

(1) #AmericanAirlines still not helping with lost baggage, no word on compensation for missed flight— 2 days late to cruise!!

(2) I will never fly on @AmericanAir ever again! Their Miami Customer Service Department is absolutely horrible!!!… https://t.co/WCvWgYpnzA

(3) London at a standstill again #londontraffic

(4) Wow you bore me and I am only reading your tweets, I feel sorry for your fat #family looking at your grotting face every day

### 4.2.3 Twitter Dataset Annotation

[10] https://www.forbes.com/sites/larryolmsted/2017/12/21/best-worst-of-aviation-2017-airlines-and-alliances
https://www.businessinsider.nl/worst-airlines-in-the-world-in-2018-ranked-airhelp-score-2018-6/

#### 4.2.3.1 Overview

The dataset of 30171 tweets was filtered and manually curated; from the resultant 22314 tweets, we selected 12000 tweets randomly with 3000 tweets from each of the four domains (airlines, traffic, politics, and personal events). For the tasks described in this section, three human coders were chosen for data annotation, different from the coders who did the previous annotation in Section 3.2.2. These 12000 tweets were manually annotated for stress scores on a scale from -1 to -5. It was found in the annotation that 8871 tweets had stress scores -2 to -5, which meant they had an expression of stress. The remaining 3129 tweets had a stress score of -1 which indicated the absence of stress. The 8871 tweets with stress expressions were annotated for the stressor. From the dataset of 12000 tweets, 2000 were randomly chosen (500 from each domain) and were annotated for the following three:

1. Presence of swearing/socially offensive language
2. Presence of sarcasm
3. Temporal intent

A detailed description of the annotation guidelines and the inter-annotator agreement for the stress strength and stressor is given in Section 4.2.3.2. The guidelines and agreement of the socially offensive language, sarcasm (Chapter 5), and temporal intent (Chapter 6) are given in the corresponding chapters.

The protocol to resolve conflicts was decided before the annotation process. For stress strength, the final score was the arithmetic mean of the scores assigned to the tweet by the three annotators. Hence there was no risk of unresolved conflicts. In the annotation task of marking stressors in the tweet, in case of a disagreement, the stressor which two of the annotators agreed upon was taken as the final annotation. There were no tweets for which all three annotators chose different stressors. For socially offensive language and sarcasm, the majority rule resolved the conflicts since there were two categories (offensive/non-offensive, sarcastic/non-sarcastic) and three annotators. While annotating temporal categories, majority rule was used to resolve conflicts. Tweets where all three annotators differed in the temporal intent class were put in the atemporal class since no consensus could be reached in that case.

#### 4.2.3.2 Annotation Guidelines

From the 22314 tweets collected, 3000 tweets were chosen randomly for each of the four domains. This dataset of 12000 tweets was annotated by a set of three coders for the stress strength.

The first step was to find tweets containing expressions of stress. The annotators assigned a stress score in the range of -1 to -5 for each input sentence, -1 denoting the least stress and -5 denoting the highest stress. The annotator instructions were the same as those given during the annotation process described in Chapter 3 for

the identical task on a different dataset (See Section 3.2.2). The tweets were not assigned relaxation scores since the rest of the study focused only on the stress expressions. The mean of the stress scores by the three annotators, after rounding to the nearest integer value, was assigned as the stress score for the tweet. Tweets scoring from -2 to -5, therefore, contain expressions of stress (Table 4.4).

**Table 4.4 Number of tweets from each domain in the stress corpus (total number of tweets 8871)**

| Domain | Share of dataset (tweets) |
|---|---|
| Politics (UK) | 21.7% (1927) |
| Traffic (London) | 29.5% (2613) |
| Airlines | 28.1% (2493) |
| Personal Events | 20.7% (1838) |

The annotators were further instructed to mark the stressor for each tweet in this dataset (of 8871 tweets). For these annotations, the inter-annotator agreement is reported in Section 4.2.3.3.

**Annotation Question**

Which among the list of possible options, most accurately describes the stressor in the tweet?

**Explanation**

The annotators were provided a tweet with a stress score from -2 to -5, and a possible list of stressors (corresponding to the domain of the tweet), constructed from the methods described in Section 4.3. They were required to find one among this potential list of stressors which can be most appropriately described as the stressor in the given tweet.

**Guidelines**

- For the given tweet, mark the stressor (from the given list of options) which most correctly describes the stressor.
- A stressor is the stimulus behind stress. It can be an event, an experience, a personal interaction, or an environment which induces the stress expressed in the tweet.
- The stressor could be directly mentioned or indirectly indicated in the tweet.
- In the absence of an exact match, choose the stressor which suits the best among the given options.

Example

4. I was so nervous before the test started that I actually forgot my own birthday

   *Possible stressors* Death, Relationships, Marriage, Children, School

*Stressor* School

Conflicts in choosing the stressor were resolved by majority voting. There was no case in which all three annotators differed in the annotation of stressors.

From the 12000 tweets annotated for stress score, 2000 were randomly chosen (500 from each domain). For each tweet in this dataset, the annotators were required to mark the presence of swearing, the presence of sarcasm, and temporal intent. The human-annotated labels after resolving conflicts by majority voting are considered as the gold standard.

#### 4.2.3.3 Annotation agreement

The inter-annotator agreements in terms of Krippendorff's alpha for the annotation of stress strength and stressor are given in table 4.5. The high annotator agreement establishes the suitability of using the annotations as the gold standard.

**Table 4.5 Inter-annotator agreement of dataset for various annotation tasks (stress strength (12000 tweets), stressor (8871 tweets)**

| Annotation task | Domain | Between A and B | Between B and C | Between A and C | Overall agreement |
|---|---|---|---|---|---|
| Stress strength | Combined | 0.821 | 0.791 | 0.834 | **0.721** |
| Stressor | Politics | 0.795 | 0.832 | 0.789 | **0.692** |
| | Traffic | 0.772 | 0.741 | 0.786 | **0.701** |
| | Airlines | 0.782 | 0.784 | 0.762 | **0.684** |
| | Personal events | 0.802 | 0.755 | 0.783 | **0.672** |

## 4.3 Methods

The method introduced in this chapter identifies the stressors expressed in tweets in two steps. A list of potential stressors is constructed for each domain under consideration. The next step selected stressors expressed in each tweet from this pre-defined list of potential stressors for the corresponding domain. The tweets were collected by the Tweepy API using specific search hashtags, as described in Section 4.2. Tweets with high-stress scores (from -3 to -5), as judged by TensiStrength, were considered for creating the list of potential stressors. These high-stress tweets were subjected to LDA topic modelling and k-means clustering to find clusters of frequently occurring topics. The resultant topic clusters were processed by their word-vector representations to generate title words to describe each cluster. The title words constituted a list of potential stressors for the tweets of that category. The tweets were processed by three new word-vector based methods to find a reason for the stress expressed within them. These were compared with reasons found by human coders to evaluate the accuracy of our proposed methods.

### 4.3.1 Constructing Potential Stressors' List for Domains

The first step was to form a list of potential stressors in a given category/domain.

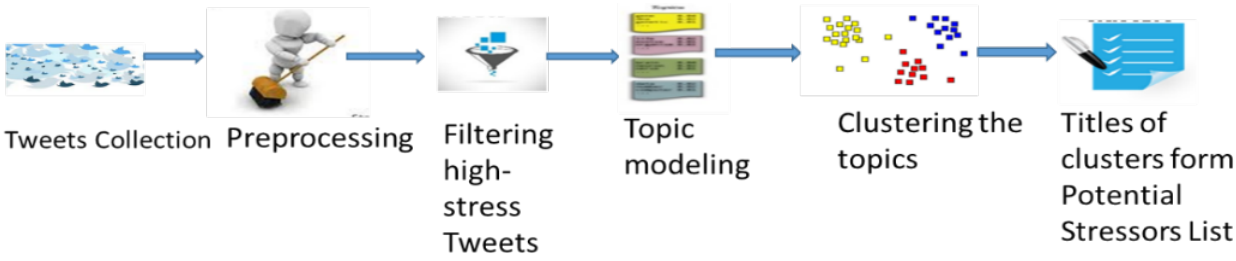


**Figure 4.1 Overview of the method to find a list of potential stressors from a set of tweets within a domain**

The tweets were pooled on the basis of hashtags to improve topic modelling. They were pre-processed through the following steps.

**Pre-processing:** Tweets are strings consisting of a maximum of 280 characters; they may consist of words, URLs, or marked words such as hashtags (starting with #) and mentions (starting with @). The tweets were tokenized using TweetTokenize() method in the NLTK TweetTokenizer package.

An example is given below:

5. @brexit_politics: #Brexit: Nearly 20 banks have committed to Frankfurt since vote to leave EU, German officials say… https://t.co/cm3SV7

*Tokens* 'RT', '@brexit_politics', ':', 'Brexit', ':', 'Nearly', '20', 'banks', 'have', 'committed', 'to', 'Frankfurt', 'since', 'vote', 'to', 'leave', 'EU', ',', 'German', 'officials', 'say', '…', 'https://t.co/cm3SV7…'

6. Almost all violent extremists share one thing: their gender. Toxic masculinity is a thing. A very bad thing https://t.co/mJxXsH6LM8RT

*Tokens* 'Almost', 'all', 'violent', 'extremists', 'share', 'one', 'thing', ':', 'their', 'gender', '.', 'Toxic', 'masculinity', 'is', 'a', 'thing', '.', 'A', 'very', 'bad', 'thing', 'https://t.co/mJxXsH6LM8RT'

The tokenized tweets were further pre-processed by removing URLs and mentions (identified by the appropriate regular expressions). Very short words (less than 3 characters) and words with only non-alphanumeric characters were removed (punctuations and emoticons removed). Hashtags were processed by removing the '#' character and splitting the constituent words if they start with capital letters (e.g. #LondonTraffic becomes london and traffic). Such hashtags in which words were concatenated in CamelCase

notation were straightforward to segment. There are supervised methods to correctly segment hashtags that do not follow the CamelCase (Reuter,2016) However, in our dataset, there were only 126 such hashtags. We manually segmented these hashtags into constituent words. English stopwords were also removed from the tweets. Stopwords occur with relatively high frequency in the corpus but carry very low informative value. Examples of English stopwords from the nltk library are pronouns, propositions, and modal verbs. The tokens were lemmatized using WordNet NLTK lemmatizer. Lemmatizing is used to reduce each word to its lexical form and is preferred over stemming in topic modelling as the lemma it returns is ensured to be an existing word in the language.

**Filtering high-stress tweets** – Tweets with stress scores of -3 to -5 were collected for finding the dominant topics (Table 4.6).

**Table 4.6 Number of tweets with stress scores from -3 to -5 in the dataset consisting of 12000 tweets**

| Domain | Tweets |
|---|---|
| Airlines | 1835 |
| Politics | 1118 |
| Traffic | 1573 |
| Personal Events | 1013 |

**Topic modelling** – Each collection of high-stress tweets (belonging to a specific domain) was subjected to an LDA-based topic modelling implementation in Python gensim package. Prior to the topic modelling, hashtag pooling was performed on the tweets. The topics for high-stress tweets in each domain were extracted in this step.

We considered two commonly used methods for topic modelling of tweets: Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSI). Steven et al. (2012) described two different coherence measures- UCI (Newman et al., 2010) and UMass (Mimno et al., 2011) and how they can be used for evaluating the performance of topic modelling algorithms.

Topic coherence i.e., the average pairwise similarity between top words of a topic is a basic evaluation criterion. UCI and UMass are two commonly used metrics to measure topic coherence, both based on co-occurrences of word pairs.

**UCI metric**

The word pairs score is defined as the Pointwise Mutual Information between two words, as given in equation 4.1.

$$UCI(w_i, w_j, \varepsilon) = \log\left(\frac{P(w_i, w_j)+\varepsilon}{P(w_i)P(w_j)}\right) \text{----------(4.1)}$$

The word probabilities are measured as word co-occurrence frequencies collected over an external corpus like Wikipedia through a sliding window method. It is an extrinsic measure since it makes use of the external corpus.

**UMass metric**

This method measures the model's learning from the data in the corpus and thus is an intrinsic metric.

$$UMass(w_i, w_j, \varepsilon) = \log\left(\frac{D(w_i, w_j)+\varepsilon}{D(w_i)D(w_j)}\right) \text{----------(4.2)}$$

where $D(w_i, w_j)$ is the number of documents with both the words $w_i$ and $w_j$ and $D(w_i)$ is the number of documents with the word $w_i$

The key difference between these metrics is that UCI uses an external dataset and Umass considers the only documents from which the topics were extracted, to calculate the pairwise co-occurrence probabilities of topic terms. Intuitively, if two words in a topic co-occur a lot, it is an indicator of the coherence of the topic and hence the performance of the model. Both metrics are based on this co-occurrence probability, the difference being that UCI is an extrinsic measure whereas Umass is an intrinsic one.

We performed LDA and LSA, implemented using Python gensim modules, on the domain-specific datasets and compared the performance based on the two measures discussed above. The quality of the model was represented as the average coherence of all topics. Based on the coherence scores, we chose the LDA topic model for the analysis of reasons (Table 4.7).

**Table 4.7 UCI and UMass coherence measures of LDA and LSA in topic modelling the dataset of 5539 tweets with stress scores from -3 to -5, belonging to domains politics, traffic, airlines, and personal events**

| Coherence measure | Politics | | Traffic | | Airlines | | Personal events | |
|---|---|---|---|---|---|---|---|---|
| | LDA | LSA | LDA | LSA | LDA | LSA | LDA | LSA |
| UCI coherence score | 0.34 | 0.23 | 1.52 | 1.27 | 1.76 | 1.45 | 0.74 | 0.66 |
| UMass coherence score | -1.43 | -1.97 | -0.53 | -0.61 | 0.85 | 0.34 | 0.71 | 0.24 |

**Clustering** – Topic modelling provides a clustering of topics, however, these topics are unstable with semantically related terms appearing in different clusters and semantically dissimilar topics appearing in the same cluster (Figure 4.2).

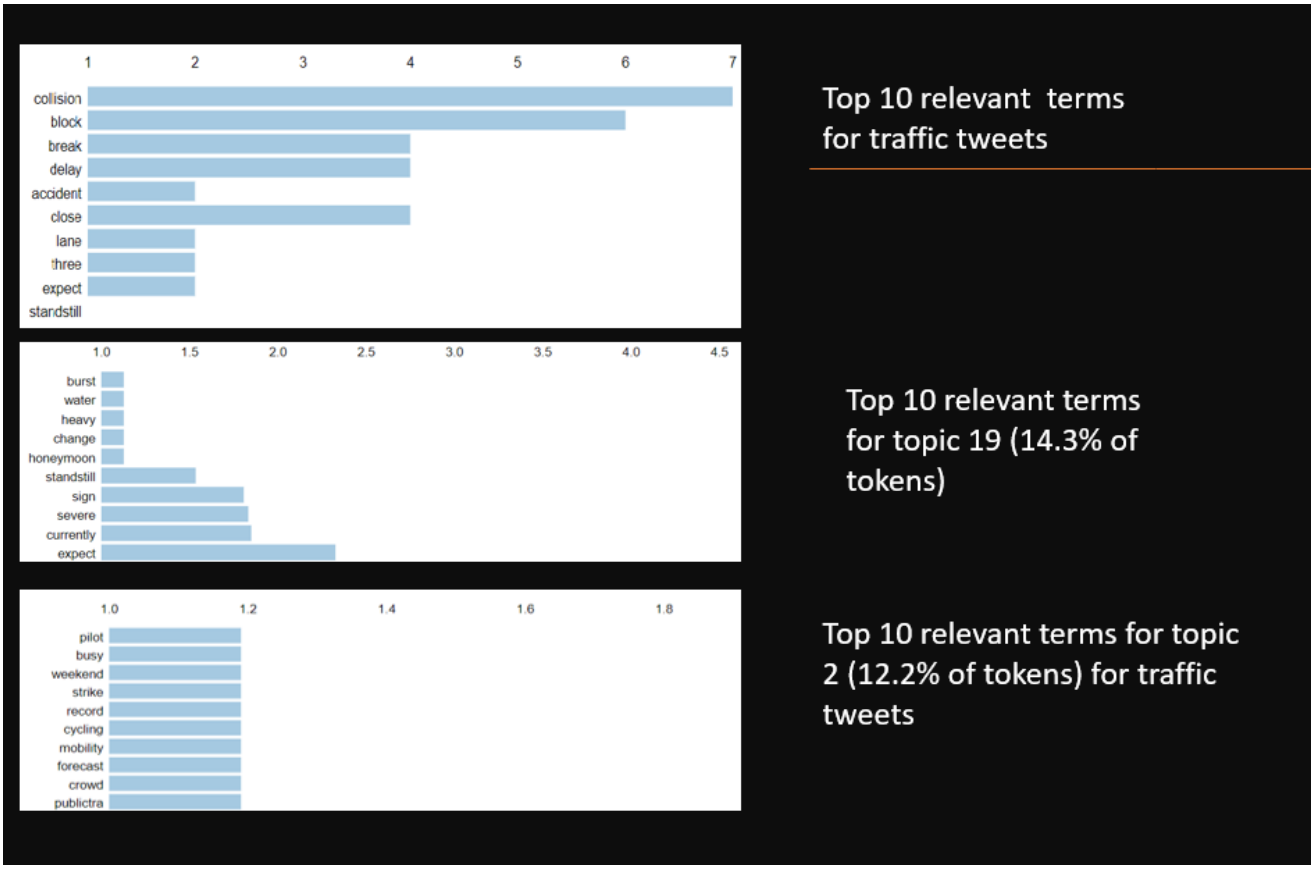


**Figure 4.2 Top 10 relevant terms in different topics extracted using LDA topic modelling on the dataset of 2173 high-stress tweets (stress scores -3 to-5) on London traffic**

This necessitates generating more coherent clusters of the topics generated by LDA, and we performed k-means clustering over the topics obtained from the previous step of topic modelling. Originally proposed in (Lloyd, 1982) K-means clustering clusters data elements based on their distance to centroids. In the first iteration of the algorithm, these centroids were fixed randomly. Each data element was assigned to the cluster of which centroid it is closest to. Once this assignment was completed, the centroid for each cluster was re-computed and the data elements were reassigned with respect to their distance to the new centroids. The algorithm converged when there were no changes in centroids in two consecutive iterations.

Since k-means clustering required numerical values, it was not directly applicable to cluster words. One way to get around this problem was to treat edit distance (Levenshtein distance) as a criterion for finding the nearest centroid for each word. This would group similarly spelt words (e.g. house and horse) into the same cluster. However, for most purposes, the semantics of words were more relevant. Specifically, in our context, we would want to group words with similar meaning into the same cluster (e.g. media, news) and dissimilar meaning (e.g. accident, march) into different clusters irrespective of their edit distances. Hence, we implemented a version of clustering using the k-means algorithm in the Python Scikit-learn package, with a Word2Vec vector representation for each word. Word2Vec is widely used for representing semantic analogies.

It generated feature vectors representing words based on the input corpus which made semantic clustering possible.

The optimal number of clusters needed to be chosen for each domain. We used silhouette analysis (Rousseeuw, 1987) to evaluate the clustering performance for a different number of clusters in each domain. Ideally, inter-cluster distance should be maximum and intra-cluster distance (between samples in the same cluster) should be minimum. We used the silhouette_score parameter in the scikit-learn metrics package to analyse this. This score is defined as the mean of silhouette coefficients of all samples. For each sample, the silhouette coefficient is

$$sl_s = \frac{(n_c - i_c)}{\max(n_c, i_c)} \text{ ----------(4.3)}$$

where, $n_c$ is the mean nearest-cluster distance

$i_c$ is the mean intra-cluster distance

The value of sl lies between 1(best) and -1(worst), and 0 indicates overlapping clusters.

We performed silhouette analysis for the clusters of word vectors representing the topics that emerged from each domain. The number of clusters that returns the highest mean silhouette score is chosen (Figure 4.3).

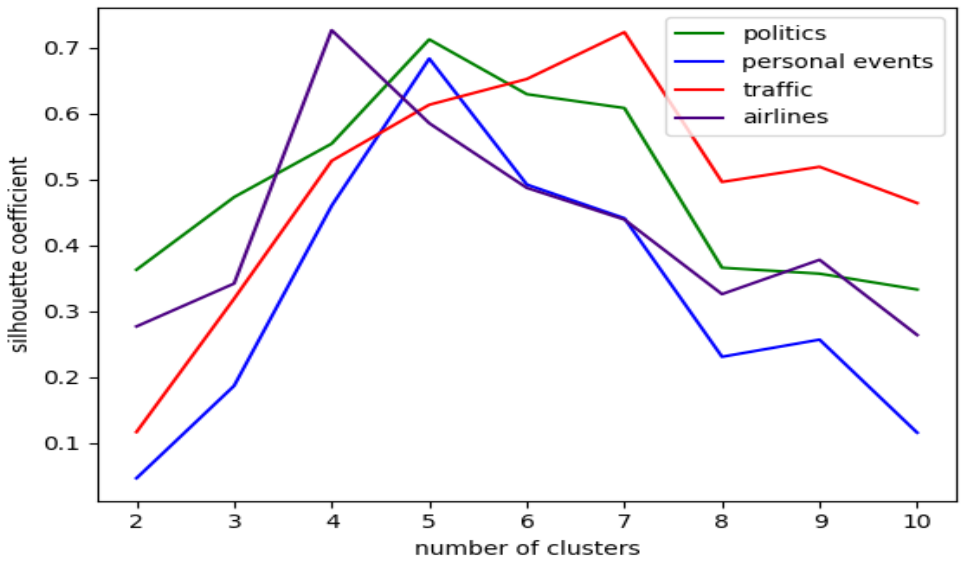


**Figure 4.3 Silhouette coefficients for different numbers of clusters in each domain**

We compared the coherence of clusters found by topic modelling alone and topic modelling combined with k-means clustering, using silhouette analysis (Table 4.8).

**Table 4.8 Mean silhouette scores of clusters derived from the two methods of clustering topic terms- topic modelling with and without k-means clustering (dataset of 5539 tweets with stress scores from -3 to -5 (domains politics, traffic, airlines, personal events)**

| Domain | Means silhouette score of clusters | |
|---|---|---|
| | Topic modelling | Topic modelling and k-means clustering |
| Politics | 0.127 | 0.701 |
| Airlines | 0.313 | 0.746 |
| Traffic | 0.112 | 0.719 |
| Personal events | 0.214 | 0.678 |

In all the domains, the coherence of the clusters has substantially increased after k-means clustering.

A title word that describes the content words in each cluster is assigned using the most_similar method in the Word2Vec Python library. The method most_similar returns the required number of words with the vectors closest to the input set of words. We set the topn parameter to 1 so as to get one word most similar to the input words. The most_similar method was run on the sample words for each of the topic clusters separately and the resultant word was assigned as the representative of the cluster and one of the stressors in the corresponding domain.

Each of these title words is considered as a stressor for the corresponding domain (Table 4.9, Table 4.10, Table 4.11, Table 4.12).

**Table 4.9 Clusters and sample topics for Politics derived from topic modelling followed by k-means clustering (dataset of 1418 tweets of stress scores from -3 to -5)**

| Cluster title (stressor) | Sample topics |
|---|---|
| Election | Majority general referendum select support leader power government voter people |
| Protest | move resign boycott challenge demonstrate chant march order |
| Violence | Hate burglary police abuse murder sniper kill bully army |
| Media | Report opinion mainstream press censor fake channel news |
| Economy | Programme plan austerity Brexit worker regulation free social reform |

**Table 4.10 Clusters and sample topics for Personal Events derived from topic modelling followed by k-means clustering (dataset of 1013 tweets of stress scores from -3 to -5)**

| Cluster title (stressor) | Sample topics |
|---|---|
| Death | Decease grandparent old accident ill cry perish suicide |
| Relationship | Love feel attraction date argument decide |
| Marriage | wed couple knot bride |
| Children | Baby parent birth pregnancy infant |
| School | Life university college graduation |

**Table 4.11 Clusters and sample topics for London traffic derived from topic modelling followed by k-means clustering (dataset of 2173 tweets of stress scores from -3 to -5)**

| Cluster title (stressor) | Sample topics |
|---|---|
| Climate | Air pollution heat heatwave burn |
| Violence | Assault attack gunman kill crime murder |
| Accident | Crash damage fatality hit vehicle emergency |
| Delay | Late hours minute expect rush |
| Congestion | Mayhem chocker block busy shambles trap crawl gridlock |
| People | Lady person punk elder woman shout yell |
| Campaign | Counsel march protest rally public crowd |

**Table 4.12 Clusters and sample topics for Airlines derived from topic modelling followed by k-means clustering (dataset of 2235 tweets of stress scores from -3 to -5)**

| Cluster title (stressor) | Sample topics |
|---|---|
| Luggage | Baggage carry issue steal miss report heavy |
| Service | Customer department reach staff pilot food unfriendly |
| Delay | Late hours issues weather time wait divert maintenance |
| Cost | Pay hundred dollar expense rise pay extra ticket rate money |

### 4.3.2 Finding Stressor from the Stressor List

The tweets were pre-processed to eliminate URLs, prepositions, interjections, conjunctions, and English language stop-words. The remaining words constitute the content words set. Three different word-vector based methods were used to find the causes of stress from the list of potential stressors.

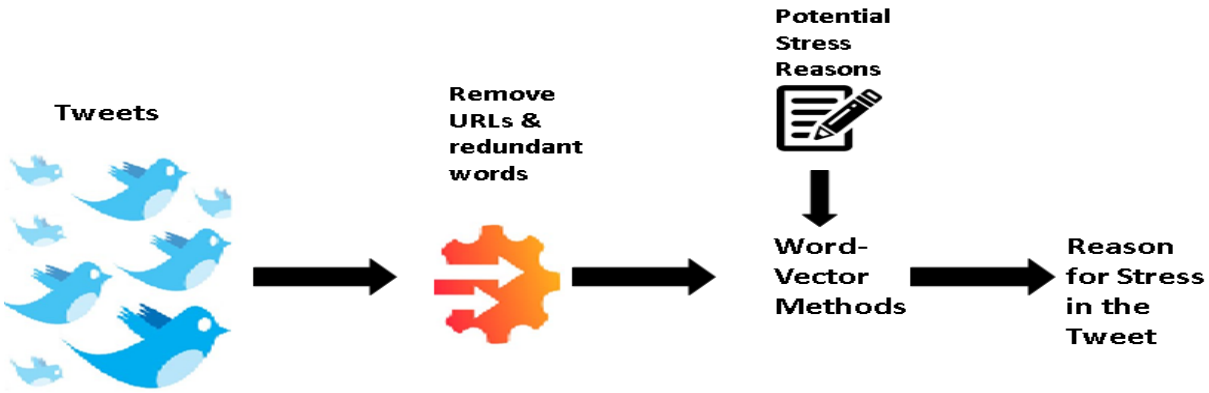


**Figure 4.4 The overview of the framework for finding stressors from tweets**

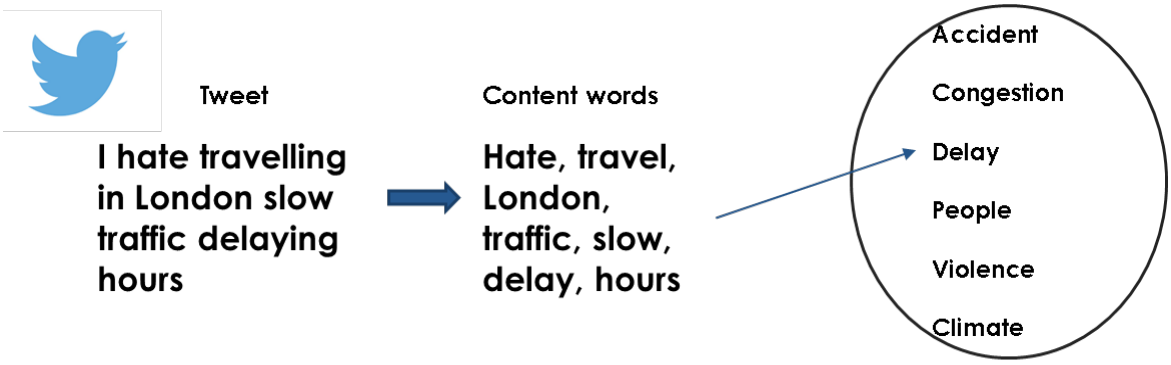


**Figure 4.5 Example of workflow for finding stressors given a tweet and potential stressors (traffic domain)**

Method 1 (maximum word similarity): The cosine similarity of each word in the content words set was calculated with each potential stressor, using corresponding word vector representations. The stressor with the highest similarity with any of the content words in the tweet was chosen as the stress cause.

Method 2 (context vector similarity): A context vector was found for each tweet by calculating the average of the word vectors of all words in the content words set. The stressor with the highest cosine similarity with this context vector was chosen as the stress cause.

Method 3 (cluster vector similarity): Each stressor was represented by a cluster vector which is the average of vectors of all words in its topic cluster. The cosine similarity of each of these cluster vectors was calculated with the context vector and the cluster with maximum similarity was chosen as the stress cause.

The three methods for finding the stressor in the following tweet are demonstrated (Figure 4.6).

7. "I hate travelling in London! Slow traffic and delaying hours"

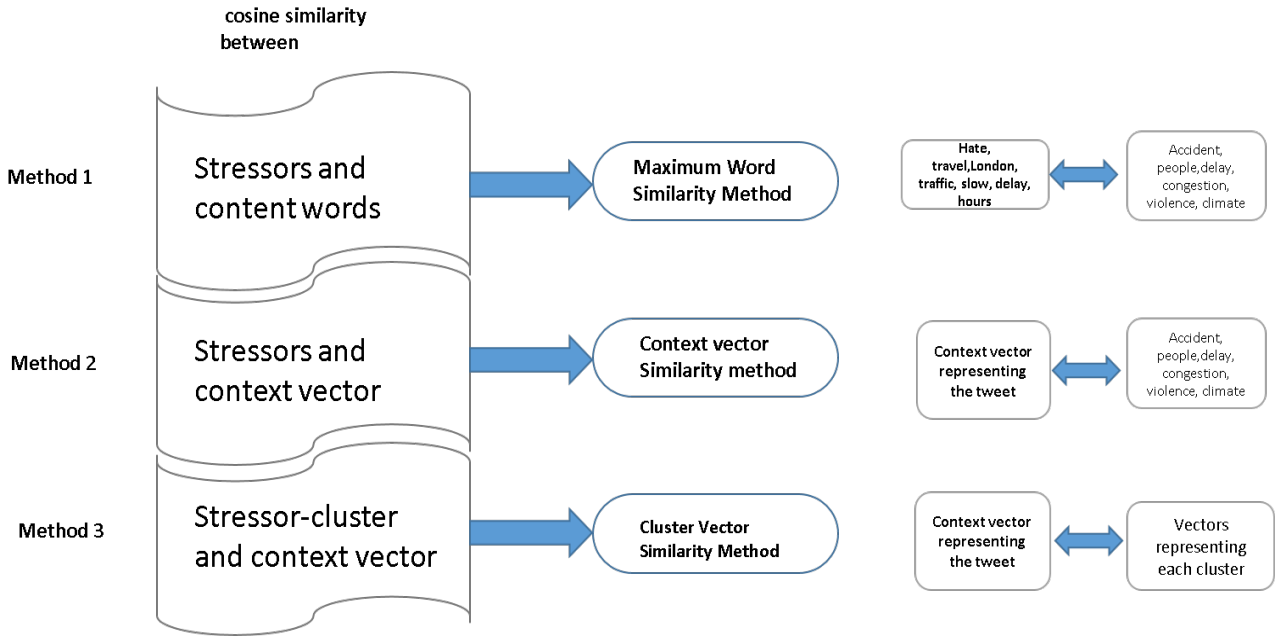


**Figure 4.6 Three word vector-based methods for finding stressor**

The words in the pre-processed tweet constituted the content words list. (hate, travel, London, traffic, slow, delay, hours). In the maximum word similarity method, the pair-wise cosine similarity method, the pair-wise similarity of content word and stressor word is calculated.

(delay, delay) has the highest similarity score (1.0), hence delay is chosen as the stressor for the tweet.

In the second method, a context vector is found by averaging the vectors corresponding to the content words. The cosine similarity between this context vector and the vector representing each stressor word is found (Table 4.13).

**Table 4.13 Cosine similarity of stressors of traffic domain with the context vector representing the tweet "I hate travelling in London! Slow traffic and delaying hours"**

| **Stressor** | **Cosine similarity with context vector** |
|---|---|
| Accident | 0.287 |
| People | 0.319 |
| **Delay** | **0.734** |
| Congestion | 0.527 |
| Violence | 0.347 |
| Climate | 0.211 |

The context vector has the highest cosine similarity to the word vector representing ‘delay’, hence it is chosen as the stressor in the tweet.

In the cluster vector similarity method, the cosine similarity between the context vector found in method 2 and each cluster vector is calculated.

**Table 4.14 Cosine similarity of cluster vectors of the stressors of traffic domain with the context vector representing the tweet “I hate travelling in London! Slow traffic and delaying hours"**

| **Stressor** | **Cosine similarity** |
|---|---|
| Accident | 0.310 |
| People | 0.271 |
| **Delay** | **0.792** |
| Congestion | 0.498 |
| Violence | 0.401 |
| Climate | 0.183 |

The context vector has the highest similarity to the vector representing the cluster belonging to the stressor ‘delay’, which is chosen as the stressor for the given tweet.

### 4.3.3 Experimental Setup

Topic modelling and clustering in the framework were performed using the scikit-learn package in Python. To train the word vectors, we used a Word2Vec model trained on 400 million tweets, originally proposed in an ACL WNUT task (Godin et al., 2015). As comparison baselines for the performance of our methods, we use three machine learning algorithms employing different approaches (Logistic Regression (LogReg), Support Vector Machines (SVM), and Adaptive Boosting Algorithms (AdaBoost)) which have been shown to be effective in sentiment analysis and text processing tasks.

**Logistic Regression** – Regularized Logistic Regression with L2 penalty, cross-entropy loss, and multi-class option set to ‘multinomial’ and lbfgs as solver. Since this is a multi-class classification problem, we set the multi-class option to ‘multinomial’. The sklearn package in Python uses cross-entropy loss for this option. For multiclass problems, only ‘newton-cg’, ‘sag’, ‘saga’ and ‘lbfgs’ handle multinomial loss; ‘liblinear’ is limited to one-versus-rest schemes. We chose lbfgs as solver with L2 regularization.

**Support Vector Machines** – Sklearn implementation of SVM provides two options for the multi_class parameter: ‘ovr’ or ‘crammer_singer’. According to the sklearn documentation ‘ovr’ trains one-vs-the-rest

classifiers. 'Crammer_singer', while optimizing overall classes, is computationally expensive, and doesn't result in significant improvements in accuracy. Hence, we set the multi_class parameter as 'ovr'. For penalty and c parameters, we retain the default values as l2 and 1.0 respectively.

**Adaptive Boosting** - DecisionTreeClassifier as base estimator, number of estimators as 100. We used the multi_class adaboosted decision trees implementation based on previous work (Zhu, et al., 2006).

Term unigrams, bigrams, and trigrams and their frequencies were used as features. Punctuation was included as a term, with consecutive punctuation treated as a single term.

The performance was assessed with precision, recall, and F-measure.

- Accuracy is the percentage of the instances where the stressor is correctly identified out of all the tweets in the dataset.
- $\text{Precision}_{\text{economy}}$ is the percentage of the instances where the algorithm correctly predicted stressor as economy out of the total instances where it is (correctly or incorrectly) predicted as economy.
- $\text{Recall}_{\text{economy}}$ is the percentage of the instances where the algorithm correctly predicted stressor as economy out of the total instances in the dataset where the stressor is actually economy.
- Mean precision is the average of the precision values over all the stressors in the given domain.
- Mean recall is the average of the recall values over all the stressors in the given domain.
- F-measure is the harmonic mean of precision and recall.
- F-measure = $\frac{2*precision*recall}{precision+recall}$

## 4.4 Results

The stressors for the dataset of 8871 tweets belonging to the four domains of politics, personal events, traffic, and airlines were found using the word vector processing methods. The cluster vector method gives the best performance in terms of all measures considered (precision, recall, accuracy, and F-measure). (Table 4.15, Figure 4.7, Figure 4.8, Figure 4.9).

**Table 4.15 Performance of the different methods in identifying the stressors for the dataset of 8871 tweets belonging to the domains of politics, personal events, traffic and airlines (accuracy)**

| Method | Accuracy | | | |
|---|---|---|---|---|
| | **Politics** | **Personal Events** | **Traffic** | **Airlines** |
| Max word | 47.81 | 49.23 | 48.3 | 50.23 |
| Context vector | 54.63 | 51.22 | 53.9 | 59.74 |
| Cluster vector | **63.41** | **66.72** | **63.8** | **67.29** |
| SVM | 58.42 | 62.31 | 62.41 | 65.15 |
| AdaBoost | 50.64 | 54.85 | 52.5 | 54.85 |
| Logistic Reg | 49.23 | 51.94 | 49.1 | 52.5 |

Since finding the stressors was a multi-class classification problem, precision, recall, and f-measure were calculated as the core measures of overall performance (Figure 4.7, Figure 4.8, Figure 4.9).

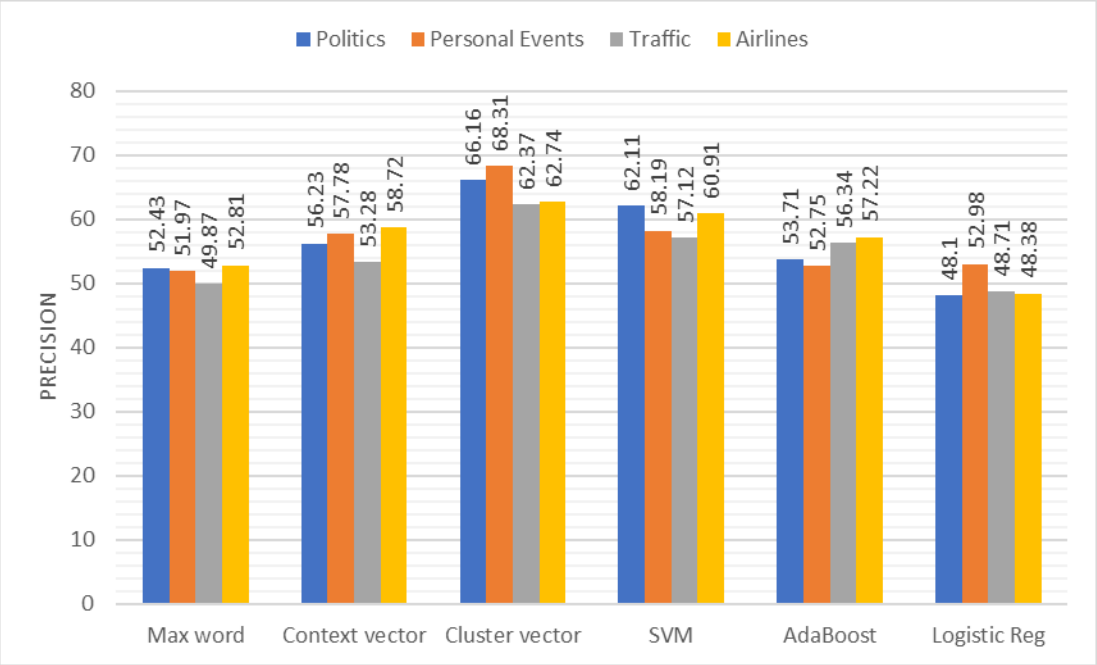


**Figure 4.7 Performance of the different methods in identifying the stressors for the dataset of 8871 tweets belonging to the domains of politics, personal events, traffic and airlines (precision)**

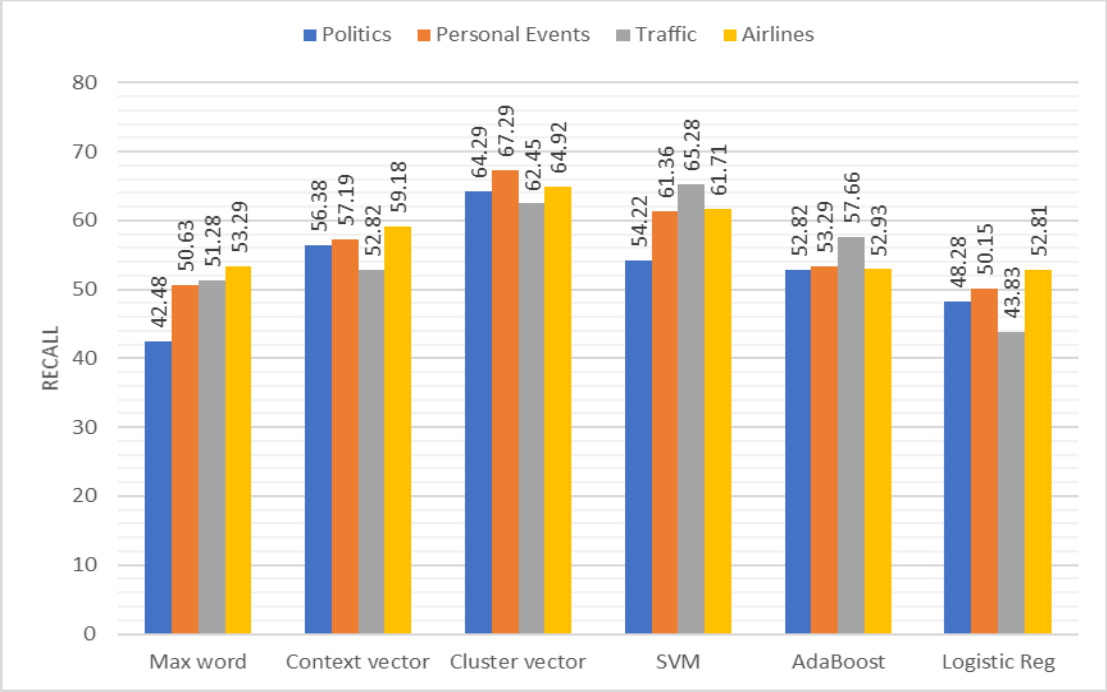


**Figure 4.8 Performance of the different methods in identifying the stressors for the dataset of 8871 tweets belonging to the domains of politics, personal events, traffic and airlines (recall).**

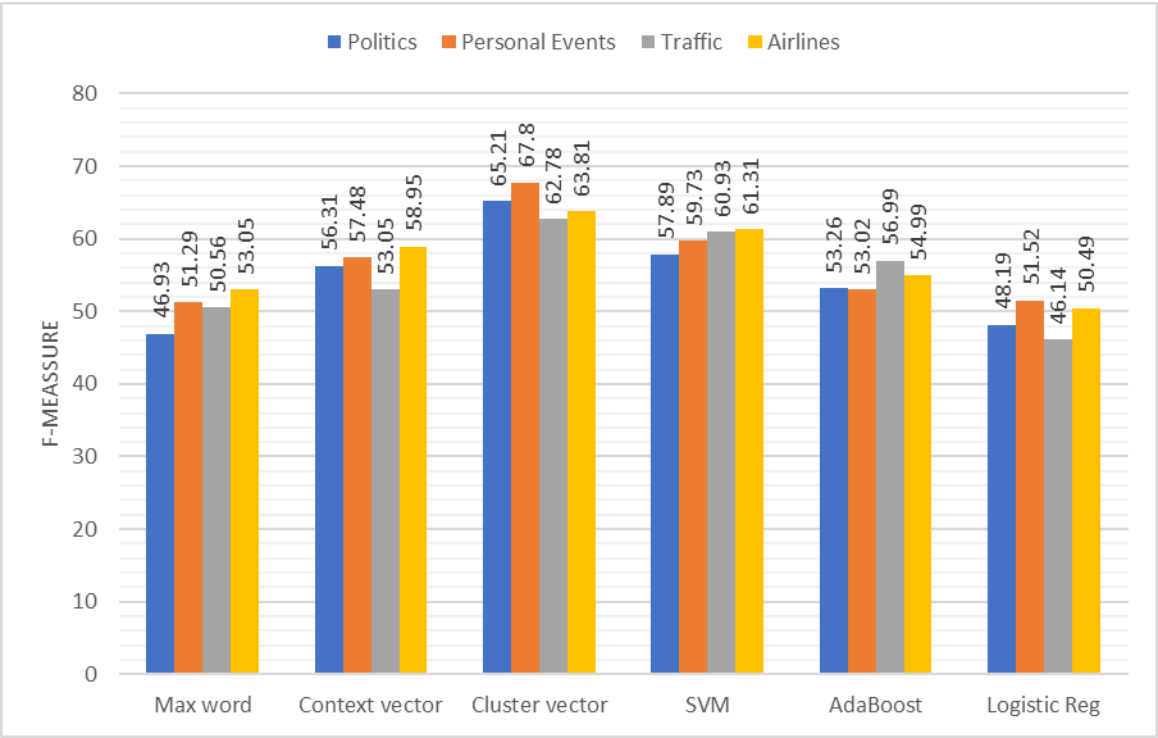


**Figure 4.9 Performance of the different methods in identifying the stressors for the dataset of 8871 tweets belonging to the domains of politics, personal events, traffic and airlines (F-measure)**

The cluster vector method outperforms the other methods in finding the stressor in tweets for all the domains. The performance of context vector and SVM classifiers was the second-best and comparable to each other.

We assumed that the reason for the good performance of the cluster vector method is that it considers the semantic similarity between the tweet and all constituent words in a cluster of stress topics. It considers the broad context of both the tweet and the stressor which might be contributing to the better predictions.

In terms of accuracy, the maximum word similarity method performed worst in all domains, because it considers only the pairwise similarity between the stressors and the words in the tweet. It performs well for tweets where one of the stressors is directly mentioned in the tweet (e.g., "I hate travelling in London! Slow traffic and delaying hours"). However, the percentage of such tweets is relatively low and the other methods perform better for the tweets in which the stressor is indirectly mentioned as demonstrated by the results (Table 4.16; Figure 4.7, Figure 4.8, Figure 4.9).

## 4.5 Discussion

This section analyses the errors in identifying the correct stressors in each of the domains. The stressors identified using the cluster vector method are compared to the stressors assigned by human annotators in the confusion matrices (Table 4.16, Table 4.17, Table 4.18, Table 4.19).

Analysing the confusion matrices for the domains provided further insights:

**Traffic**

**Table 4.16 Confusion matrix with different stressors for the domain Traffic (2613 tweets)**

| **Predicted Reason** | **Human annotated reason** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Accident** | **Campaign** | **Climate** | **Congestion** | **Delay** | **People** | **Violence** |
| Accident | 726 | 7 | 4 | 41 | 20 | 6 | 4 |
| Campaign | 15 | 66 | 22 | 28 | 5 | 11 | 2 |
| Climate | 35 | 17 | 248 | 54 | 3 | 26 | 3 |
| Congestion | 25 | 52 | 6 | 209 | 119 | 18 | 11 |
| Delay | 118 | 36 | 56 | 108 | 198 | 25 | 2 |
| People | 57 | 21 | 35 | 18 | 35 | 74 | 6 |
| Violence | 129 | 36 | 27 | 24 | 31 | 11 | 21 |

The highest number of misclassified stressors occur between the classes congestion and delay. Out of the 814 tweets where congestion was the actual stressor as per human annotation 318 were correctly classified but 218 were classified as delay. Similarly, out of the 646 tweets with delay as stressor, 318 were correctly classified and 189 were incorrectly labelled with congestion.

Example:

(16)Rushed late to a client in traffic jam in Central London and forgot where I parked my car! I need a holiday please!

Human annotated reason: Delay

Reason from the cluster vector method: Congestion

The high number of misclassifications between these two stressors point to the possible co-occurrences of words indicating delay and congestion in the same tweet. Since congestion often results in being late or slow-moving traffic, it is hard to differentiate the exact stressor even for human annotators in those tweets.

Overall, the stress class 'climate' had the highest recall (0.683) and congestion (0.416), people (0.454), and delay (0.473) had low recall. Tweets in the stressor category had explicit climate-related terms such as 'heat' and 'heatwave' which helped the word vector methods relying on semantic similarity to correctly identify those.

**Personal Events**

**Table 4.17 Confusion matrix with different stressors for the domain Personal events (1838 tweets)**

| **Predicted reason** | **Human annotated reason** | | | | |
|---|---|---|---|---|---|
| | **Children** | **Death** | **Marriage** | **Relationship** | **School** |
| Children | 552 | 11 | 28 | 25 | 55 |
| Death | 5 | 84 | 19 | 54 | 32 |
| Marriage | 53 | 14 | 152 | 106 | 12 |
| Relationship | 31 | 4 | 121 | 236 | 45 |
| School | 38 | 5 | 16 | 14 | 126 |

In tweets discussing personal events, the stressors 'marriage' and 'relationship' were misclassified as each other. These classes had relatively low recall measures (marriage: 0.474 and relationships: 0.541) The other three classes- children, death, and school - had relatively higher recall values (0.724, 0.737, and 0.621) respectively. Tweets in this domain mostly discuss any one of the stressors and hence there is less overlap between different stressor classes. Hence the vector similarity of the tweet is substantially higher with a specific cluster, which is correctly chosen as the stressor for the tweet.

**Politics**

**Table 4.18 Confusion matrix with different stressors for the domain Politics (1927 tweets)**

| Predicted reason | Human annotated reason | | | | |
|---|---|---|---|---|---|
| | **Economy** | **Election** | **Media** | **Protest** | **Violence** |
| Economy | 263 | 7 | 36 | 45 | 42 |
| Election | 55 | 372 | 43 | 8 | 43 |
| Media | 23 | 33 | 224 | 8 | 42 |
| Protest | 18 | 24 | 32 | 112 | 82 |
| Violence | 28 | 8 | 7 | 88 | 264 |

Politics had the following recall values for the stressor classes - economy (0.653), election (0.665), media (0.551), protest (0.527) and violence (0.598). Protest and violence had a high number of mutual misclassifications (261 tweets with stressor protest-88 misclassified as violence; 473 tweets with stressor violence-82 misclassified as protest). Similar to the other domains, this was due to the occurrence of words which are semantically similar to violence occurring in the tweet with protest as the classifier and vice versa.

**Airlines**

**Table 4.19 Confusion matrix with different stressors for the domain Airlines (2493 tweets)**

| Predicted stressor | Human annotated stressor | | | |
|---|---|---|---|---|
| | **Cost** | **Delay** | **Luggage** | **Service** |
| Cost | 157 | 93 | 44 | 43 |
| Delay | 14 | 613 | 163 | 68 |
| Luggage | 45 | 111 | 283 | 54 |
| Service | 74 | 113 | 36 | 307 |

The recall values of the stressors of the airlines domain - cost, delay, luggage, and service - were 0.54, 0.66, 0.53, and 0.65 respectively. There is a high number of misclassifications (n=163) where the annotated stressor was luggage but the predicted stressor was delay. An example is

(17)#turkishairlines so pathetic the baggage has still not arrived. Why this delay?????

In this example, the stressor is luggage, but the word vector methods chose delay instead.

### 4.5.1 Error Analysis

There are some systematic reasons for the methods failing to find the correct stressor.

**Indirect Expressions:** Tweets with indirect expressions of stress pose a challenge to our methods. Examples

(18) The joys of London traffic; not moving for last one hour is killing me

(19) Lea Bridge rd E10 traffic is murder really London counsel

In these examples, the stressor is congestion but is detected as violence.

(20) #ryanair You lot should be ashamed of yourself. Pathetic, rude staff. Cost me my peace for whole 2 hours of flying!

The human-annotated stressor is "service" but cost is a misleading word which makes the maximum word similarity method to choose the stressor as "cost". Context vector and cluster vector methods, which consider the aggregated tweet vector instead of the vectors of individual words, correctly identified the stressor as "service".

**Multiple stressors:** Tweets in which there are multiple stressors. In those examples, the human annotation is more subjective and harder to reach a consensus as there is more than one stressors which look equally probable. Also, the vector similarity of the tweet to two or more clusters will be high, however the algorithm will simply choose the one with the highest similarity.

(21)#AmericanAirlines lost baggage!!! Will never fly again! 2 delays in 1 night too.

The human annotation identified the stressor as luggage though delay is also mentioned based on the relative placement of the two stressors (luggage and delays). However, the cluster vector method identified the stressor as delay. Again, this is an example of multiple stressors in the same tweet. It is to be noted that the vector similarity between the tweet and the cluster of luggage (0.772) was very close to the vector similarity between the tweet and the cluster of delay (0.788). In such marginally different cases, predicting multiple stressors could be a more reasonable option.

(22) Vehicle emissions rise during rush hours, making the traffic jams hellish

Similar to (23) this tweet has two stressors, "pollution" and "congestion".

A possible solution will be to expand the methods to accommodate multiple stressors.

### 4.5.2 Limitations

As shown in the error analysis, the methods are challenged by the tweets with indirect expressions. For tweets which have misleading words, the word vector methods often report stressors corresponding to the literal rather than implied meaning. Another factor is the presence of multiple stressors in the same tweet which is overlooked by the present method. A modified method which predicts the confidence interval for multiple stressors for a single tweet would possibly be more realistic for handling such tweets. However, given the

brevity of the tweets, it would depend on the dataset whether such an adaptation would indeed perform better.

Furthermore, there could be tweets with no specific stressor. The annotators simply labelled all the tweets with the best matching stressor. In datasets from the other domains or those collected randomly, there could be a high presence of tweets without a specific stressor. This would need to be addressed by adding a 'no-stressor' class to the annotation labels. If the vector-based similarities of the tweet to all the clusters are below a specific threshold, the tweet could be assigned to the no-stressor class. Also, the annotators can potentially find a stressor not present in the list of stressors given in the annotation instructions. The present approach does not address that possibility.

The methods for extracting and assigning the stressors have been evaluated on tweets collected with generic, broad domain topics. Though they can be easily adapted to other domains, their performance on more granular, focused domains has to be validated yet.

Another potential limitation is the approach of treating tweets as bag of words. The order and the relative distance between words could have been useful to better identify the correct stressor, especially in tweets with multiple stressors present.

Lastly, the tweets were gathered based on domain-specific hashtags to ensure a wider selection. There are very little explicit self-declarations of stress or the underlying reason. As evidenced from basic pattern matching, there are no tweets which contain phrases like "I feel stressed because…" or "I am stressed due to…". With this lack of explicit declarations, the reasons/stressors are implicitly mentioned. Hence, the gold standard for stressors is constructed based on the annotators' perception of the tweet rather than explicitly mentioned causality within the tweets. However, the high agreement metrics between the multiple coders validate the usage of the annotated labels as the evaluation criteria.

## 4.6 Conclusions

The chapter addressed the problem of developing a generic framework for finding stressors from tweets belonging to different domains. Intuitively, for each domain the stress-inducing factors would be different. Hence, a two-step method was adopted to construct a list of potential stressors for each domain and to map a tweet to one of the stressors. A list of potential stressors for each domain was constructed from high-stress tweets using LDA topic modelling and k-means clustering. We proved, using silhouette analysis that combining topic modelling with k-means clustering improved the coherence compared to topic modelling alone. This approach could be used in improving the coherence of word clusters in future research.

For each tweet, the stressor was found from this list using three different word-vector based similarity methods. The word vector similarity method compared the pairwise cosine similarity of the vectors representing words in the tweet and the stressor words. The pairwise similarity between the tweet vector and the stressor words was found in the context vector similarity method. In the third, cluster-vector similarity

method, the stressor is represented by the average of the words in the cluster representing it; the stressor cluster with the highest similarity with the tweet vector is chosen as the stressor.

Among these, the cluster-vector method performed the best in terms of various metrics (accuracy, precision, recall and F-measure), in the datasets of the four domains- traffic, airlines, personal events and politics. Context vector method and SVM classifier give the next best performance. For the domains of politics, personal events, traffic and airlines, the F-measures for the cluster vector method in identifying the stressor are 65.21, 67.80, 62.78 and 63.81 respectively whereas the F-measures for the best-performing machine learning classifier (SVM) using n-grams is 57.89, 59.73, 60.93 and 61.31 respectively. Because of the consistently better performance of the cluster-vector method in all the four domains, it could be potentially adapted to extract the stressors from other domains.

It is possible that with more advanced features and a larger dataset, the machine learning classifiers could give better performance. However, given the simplicity of implementation and adaptability across multiple domains, cluster vector similarity method is a promising possibility for identifying stressors, especially from smaller datasets.

This can also be useful as a pre-annotation step. This is specifically significant in the task of stressor identification where human-annotated, domain specific datasets are rare to non-existent. The cluster vector similarity method might be helpful to address this issue, by providing pre-annotations for the possible stressor in the given tweet. Future experiments can explore the potential time saving and bias effect by this pre-annotation step.

Tweets with indirect expressions or multiple stressors were challenging to classify by the word vector methods. We proposed modifying the methods to predict the confidence interval of stressors rather than to identify a single stressor for different tweets. This might be helpful in accommodating the cases where there is more than one stressor for a single tweet. Also, the word vectors used in the methods were trained using the same Word2Vec model. Domain-specific Word2Vec models trained on separate topic-based collections tweets might improve the word vectors; this could be explored in future research.

# Chapter 5

## SWEARING AND SARCASM IN HIGH-STRESS TWEETS

This chapter detects potential manifestations of stress in Tweets through swearing and sarcasm as possible outlets of stress in social media language. In the previous chapter (chapter 4), a method to identify the potential stressors from tweets was described. In particular, stressors belonging to different domains of airlines, traffic, politics and personal events were studied separately. Using the same dataset as was used for analysis of stressors, we continue to analyse specific linguistic features in chapters 5 and 6. Swearing and sarcasm have been studied in traditional psychology as an expression or alleviator of stress. Inspired from that research background, this chapter further investigates their presence in the tweets dataset – highlighting its difference in high-stress and low-stress tweets. For each of the domains, we analyse these linguistic features.

Due to the uncensored nature of social media content, users may use swear words. Swearing is the use of socially offensive language. The Cambridge Dictionary[11] defines 'offensive' as 'causing someone to be upset or have hurt feelings'. This is a subjective definition and, in many contexts, depends on the perception of the reader/listener. Swear words are often taboo in more formal contexts. The perception of taboo words changes with the developmental course as well with children, adolescents, and adults differing in their perceived offensiveness of the same word (Jay & Janschewitz, 2008). This subjectivity makes offensive language hard to define and, thus, hard to detect. It could be in one of the three forms: 1) untargeted swear words 2) swear words targeted towards an individual 3) hate speech with or without swear words targeted at a group (based on gender, sexual orientation, religion or race). Swearing can also be used in non-offensive contexts. In some contexts, swearing can be used within informal conversation as a method of signaling friendship (e.g., "Modern Family was bloody amazing, mate"). In this common use, it signals relaxation more than stress but detecting such contexts is hard within tweets due to a lack of knowledge of the relationship between the tweeter and intended audience. Thus, this chapter, as a first approach, ignores non-stress uses of swearing. This is a limitation that future research is needed to address. This simplification is broadly appropriate for the domains analysed (airlines, personal events, traffic, politics).

Swearing, the usage of socially offensive language, can be detrimental to a healthy social networking experience, but psychological research treats swear words as an inherent component in language. It is viewed as a result of the speaker's emotional state and an effort to convey that emotional state to the listener (Jay & Janschewitz, 2008). Swearing can also be a coping mechanism or response to stress. A survey of the stress and

[11] https://dictionary.cambridge.org/dictionary/english/offensive

coping strategies of students at Gulf Medical University, United Arab Emirates, found that 32.5% reported that they use verbal expressions including swearing to vent stress (Gomathi & Ahmed, 2012). Swearing fluency, the number of swear words generated by a participant in a minute, was found to be greater with raised emotional arousal (Stephens & Umland, 2011). Moreover, swearing is useful in reducing stress and aggressive drive (Vingerhoets et al., 2013). This chapter draws inspiration from these research works and analyses the presence of socially offensive/swearing posts in the dataset of high-stress tweets.

Sarcasm is a form of speech in which the intent of a message is implicit and is contradictory to the literal meaning. Understanding sarcastic language is a non-trivial task for even human readers. It is often used to express negative sentiment about a situation, as an outlet of frustration. Like offensive language, sarcasm can also be used as a coping mechanism for stress. Sarcasm decreases the hormone cortisol, an indicator of psychological stress, in the speaker and has been shown to be used as a behavioural response to cope with stress (Roubinov, Hagan & Luecken, 2012). This study specifically considered only verbal usage of sarcasm whereas the relation between written expressions of sarcasm and psychological stress is yet to be explored and established. Interestingly, correctly identifying sarcasm indicators has been shown to improve the performance of sentiment detection, which is a task comparable to stress detection, in the context of tweets (Maynard & Greenwood, 2014).

The rest of the chapter is organized as follows: Section 5.1 reviews the existing work in the tasks of identifying swearing and sarcasm in social media. Section 5.2 describes the methods for analysing offensive language and sarcasm in high-stress tweets. The results are presented in Section 5.3, with a discussion of the errors and limitations in Section 5.4. The chapter concludes with a brief summary and discussion of future work in Section 5.5.

## 5.1 Related Work

### 5.1.1 Detection of Socially Offensive/Swearing Posts

As discussed in the introduction, offensive language can take different forms. It can be the usage of profanity, bullying targeted at an individual, hate speech aimed at a specific individual or group based on their gender, religious, national or racial identity. It has been found to peak after violent events, reflecting high emotional arousal. A study of gender differences of swearing in MySpace users, using a lexicon-based approach, revealed more usage of very strong swear words by male users compared to female users in the UK but similar usage of strong swear words by both groups (Thelwall, 2008). The ubiquity and usages of cursing in Twitter, together with the gender and social roles of the participants of the conversation, have been studied using a lexicon approach (Wang et al., 2014). The masking numbers and special characters (e.g. S#it) are handled using pre-processing. This approach results in high precision and low recall in a large dataset of 51 million tweets. It

relies upon annotating the list of swear words for their probability of being used in offensive and non-offensive sense and considering only those swear words which were labelled as 'mostly used for offensive' usage. This approach doesn't consider whether a swear word has been indeed used in offensive meaning in a specific tweet. Hence, we further examined the other approaches to identify swearing in tweets.

Analyses of offensive posts in social media largely focus on hate speech targeted at a specific group. Examples include the London Riots of 2011 (Lewis et al., 2011) and the Woolwich incident of 2013 (Williams & Burnap, 2015). Various features and machine learning and deep learning methods have been used to identify the presence of offensive/swearing posts in social media and online discussion forums (Table 5.1).

Lexical features (such as word n-grams, character n-grams, word TF-IDF) (Williams & Burnap, 2015; Nobata et al., 2016; Malmasi & Zampieri, 2017; Badjatiya, 2017; Warner & Hirschberg, 2012; Mehdad & Tetreault, 2016;) and sentiment scores (Schmidt & Wiegand, 2017) are typically used in identifying social media text as offensive or not. User characteristics such as gender, Author Historical Salient Terms and behavioural tendency to engage in hatred expressions have been explored to improve the classification accuracy of hate speech (Pitsilis, Ramampiaro, & Langseth, 2018; Waseem & Hovy, 2016). Models using character n-grams as a feature were found to outperform those which do not. In one instance (Malmasi & Zampieri, 2017), the classification model using only character 4-grams outperformed the model using it in combination with several other character and word n-gram features. Character-based systems are better in identifying obfuscated profanity and racial slurs, potentially with a combination of letters and digits. Models using behavioural tendencies of users were also observed to perform better but this requires the historical information of the posting behaviours of users, which may not be available in many datasets.

Supervised learning classification methods, specifically Support Vector Machines (Nobata et al., 2016; Malmasi & Zampieri, 2017) and Logistic Regression (Djuric et al., 2015) have been employed in offensive language detection. An ensemble classifier using SVM and Random Forest Decision Trees analysed hate speech in tweets related to the Woolwich incident, using typed dependencies and unigram, bigram frequencies (Williams & Burnap, 2015). Deep learning methods like CNN (Park & Fung, 2017) and LSTM (Badjatiya, 2017) were successfully used too in abusive language detection. A direct comparison of these deep learning methods is presented in (Badjatiya, 2017), in which LSTMs outperformed CNN based models. However, the authors attribute this better performance to the ability of RNNs like LSTM to capture long-range dependencies. It is useful in identifying hate speech, but may not be relevant for other types of socially offensive language. Also, the learning curve of the model is not presented, hence it is not clear if the better performance will hold in a smaller dataset. In a broad study, considering different types of offensive language, not limited to hatespeech, in a dataset of 14100 tweets, CNN is found to outperform BiLSTMs (Zampieri et al., 2019).

In recent years, research in offensive language has further explored gender differences in the usage of profanity (Wong, Teh & Cheng, 2020), hatespeech in the context of Covid-19 (Fan, Yu, & Yin, 2019) anti-Semitic content (Ozalp & Williams, 2019). Together with neural network models, supervised machine learning classifiers are still used for their simplicity and promising results in existing research. The effect of user-specific information such as gender, profile activity and networking in offensive language identification was examined but it was found not to have a significant impact (Unsvåg & Gambäck, 2018).

**Table 5.1 Examples of research on offensive language and hate speech detection from social media posts**

| Authors | Focus | Method | Features |
|---|---|---|---|
| Fan, Yu & Yin, 2019 | Hate speech in covid-19 related tweets | Binary decision trees | Emotions |
| Ozalp & Williams, 2019 | Anti-semitic hatespeech | SVM | BoW features |
| Wong, Teh & Cheng, 2020 | Gender differences in using profanity | Lexicon models | |
| Zampieri et al., 2019 | Type and target of offensive language – presented OLID a dataset annotated at three levels – offensive/not, targeted offense/untargeted, targeted at an individual/ a group/other | SVM, BiLSTM, CNN (best performing) | Word unigrams and word embeddings |
| Kshirsagar, 2018 | Hatespeech (sexist/racist) | MLP | Word embeddings |
| Unsvåg & Gambäck, 2018 | Hatespeech | Logistic Regression | User features (sender, network, activity and profile information) with word and character n-grams |
| Holgate et al., 2018 | Purpose of swearing words | Logistic regression | GloVe vectors, sentiment, PoS, Brown clusters |
| Thelwall, 2008 | Gender differences in swearing by UK and US users | Lexicon | |
| Wang et al., 2014 | Differences of gender and social roles in swearing | Lexicon | |
| Williams & Burnap, 2015 | Hate speech in relation to events (2013 murder of drummer Lee Rigby, Woolwich, London) | Ensemble classifier (Bayesian logistic regression) SVM and Random Forest decision trees | Typed dependencies (Stanford Lexical Parser) frequency of unigrams and bigrams |
| Djuric et al., 2015 | Abusive language | Logistic regression | Text embedding using Paragraph2Vec |

| | | | |
|---|---|---|---|
| Nobata et al., 2016 | Hate speech, derogatory, profanity in Yahoo! News, and Finance | SVM | n-grams, linguistic, syntactic, embedding-based features |
| Malmasi & Zampieri, 2017 | Hate speech and offensive language in Twitter | SVM | n-grams, word skip-grams, Brown clusters |
| Pitsilis, Ramampiaro & Langseth, 2018 | Hate speech in Twitter (racism and sexism) | An ensemble of LSTM classifiers | users' tendency towards hatred |
| Park & Fung, 2017 | Hate speech (racism and sexism) | Hybrid CNN and logistic regression | Character and word features |
| Badjatiya, 2017 | Hate speech (racism and sexism) | FastText, CNNs, and LSTMs | Character n-grams, word TF-IDF, BoW vectors over GloVe |
| Warner & Hirschberg, 2012 | Hate speech in Yahoo! Group comments | SVM classifier | Unigrams |
| Waseem & Hovy, 2016 | Hate speech (racism and sexism) | Logistic Regression | Ngrams and user features (gender and location) |
| Schmidt & Wiegand, 2017 | Hate speech (race/gender/ sexual orientation) | Survey on various methods | Lexicons, sentiment, word representations, linguistic features |
| Mehdad & Tetreault, 2016 | Offensive (hate speech and profanity) | RNN, Language models, SVM with Naïve Bayes features | n-grams and character n-grams |

There have been several academic events and shared tasks in recent years focusing on the identification of abusive or aggressive language on social media. Events like the first and second Workshop on Abusive Language (AWL)[12], and the First Workshop on Trolling, Aggression, and Cyberbullying (TRAC-1)[13] have garnered enthusiastic attention from the NLP research community. TRAC-1 focussed on automatic detection of aggression and included a shared task in which the participants were required to provide a classifier to identify overtly aggressive, covertly aggressive and non-aggressive texts from datasets from Facebook. The major contribution of the second edition of this shared task (TRAC-2020)[14] was extending the aggression detection task to multi-lingual datasets. AWL brought together research works on abusive language on diverse dataset sources like Twitter, WhatsApp and news comments.

'OffensEval' shared tasks, organized as part of the SemEval workshops, focusing on offensive language in Tweets, are specifically noted for the high participation rates and contribution of datasets. They are modelled over a hierarchical annotation scheme based on the type and target of the offence (Zampieri et al., 2019). The

[12] https://sites.google.com/site/abusivelanguageworkshop2017/
[13] https://www.aclweb.org/anthology/volumes/W18-44/
[14] https://sites.google.com/view/trac2/home

main contribution of the three-level annotation provided with the dataset (OLID – Offensive Language Identification Dataset) is that it considers the problem of offensive language detection on the whole rather than on specific cases such as hate speech and cyberbullying. Among the three subtasks of OffenseEval-2019 (Zampieri et al., 2019a), the first one A required the participants to distinguish between offensive and non-offensive tweets. Subtask B was to identify the type of the offensive language as targeted or untargeted and the third subtask C was to find whether the target of the offence is an individual, a group or others.

The best-performing methods by the top-ranking teams in Subtask A used BERT (Table 5.2).

**Table 5.2 Best performing teams in Task 6, Subtask A, OffensEval-2019. Performance is measured in terms of f-measure**

| System | Paper | Method | F-measure |
|---|---|---|---|
| NULI | Liu, Li & Zou, 2019 | Transfer learning using BERT | 0.829 |
| Vradivchev_anikolov | Nikolov & Radivchev, 2019 | BERT with pretrained GloVe vectors | 0.815 |
| UM-IU@LING | Zhu, Tian & Kubler, 2019 | BERT with pre-trained word embeddings | 0.814 |
| Embeddia | Pelicon, Martinc & Novak, 2019 | BERT fine-tuned on the OLID dataset | 0.808 |
| MIDAS | Mahata et al., 2019 | An ensemble of CNN, Bidirectional LSTM with attention, Bidirectional LSTB, Bidirectional GRU | 0.807 |

The four top-performing systems use Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019).

The second edition of the OffensEval task (OffensEval-2020, Zampieri et al., 2020) followed the same three-subtasks structure of OffensEval-2019, but provided multi-lingual datasets in five languages – English, Danish, Greek, Arabic and Turkish. Similar to OffensEval-2019, the best performing teams used variants of BERT-like pretrained transformers.

BERT is a language model that pretrains bidirectional representations. It alleviates the issue of unidirectionality in applying pre-trained models for sentence-level or token-level tasks. Unidirectional models use contextualized representations based on either the left or the right context of a word, but not both. In contrast, BERT, as a bidirectional model, considers both left and right context of a word while creating its representation. BERT is shown to achieve superior performance in several standard NLP tasks such as Part-of-speech tagging, parsing and semantic role labelling (Zhang et al., 2020), Question Answering (Devlin et al., 2019) and Aspect Based Sentiment Analysis (Sun, Huang & Qiu, 2019).

As in the other NLP tasks, BERT also brought a further breakthrough in offensive language detection of social media (Mozafari, Farahbakhsh & Crespi, 2019). To address overfitting, there was also an interesting study in aggregating multiple BERT models (Risch & Krestel, 2020). HateBERT, a BERT model trained specifically on a dataset of offensive Reddit (a social media website) comments, showed superior results over the general BERT model (Caselli, Basile & Mitrovic, 2020).

However, this is a very recent technique. For the analysis of swearing/socially offensive language in high-stress tweets, we used a model based on Convolutional Neural Networks, which has been used successfully in offensive posts identification as described in the related works. Thus, though the models used in our study were appropriate and useful for the identification of swearing/socially offensive language, in the light of recent advancements, they cannot be considered state of the art and hence not a contribution to the tasks themselves.

Before proceeding to the description of the experiments, it is critical to clearly define the scope of our study, in the light of the different focal points in the existing research. As is evident from the literature, the term 'offensive language' is overarchingly used for any unacceptable language (Zampieri et al., 2019). This can take multiple forms. Usage of swear words in a harmful sense make online platforms unsafe for children and distasteful for a very large audience. Repeated or intentional insults, threats or profanity directed at a specific individual is classified as cyberbullying (Chun et al., 2020). When the harmful content is targeted at a specific group based on its gender, sexual orientation, religion, ethnicity, race or any other collective identity feature of the members, it is typically recognised as hate speech. For the purpose of our research, we consider the presence of swear words in tweets and do not explore whether or not they have a target. To reinforce the idea that certain words can be used in offensive or non-offensive meaning depending on the context, and that we only consider the former, we use the term 'socially offensive' together with 'swearing' in the later parts of this thesis.

### 5.1.2 Sarcasm Detection

Irony is a form of written or verbal language in which the intended meaning and the apparent meaning are different, often contradictory to each other. It has been described as a form of indirect negation (Giora, 1998) and an 'overt flouting' (Dynel, 2017) of Grice's first maxim of Quality in cooperative principle of conversations (Grice, 1975) which stipulates the speaker says what they believe is true. In linguistics, sarcasm is sometimes defined as 'verbal irony targeted at a victim' (Jorgensen, 1996). Thus, they are related to each other, though not synonymous. Sarcasm is also viewed as a 'special case of irony with an indirect insult' (Dews, Kaplan & Winner, 1995). In the context of tweets, an early quantitative and qualitative study pointed out that irony can be used in two senses – aggressively attacking a specific target or describing an event or a situation (Wang,

2013). The first sense is the sarcastic use of irony. In other linguistic studies, sarcasm has also been defined as a negative, critical form of irony (Colston & Gibbs, 2007).

Thus, though there is a distinction between sarcasm and irony in linguistics, there is no clear consensus on their demarcation. In social media, the hashtags #irony and #sarcasm are often used interchangeably (Wang, 2013). Sarcasm and irony are not distinguished from each other in several studies on their presence in tweets (e.g. Maynard & Greenwood, 2014; Joshi et al., 2016; Tungthamthiti, Shirai & Mohd, 2014). The second Workshop on Figurative Language Processing (FigLang-2020) at ACL-2020 treats sarcasm and verbal irony as the same, characterised by the contrast between the literal and intended meaning. For the purpose of our research, we consider sarcasm as a figure of speech in which the implied negative meaning of the text differs from its literal meaning, and hence do not specifically distinguish between irony and sarcasm.

The characteristic traits of sarcastic tweets have been used as features in studies for sarcasm identification.

**Lexical Features** As for other text classification tasks, statistical bag-of-words features have been used extensively for the identification of sarcasm. N-grams (unigrams, bigrams and trigrams), skip-gram and character n-gram based features are examples.

**Contradictory sentiments**

(23) Isn't it great when your girlfriend dumps you?
(24) I just love when you make me feel like shit.

In tweets like examples 23 and 24, the sarcasm/irony is evident with the expressions of the conflicting emotions. The left and right halves of the same sentence express positive sentiments (Isn't it great, I just love) and negative situations (when your girlfriend dumps you, when you make me feel like shit). The positive sentiments expressed cannot be truthful responses to the negative scenarios. This contradiction is a key feature in identifying sarcasm.

**Hyperbole**

(25) It's Friday and still no news of OC? I'm shocked!!! #sarcasm

In example 26, without the #sarcasm marker, the sarcasm would be less explicit, with the hyperbolic expression ("I'm shocked") as a subtle indicator.

**Other indicators**

Apart from lexical features, concept-level knowledge and coherence between sentences also have been considered in sarcasm identification (Tungthamthiti, Shirai & Mohd, 2014). It is useful in correctly identifying sarcastic sentences like:

(26) It is bliss to work on holidays.

where there is no contradiction between sentiments. The information about the concepts of 'holiday' and 'work' and their perceived sentiments are necessary to identify the sarcastic intent. Also, there could be non-sarcastic tweets with contrasting sentiments like

(27) I like museums, my husband hates paintings.

This approach used the concept lexicon ConceptNet[15], conflicting sentiment scores, lexical features (n-grams, punctuation and special symbols) and a rule-based sentence coherence measure. Tweets were classified as sarcastic or non-sarcastic based on these features using an ensemble of two SVMs. However, the non-sarcastic tweets in the gold-standard dataset were collected using keywords from WordNet; it is not clear if further filtering is performed to ensure sarcastic tweets are not wrongly and inadvertently included in the non-sarcastic gold standard. Also, the conceptual information from ConceptNet could have been more useful for a dataset collected using WordNet keywords, since ConceptNet uses WordNet as a knowledge resource. It is not clear how useful this information will be in datasets collected with different search criteria.

**Table 5.3 Existing research on sarcasm detection from social media posts**

| Authors | Features | Method |
|---|---|---|
| Dong, Li & Choi, 2020 | | BERT-based models |
| Lemmens & Burtenshaw, 2020 | LSTM – hashtags, emoji representation,<br>CNN-LSTM – casing, stopwords, punctuation, sentiment<br>MLP -sentence embeddings<br>SVM- stylometric/emotion based features | Ensemble model with LSTM, CNN-LSTM, MLP and SVM as components |
| Kolchinski & Potts, 2018 | Text features, Author propensity to sarcasm | Bi-directional RNN with GRU cells |
| Subramanian et al., 2019 | Embeddings of emojis and words | Bi-directional RNN with GRU cells |
| Cai, Cai & Wan, 2019 | Image information, metadata of images, text features | Hierarchical fusion model |

[15] http://conceptnet.io/

| | | |
|---|---|---|
| Hazarika & Poria, 2018 | content based features and stylometric and Big-five personality features of users | CNN |
| González-Ibáñez, 2011 | **Lexical** n-grams, dictionary-based (LIWC, WordNet Affect, interjections, and punctuations) **Pragmatic** (emoticons, user mentions) | SVM with Sequential Minimal Optimization and Logistic Regression |
| Reyes, Rosso & Veale, 2013 | Punctuation, emoticons, contradiction indicators, temporal compressors, c-grams, skipgrams, dictionary-based (Whissel's dictionary of affect in language) | Naïve-Bayes and Decision Trees |
| Karoui et al., 2015 | Negation markers/negative polarity markers | SVM with Sequential Minimal Optimization (Weka Toolkit) |
| Riloff et al., 2013 | **Contrast sentiments** positive verb phrases, positive predicative expressions, and negative situation phrases **Sentiment Lexicons** | SVM classifier using LIBSVM |
| Wang, 2013 | Opinion Lexicons | Lexicon-based |
| Rajadesingan, Zafarani & Liu, 2015 | **Behavioural** contrast of sentiments, complexity of expressions, familiarity with persons and language | L1 regularized Logistic Regression |
| Tungthamthiti, Shirai & Mohd, 2014 | **Concept knowledge** ConceptNet **Contradicting sentiments, Lexical** n-grams punctuation special symbols **Rule-based coherence measure** | Ensemble model of two SVMs |
| Liebrecht, Kunneman & Van den Bosch, 2013 | uni, bi, tri-grams weighted by $X^2$ metric | Balanced Winnow classifier |
| Maynard & Greenwood, 2014 | Hashtag tokenization and analysis | rule-based classification (based on GATE and hashtag analysis) |
| Zhang, Zhang & Fu, 2016 | Local and contextual features of tweets | Bidirectional gated RNN |
| Poria et al., 2016 | Sentiment, emotion and personality features | Convolutional Neural networks with SVM |

Sarcasm has also been modelled behaviourally, with sentiment contrasts, expression complexity, prosodic and structural variations and familiarity with language and persons as features, using L1 regularized logical regression (Rajadesingan, Zafarani & Liu, 2015). This method gave 83.46% accuracy on a dataset of 9104 tweets collected using the Twitter streaming API, using the hashtags #sarcasm and #not as keywords. Though

these are promising results, the top 10 predictive features for sarcasm included users' past behaviour such as number of past sarcastic posts and percentage of past posts with specific sentiment scores.

More recent works also leverage user-specific information such as stylometric features and big-five based personality classes (Hazarika & Poria, 2018). Author propensity to sarcasm as evidenced from past messages is also a useful feature in sarcasm detection (Kolchinski & Potts, 2018). However, since our dataset does not focus on user-specific information, these features are not relevant for our models for sarcasm detection. Including emoji representations also considerably enhanced sarcasm detection in tweets and Facebook posts (Subramanian et al., 2019). There have also been efforts to utilize the multi-modality of social media posts in the task of sarcasm detection, with image-based attributes (Cai, Cai & Wan, 2019). Though the analysis of images together with the text features resulted in better performance, it also came at greater computational complexity and cost. In our research, the focus is not on improving the sarcasm detection performance but to analyse the presence of sarcasm in high-stress and low-stress tweets. Hence, we consider only text features. Including emojis and image-based features in the sarcasm analysis could be a future improvement.

Table 5.3 summarises the lexical, sentiment and behavioural features and machine learning and deep learning methods used for sarcasm detection from social media posts.

While the models were assessed on different datasets and the results are not directly comparable, many use some common features which we utilize in designing our classifier for sarcasm identification in our dataset.

Considering the proliferation of research in irony/sarcasm detection methods, several related shared tasks have been organized too. In SemEval-2018, task3 (Van Hee, Lefever & Hoste, 2018) focused on irony detection in English tweets. It proposed two challenges to the participants: identify tweets as ironic or non-ironic (subtask A) and further identify the category of irony (polarity clash, situational, another type or non-ironical) (subtask B). Similar to the approach by the majority of related works, this task doesn't distinguish between irony and sarcasm and the tweets were collected using the hashtags #sarcasm, #irony and #not. We adopt this approach in our study too.

Deep learning methods and ensemble methods were found to give the best performance (Table 5.4).

**Table 5.4 Performance of the top-performing teams for an irony identification task in SemEval-2018**

| Team | Paper | Method | F-measure |
|---|---|---|---|
| THU_NGN | Wu et al., 2018 | Densely connected LSTMs with pre-trained word embeddings, sentiment features using AffectiveTweet package and syntactic features (POS tags and sentence embedding features) | 0.705 |
| NTUA_SLP | Baziotis et al., 2018 | Ensemble classifier of word and character-based bidirectional LSTMs to capture syntactic and semantic information | 0.672 |
| WLV | Rohanian et al., 2018 | Ensemble voting classifier (SVM and Logistic Regression) with pre-trained word and emoji embeddings, sentiment contrasts between elements in a tweet | 0.650 |
| IITBHU-NLPRL | Rangwani, Kulshreshtha & Singh, 2018 | XGBoost classifier with pre-trained CNN activation using DeepMoji and hand-crafted features like polarity contrast, context incongruity, WordNet similarity, and URL counts | 0.648 |
| NIHRIO | Vu et al., 2018 | Neural-networks-based architecture (i.e. Multilayer Perceptron) using lexical, syntactic and semantic (GloVe word embeddings) polarity features (from the Hu and Liu Opinion Lexicon) Brown Cluster features | 0.648 |

## 5.2 Methods

The approaches for detecting the socially offensive posts/swearing and sarcasm in tweets were similar. As described in Chapter 4, we had a dataset consisting of 2000 tweets which were annotated for swearing/socially offensive content and sarcasm. Based on the analysis of the related works, we chose a few classification models for both these tasks. The performances of these methods were evaluated on the annotated dataset of 2000 tweets, and the best-performing model was chosen for classifying 10000 tweets for both swearing/socially offensive content and sarcasm.

### 5.2.1 Dataset Annotation

The presence of socially offensive/swearing posts in the dataset consisting of 12000 tweets (described in Section 4.2.3), in which 8871 tweets were identified as stressful tweets and the rest no-stress tweets, was examined. A collection of 2000 tweets (500 from each domain – traffic, airlines, politics and personal events), randomly chosen from the 12000 tweets, was annotated for presence of socially offensive content. The annotation process was described in the following section.

#### 5.2.1.1 Socially offensive/swearing Posts

**Annotation Question**

Does the tweet contain swearing/socially offensive language?

**Explanation**

The annotators were provided with a tweet and were required to mark whether the tweet contained socially offensive language. This was difficult since this classification was often subjective. What one individual perceived as profanity could have been normal language for another. This necessitated clear measures to ensure the annotators reached acceptable agreement in their judgments.

Ambiguous words like hell, suck can be used with socially offensive and non-offensive intent; the annotators were required to use their judgment in such contexts. The 'socially offensive' term is used together with 'swearing' throughout the thesis to highlight this demarcation.

**Guidelines**

- Mark the tweet as socially offensive/swearing if:
    - it contains the usage of one or more swear words in a socially inappropriate way or
    - it contains words which might have other non-offensive senses, but used in the context of the tweet to insult, hurt the sentiments of, or downgrade the target
- In case of socially offensive tweets, mark the specific word which had the offensive intent.

**Examples**

8. Getting sucked into a defensive stance on a bullshit 'border security' discussion only helps #Conservatives (Offensive-but 'bullshit' is the offensive word)
9. @username: @username Aaah look. Another #corbynite with the same bullshit. Originality, not your strong point? (Offensive; 'bullshit')
10. Two days till the exam and I can't believe I wasted one full morning looking at videos of blue whale sucking in bellyfuls of krill!! (non-offensive)

#### 5.2.1.2 Sarcasm

**Annotation Question**

Does the tweet contain sarcasm?

**Explanation**

The annotators were provided with a tweet and required to mark whether the tweet contains sarcasm or not.

It is difficult even for human annotators to correctly identify sarcastic tweets. However, there are markers like #irony and #sarcasm which unambiguously and explicitly categorize a given tweet as sarcastic (Maynard & Greenwood, 2014). In the absence of explicit markers, contradicting sentiment expressions and hyperbolic adjectives (Liebrecht, Kunneman & Van den Bosch, 2013) are likely indicators of sarcasm.

**Guidelines**

- For the purpose of this annotation, sarcasm is defined as a figure of speech in which the implied negative meaning of the text differs from its literal meaning.
- Typical sarcasm markers are hashtags (including but not limited to #sarcasm, #irony, #not and #notreally), hyperboles, apparently positive responses to a negative event or circumstance.

**Examples**

11. I fell down on the sidewalk and now my hip hurts badly; awesome day so far
12. Excellent weather, it hasn't stopped raining since morning

### 5.2.2 Socially Offensive Posts/Swearing Detection

Out of these annotated tweets, the presence of offensive content targeted at a specific ethnic/gender/sexual orientation was found to be low (0.75%). Hence, we focused our study on the presence of socially offensive posts and not specifically on hate speech. That is, the tweets with a socially inappropriate usage of a swear word or with an intension of insulting/hurting/downgrading a target are considered in our study. We do not examine whether the subject of offence, if any, is targeted because of aspects of their identity, such as ethnicity, religion, gender or nationality.

Though the objective was to identify socially offensive posts, we decided against using a lexicon-approach, because it might not be useful in differentiating the offensive and non-offensive usage of the same word. Inspired from our analysis of classification models for the related problem of hate speech, we implemented three classifiers- CNN, SVM and Logistic Regression- and compared its performance on the annotated dataset of 2000 tweets and the OffensEval dataset. The intention was to choose the best performing classifier and use it in the classification of the 10000 tweets which are not annotated.

CNN - We implemented a Convolutional Neural Network using Keras. The CNN was designed based on an existing CNN architecture which is simple but achieved good results in several NLP tasks including sentiment analysis and question answering (Kim, 2014). The implemented CNN has one embedding layer, one convolution layer (activation function as ReLU), one maxpool layer, one fully connected layer and one softmax layer with dropout. The shallow design was adopted in previous comparable research to successfully address the overfitting in small datasets (Park & Fung, 2017). The parameters were set by the Random Search method, using RandomizedSeachCV class in Keras classifier wrapper in Sci-kit learn API. Words in the tweet were represented by pretrained vectors from Word2Vec model trained on 400 million tweets (Godin et al., 2015).

Logistic Regression- liblinear library, l2 regularization

Support Vector Machine – LibSVM based implementation, linear kernel

The tweets were tokenized using NLTK TweetTokenizer package. URLs and usernames were removed. In some of the words swearing was present in the hashtags (e.g. #f**kUnited). Hashtag symbol was removed and the constituent words were split based on CamelCase notation, similar to the pre-processing approach in Chapter 4. The hashtags which did not follow CamelCase were manually segmented. For hashtag segmentation, we did not use automatic segmentation tools, as the number of unique non-CamelCase hashtags was rather low (268 in the dataset of 12000 tweets). Words which were in the NLTK stopwords list were removed.

In the 2000 tweets used for annotation, the annotators noted down the obfuscated swear words and their presence was quite low (87 out of 2000 tweets). Hence, we adopted a simpler approach of pre-processing the tweets using hand-crafted rules. The swear words are normalized by converting digits or symbols in a word to its letter form ($->s, #->h, !->i, + -> t, 0->o, 9->g, @->a), similar to the approach adopted in a study on cursing in Twitter (Wang et al., 2014). The symbol '#' at the start of a sentence was not converted as it was the norm for a hashtag.

The features used for the classification models were the following:

- N-grams - Features are encoded as binary scores based on the presence or absence of word n-gram and character n-grams in a Tweet. We consider top 1000 uni-, bi-, and trigrams based on tf-idf for words and characters.
- The number of capital letters, emoticons, and punctuation marks.
- Number of characters in the tweet and average number of characters in the words in a tweet

- Lexicon – We constructed a list of English swear words from existing research lexicons (Thelwall, 2008; Wang et al., 2014). The number of words in the tweet which are present in the swear word lexicon is taken as a feature. To account for obfuscation which is unresolved by the pre-processing (e.g. s%it), we consider the edit distance of words in the tweet to those in the lexicon. An exact match adds 1 to the count, whereas an edit distance of 1 (e.g. bitch and b_tch) adds 0.5.

We compare the performance of these methods (CNN, SVM and Logistic Regression) on two datasets – the dataset from the OffensEval task of SemEval-2018 and the human-annotated tweets dataset from Chapter 4, based on the F-measure (Tables 5.5, 5.6). The scores are averaged over 5-fold cross-validation.

**Table 5.5 Performance of the CNN to detect swearing/ socially offensive tweets in comparison with the baselines, on the OffensEval-2018 dataset**

| Method | Performance (F-measure) |
|---|---|
| CNN | 0.796 |
| SVM | 0.754 |
| LogReg | 0.748 |

**Table 5.6 Performance of the CNN to detect swearing/ socially offensive tweets in comparison with the baselines, on the human-annotated tweets corpus (2000 tweets)**

| Method | Performance(F-measure) |
|---|---|
| CNN | 0.803 |
| SVM | 0.724 |
| LogReg | 0.726 |

Based on this better performance of CNN on both datasets, we use it for classifying the remaining 10000 tweets as socially offensive or non-offensive.

### 5.2.3 Sarcasm Detection

Based on our analysis of related work on sarcasm detection, we chose to implement a Multi-Layer Perceptron (MLP), an SVM and a Logistic Regression model for the classification of tweets as sarcastic or not. An LSTM-based architecture was used by the best performing team in the shared task 3 for SemEval-2018, but it also made use of extensive features. MLP gave comparable results with a far simpler architecture and features. We examined the same dataset of 12000 tweets (within which 8871 tweets were stressful and the rest no-stress) for presence of sarcasm, as we did for socially offensive language. As described in Section 5.2.1 , a collection of 2000 tweets (500 from traffic, airlines, politics and personal events), was randomly chosen from the 12000 tweets, and was annotated as sarcastic or not. The performance of the three models was evaluated on the

SemEval-2018 task 3 dataset and the human-annotated dataset of 2000 tweets. The best performing model was selected for the classification of the remaining 10000 tweets.

The tweets were pre-processed in two steps - tokenization and normalization. The tweets were tokenized using NLTK TweetTokenizer package. URLs and usernames were removed. Hashtags were segmented based on CamelCase notation; the ones which did not follow camelCase notation were manually segmented, as they were quite low in number.

Multilayer Perceptron was implemented based on the NIHRIO solution (Vu et al., 2018) for irony detection at the SemEval-2018 task 3. The architecture consisted of an input layer, two hidden layers with ReLU activation and a softmax output layer. A feature vector represented the tweet as the concatenation of the following lexical, syntactic, semantic and polarity features.

**Lexical features:** Inspired from the MLP implementation, we use 2 and 3-grams at the word level. The top 1000 n-grams based on tf-idf are chosen for each type of n-grams. The numbers of capital letters, punctuations, emoticons, words and characters in a tweet were also considered as features.

**Syntactic features**: POS-tags and their tf-idf values are syntactic features and feature values. NLTK toolkit is used to annotate POS tags.

**Semantic features:** The tweet embedding is calculated as the average of word embeddings. Each word in the tweet is represented by the Word2Vec vector representation.

**Polarity features:** Many sarcastic tweets are characterised by an apparently positive response to a negative situation. Using the SentiStrength lexicon (Thelwall et al., 2010), we find the positive and negative sentiment score of each tweet. SentiStrength assigns two scores to each tweet: positive sentiment in the range of +1 to +5 and negative sentiment in the range of -1 to -5. These scores were used as the polarity features.

This MLP implementation and the baselines (SVM and Logistic Regression with the same feature set) are evaluated on two datasets for sarcasm detection task – Test dataset for SemEval-2018 task 3, human-annotated corpus of 2000 tweets.

**Table 5.7 Performance of the Multi-Layer Perceptron to detect sarcastic language in comparison with the baselines, on the SemEval-2018 task-3 dataset.**

| Method | F-measure |
|---|---|
| MLP | 0.63 |
| SVM | 0.61 |
| LogReg | 0.58 |

**Table 5.8 Performance of the Multi-Layer Perceptron to detect sarcastic language in comparison with the baselines, on the annotated tweets' corpus (2000 tweets).**

| Method | F-measure |
|---|---|
| MLP | 0.69 |
| SVM | 0.65 |
| LogReg | 0.63 |

The Multi-layer Perceptron gave a clearly better performance compared to the baselines and we chose it for classifying the remaining 10000 tweets as sarcastic or non-sarcastic. Over 5-fold cross-validation, the MLP's parameters were tuned for optimum performance (epochs 100, learning rate $10^{-4}$). L2 regularization was chosen to minimise cross entropy loss.

The annotators had marked the linguistic cues which helped them identify sarcasm in the dataset of 2000 tweets (Table 5.9).

**Table 5.9 Linguistic cues present in the sarcastic tweets in the annotated dataset of 2000 tweets**

| Linguistic cue | Percentage of tweets |
|---|---|
| Exclamation/repeated punctuation (only) | 9.4% |
| Hyperbole (only) | 7.5% |
| Explicit markers (only) | 5.7% |
| Contrasting emotions (only) | 14.4% |
| punctuation+hyperbole | 11.8% |
| punctuation+explicit markers | 5.6% |
| Punctuation+hyperbole | 6.3% |
| Hyperbole+explicit markers | 7.1% |
| Hyperbole+contrasting emotions | 18.2% |
| Explicit markers+contrasting emotions | 7.9% |

| | |
|---|---|
| Explicit markers+contrasting emotions+hyperbole+punctuation | 3.1% |
| Other | 3.2% |

Based on this distribution, 46.5% of sarcastic tweets in the dataset of 2000 tweets contained hyperbole and 43.2% contained contrasting emotions. Feature weighting could likely improve the performance of sarcasm classifiers; however, due to the small size of the dataset, we decided against it to avoid overfitting. With a larger training dataset, assigning weights to the more relevant features could likely improve the performance of MLP classifier for sarcasm.

## 5.3 Results

### 5.3.1 Prevalence of Socially Offensive/Swearing Posts and Sarcasm

We use the CNN (Section 5.2.2) for classifying the dataset of tweets as socially offensive or non-offensive. Similarly, the MLP described in Section 5.2.3 is used for the sarcasm classification of the same dataset.

**Table 5.10 Percentage of tweets which are sarcastic or swearing (in the dataset of 12000 tweets - 8871 tweets with stress scores -2, -3, -4 or -5 and 3129 tweets stress score -1)**

| Domain | Type | % of swearing/socially offensive tweets | % of sarcastic tweets |
|---|---|---|---|
| Airlines | High-stress | 14.72 | 9.64 |
| | No-stress | 6.01 | 1.94 |
| Personal events | High-stress | 21.9 | **12.05** |
| | No-stress | 10.26 | 7.73 |
| Politics | High-stress | **18.56** | 10.68 |
| | No-stress | 9.61 | 6.67 |
| Traffic | High-stress | 11.59 | 9.36 |
| | No-stress | 10.16 | 7.23 |

In all domains, high-stress tweets have a higher percentage of socially offensive tweets and sarcastic tweets, compared to no-stress tweets (Table 5.10). Also, the percentage of socially offensive tweets is significantly higher in all domains compared to sarcastic tweets. This indicates socially offensive language being used more in the dataset to express stress, in comparison to sarcasm. The largest difference in the percentage of socially offensive tweets between high-stress and low-stress corpus, occurs for personal events (11.64%) and politics (8.95%). Politics has the highest percentage of socially offensive language in the high stress corpus, whereas the personal events corpus has the highest percentage of sarcastic language (both marked in bold in Table 5.10).

The stressors identified using the annotation scheme in Chapter 3 for stressors are used for the further analysis of this classification results (Figures 5.1 to 5.4). The figures present, for each stressor, the percentage of all the tweets that was classified as offensive and sarcastic.

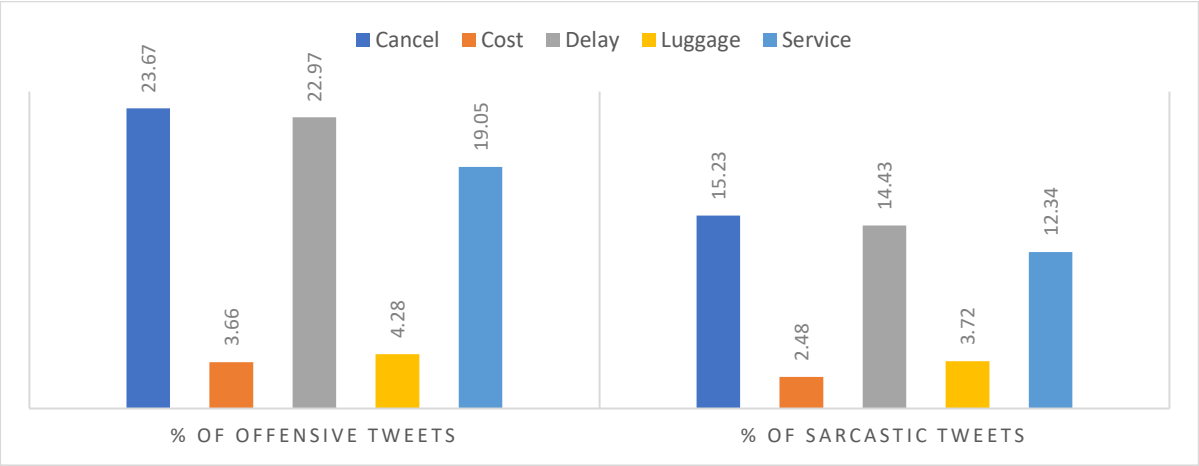


**Figure 5.1 Percentage of offensive and sarcastic Airlines tweets (n=2493 tweets)**

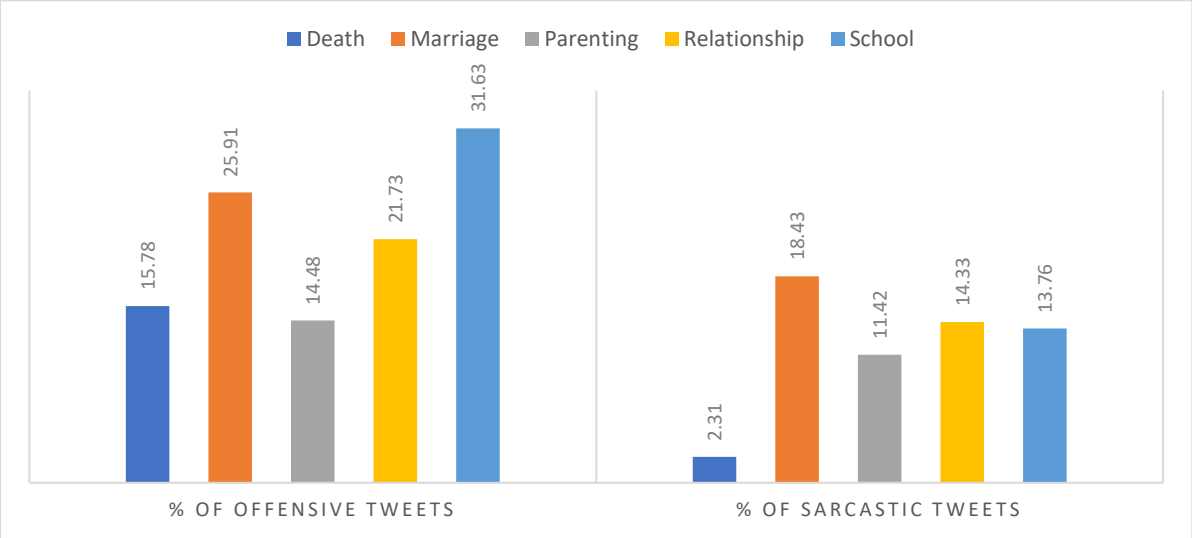

**Figure 5.2 Percentage of offensive and sarcastic Personal Events tweets (n=1838 tweets)**

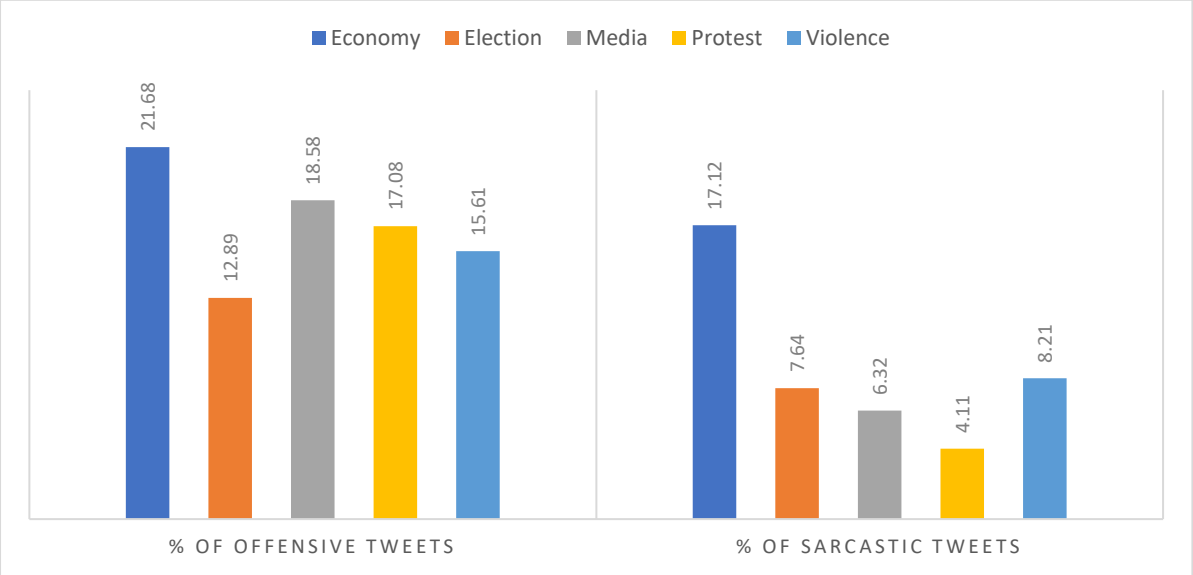


**Figure 5.3 Percentage of offensive and sarcastic Politics tweets (n=1927 tweets)**

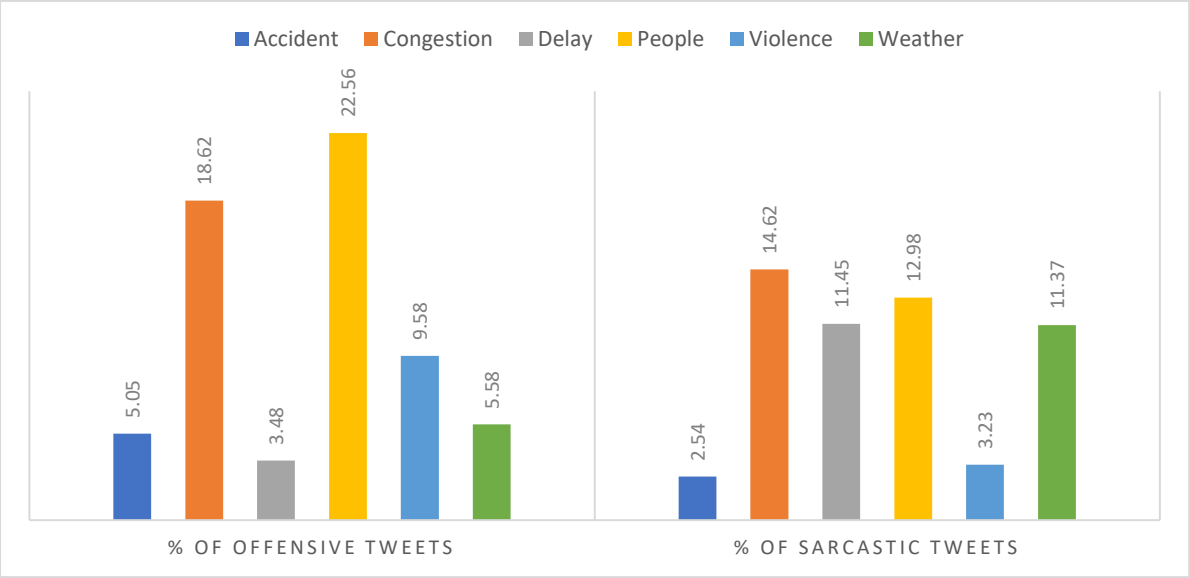


**Figure 5.4 Percentage of offensive and sarcastic Traffic tweets (n=2613 tweets)**

In airlines, cancellation of flights is the stressor with the highest percentage of offensive and sarcastic tweets, followed by delay and service. (euphemistic spelling for swear words with asterisk symbol (*))

(28)@British_Airways Excuse my French: is this a F***ING JOKE? This flight is delaying an hour every past minute!!! (delay)

(29)Thanks for f***ing up my day with your cancellations @united ... Never again. #f***united #unitedAIRLINES #annoyed (cancellation)

It can be noted that (30) uses both socially offensive language and sarcasm to express stress.

Among political tweets, the subset with economy as the stressor has the highest percentage of socially offensive and sarcastic tweets.

(30)@SkyNews Here we go again!! #TheresaMay lying through her teeth about the economy!!! She couldn't give a rats a** about anything

In tweets related to personal events, marriage has the highest percentage of socially offensive tweets, and tweets on relationships have the highest percentage of sarcastic tweets.

(31)I hate her!! what a shame the poor groom's bride is a w****

Finally, in traffic tweets, the highest percentage of offensive tweets is observed in the category of tweets with 'people' as the stressor.

(32)The joys of being stuck in London rush hour traffic after an onsite visit #HaventMovedForAnHour (sarcasm) congestion)

(33)You retard you will have to explain that to me young man (offensive) (people)

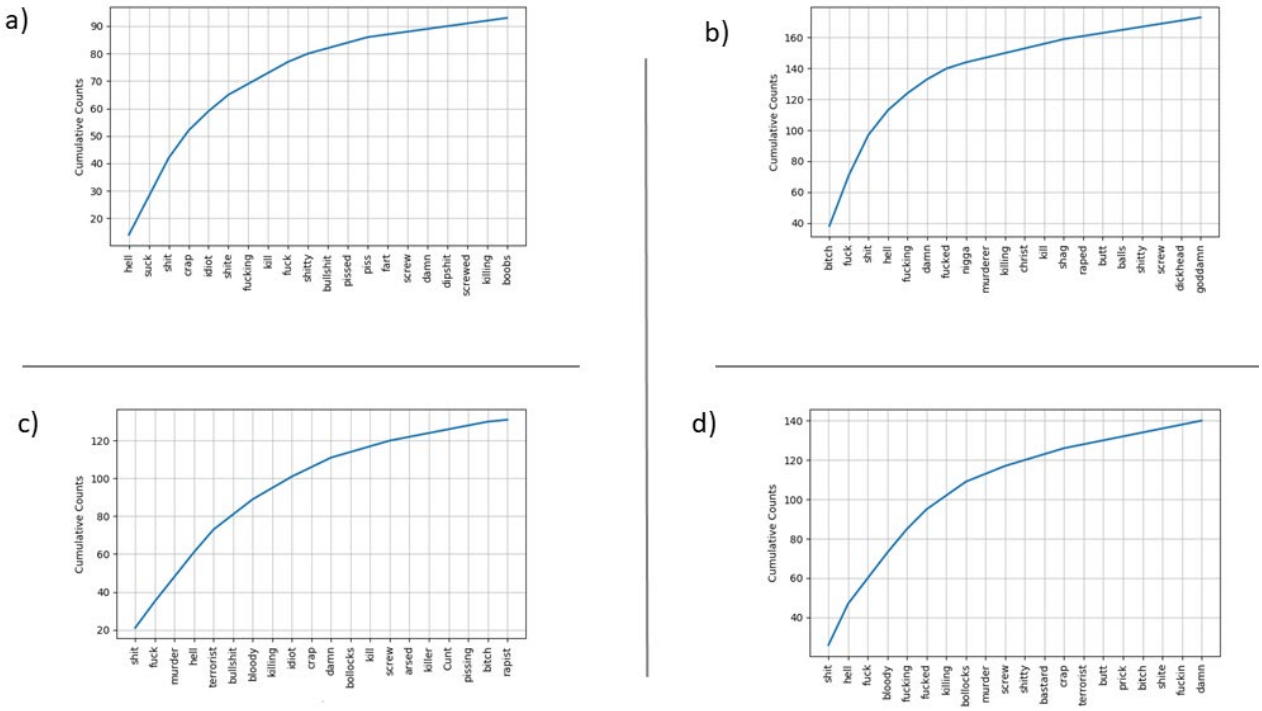

**Figure 5.5 Cumulative frequencies of the top 20 swear words in the high-stress (8871 tweets from four domains) corpus for a) airlines b) personal events c) politics and d) traffic**

The swear words used in high-stress corpus differ according to the domain (Figure 5.5).

### 5.3.2 Lexical Analysis of Stress-Related Terms

A lexical analysis was done to extract and analyse words which occur characteristically in the high-stress corpus. We used the Python implementation of SAGE (Sparse Additive Generative Model) (Eisenstein, Ahmed & Xing, 2011) extracting the keywords. SAGE compares the given text to a background distribution and chooses the discriminating keywords. In each domain, the high-stress tweets (TensiStrength score -2 to -5) and low-stress tweets (TensiStrength score -1) were put in separate text files. The distinguishing keywords extracted from the high-stress corpus in each domain are given in Figure 5.6, Figure 5.7, Figure 5.8,Figure 5.9.

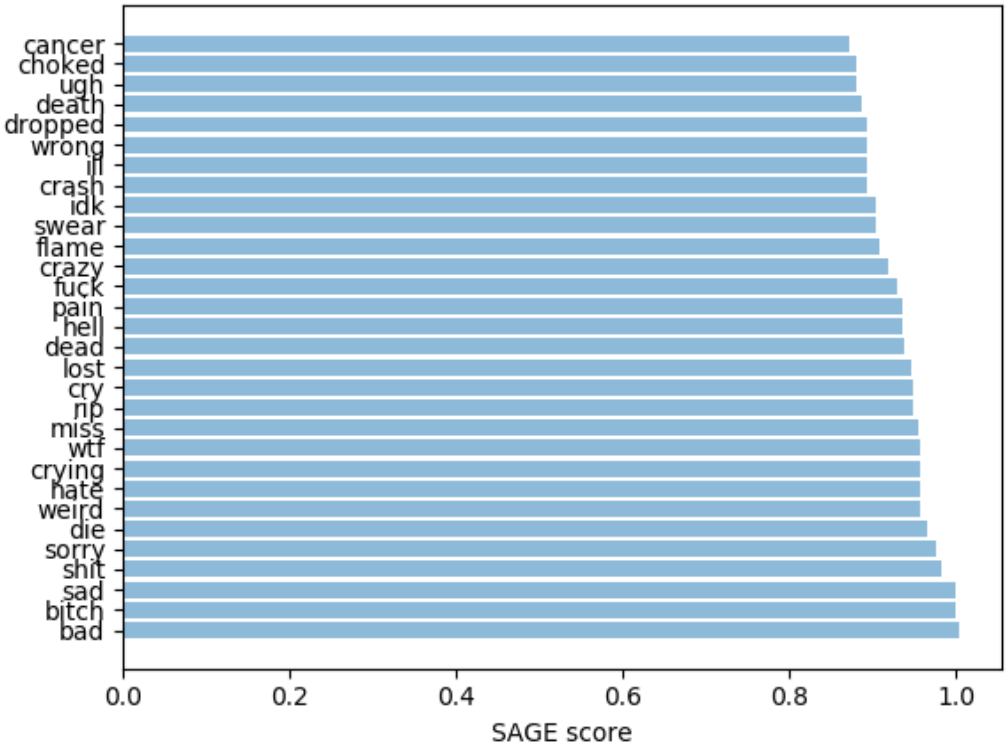


**Figure 5.6 Top discriminating unigrams in high-stress tweets compared to low-stress tweets in the Personal Events domain**

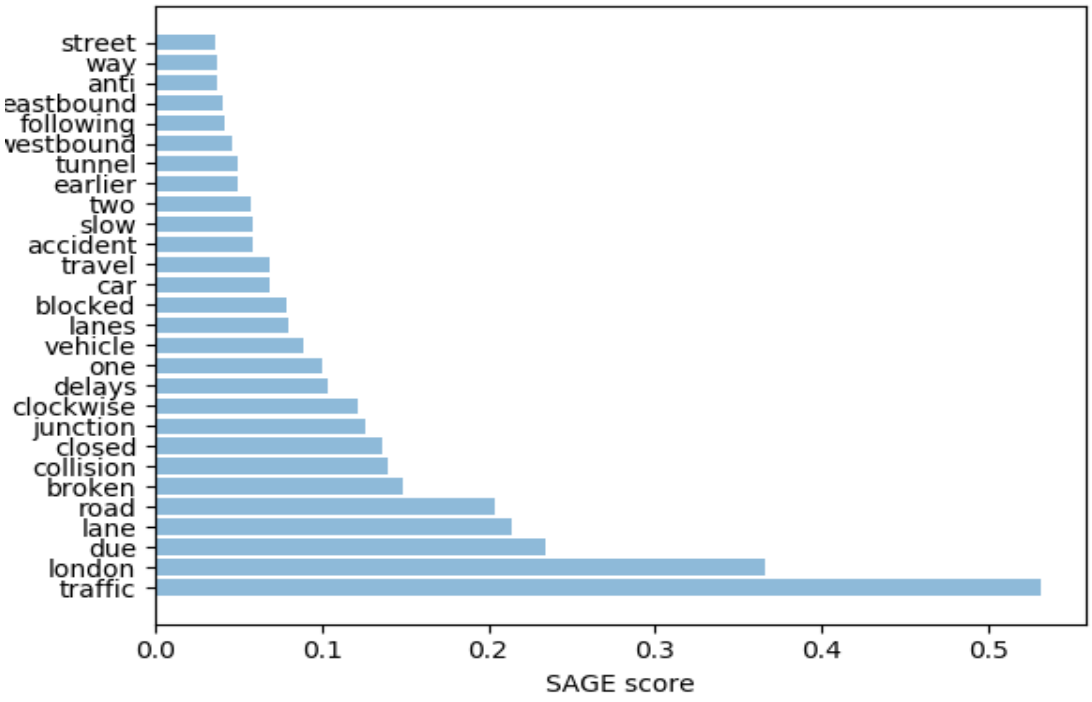


**Figure 5.7 Top discriminating unigrams in high-stress tweets compared to low-stress tweets in the Traffic domain**

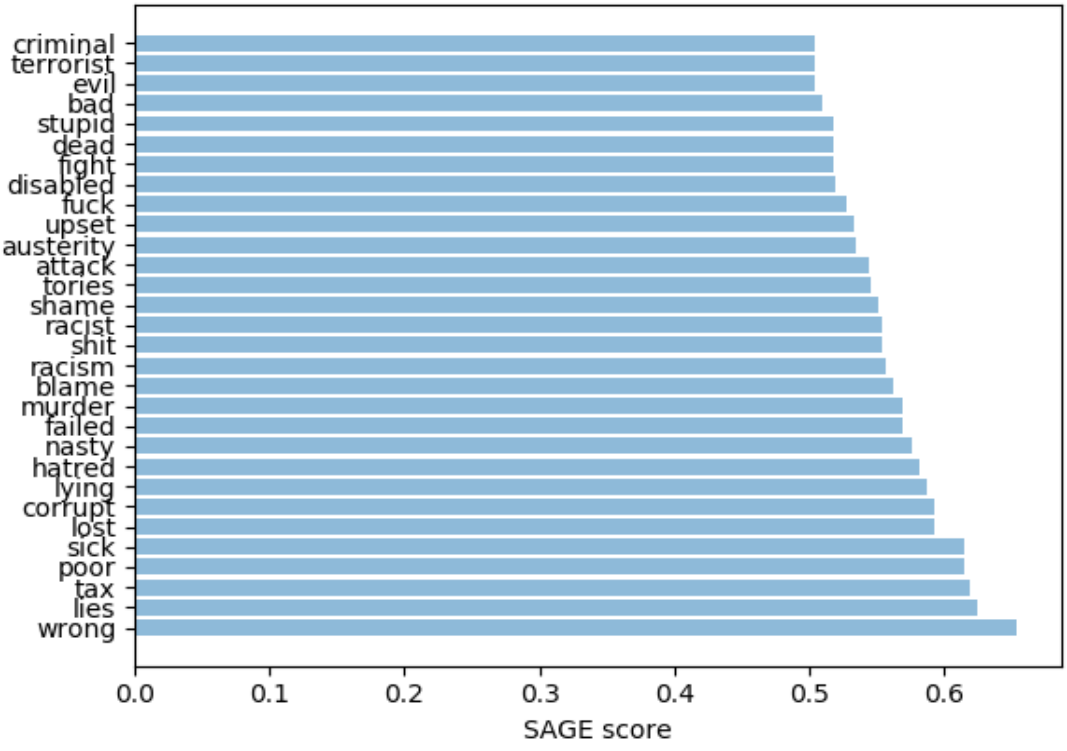

**Figure 5.8 Top discriminating unigrams in high-stress tweets compared to low-stress tweets in the Politics domain**

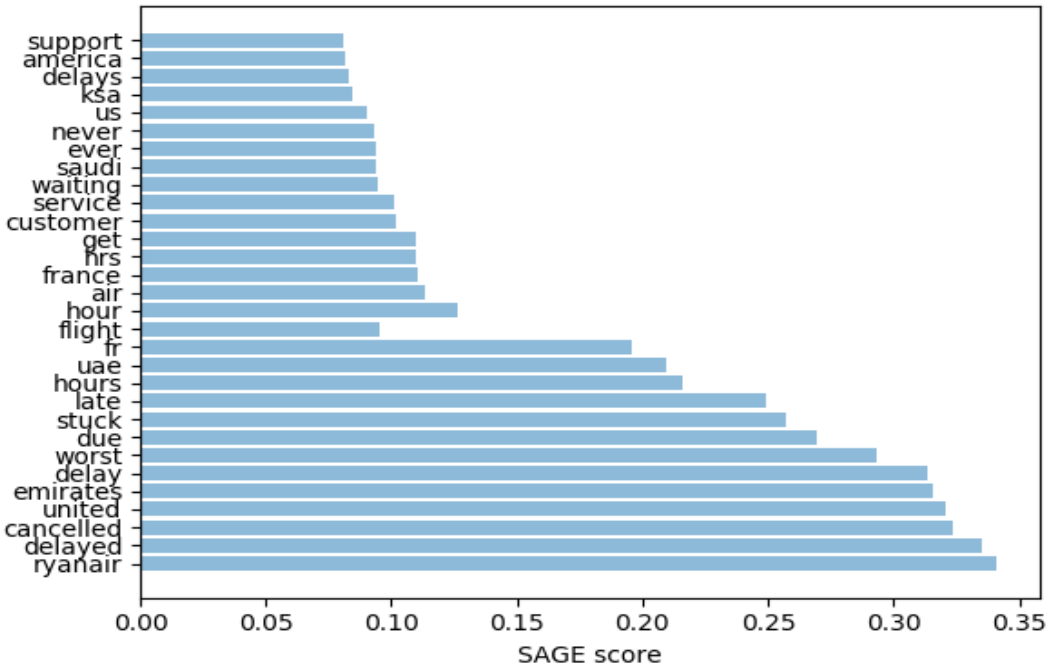


**Figure 5.9 Top discriminating unigrams in high-stress tweets compared to low-stress tweets in the Airlines domain**

There is a deviation from the observed linguistic patterns from existing research on the language of stress. Previous research (Guntuku et al., 2018) has pointed out a significant correlation between self-oriented linguistic constructs ('me', 'I', 'I don't', 'I hate') and high scores of psychological stress. Also, there were direct references to mental health issues in high-stress Facebook posts analysed in that research, such as 'depression', 'depressed', 'anxiety','bipolar'. However, these were not found in our analysis. One reason could be that the analysis had focused on self-diagnosed and prolonged stressful conditions whereas our study does not consider individual psychological health and is concerned with short-term stressful situations. Though psychological conditions like depression, anxiety are not reflected in the lexical analysis, some words describe physical health (Personal Events domain: cancer, death, ill, pain, dead). Compared to low-stress tweets, high-stress tweets have more negative affect terms, in politics and personal events domains. In terms of emotions, these negative affect terms appear to express sadness in the personal events domain (cry, pain, miss, hate, sad) and aggressiveness in the politics domain (shame, blame, nasty, fight). Other categories of terms observed in SAGE analysis of high-stress politics tweets are crime-related (criminal, terrorist, attack) and economic concerns (corrupt, austerity, tax). The analysis also points to a high occurrence of geographical

entities in the high-stress corpus of travel-related domains (airlines: us, Saudi, ksa, uae, America; traffic: clockwise, anticlockwise, southbound, westbound).

**Table 5.11 Examples of top discriminating unigrams in the corpus of 8871 tweets with stress expressions, according to the SAGE analysis (broad categories of unigrams given in bold)**

| Domain | Words |
|---|---|
| Airlines | **geographical** us, Saudi, ksa, uae, America **conflict** aggression delay cancel hours late |
| Personal Events | **Illness/death** cancer death ill pain **Offensive terms** swear fuck wtf shit bitch hell **abbreviations** wtf idk rip |
| Politics | **Violence** criminal terrorist fight murder attack **Adjective with negative affect** evil bad dead wrong poor sick list **Offensive Language** stupid fuck nasty |
| Traffic | **Directions** clockwise anticlockwise southbound eastbound westbound **travel** vehicle road lanes tunnel lane street **Conflicts** delay block collision broken |

There are offensive terms in the unigrams identified using SAGE in two domains – politics and personal events. This agrees with our finding that the highest percentage difference in the presence of offensive terms between the high-stress and no-stress corpora is in these two domains (Table 5.10).

## 5.3 Discussion

### 5.3.1 Error Analysis

Misclassification by our classification models in the annotated dataset of 2000 tweets is discussed here.

#### 5.3.1.1 Socially Offensive/Swearing Posts

Some tweets with swear words which had multiple character replacements not identified by the obfuscation pre-processing and which were often misclassified as non-offensive.

(34) WARNING M25 cw into Dartford tunnel is totally f%%ked! I lost an hour

(35) These are dirty c##nts can't even say when they can reimburse my tickets for cancellation

The pre-processing only considers a list of typical character replacements (Section 5.2.2 ). Swear words which do not follow these patterns were challenging for our classifier.

Another reason for misclassification was the occurrence of rare swear words or modified swear words which occurred sparingly in the dataset.

(36) Those faggots who run this country!! Austerity is social murder

(37) Those supporters are intelligent enough to see #Corbyn for the deceitful, evasive cretinette that he is

Code-mixed tweets, though rare in the dataset, had swear words in other languages which were not identified by our classifier.

(38) Sab ch**iye hain! They just keep making excuses for this delay #qatarairways

#### 5.3.1.2 Sarcasm

The MLP classifier failed to identify certain expressions of sarcasm in the annotated dataset of 2000 tweets. Among the incorrect classifications, there were also non-sarcastic tweets with hyperbole or contrasting sentiments.

(39) Delighted to travel again after chemo! Fatigue has set in but life continues #londontraffic

This example is from the traffic dataset. The word 'fatigue' indicates negative and 'delighted' indicates positive affect. Though there is no sarcasm in this tweet, it was incorrectly classified as sarcastic. This could be due to the presence of contrasting sentiments which is a common feature of sarcastic texts. It is noteworthy that 'delight' is expressed in the context of 'chemo', a typically stressful medical treatment and yet the tweet overall is non-sarcastic. It is contrary to the assumption in earlier works (Riloff et al., 2013) which take 'positive sentiment in a negative context' as a predictive feature for sarcasm identification.

(40) These traffic rules are hilariously useless

This tweet was annotated as sarcastic by one of the three annotators. The combination of positive and negative adjectives together is misleading even for human annotators. In the gold-standard, the label of this tweet was 'non-sarcastic' based on majority voting. The MLP classifier incorrectly classified it as sarcastic, possibly because of the contrasting sentiments.

(41) You lot should be ashamed of your service! Absolutely terrible@!!!!!!!!!!!!

This was another example of a non-sarcastic tweet misclassified as sarcastic. Contrary to the previous examples, there is no contrasting context or sentiment, yet repeated punctuations and hyperbolic expressions possibly mislead the classifier.

There were sarcastic tweets which were classified as non-sarcastic by the MLP classifier.

(42) It is not urgent, if you are not too busy, please hurry up and fix your website- It's been HOURS. Thank you! #united

Here, fixing the airlines website is an urgent requirement for the user, but the displeasure is expressed through sarcasm.

(43) We have been delayed 3 hours in departure. Think there will be any more delay? Thank you in advance

'Thank you' in response to a delay is a sarcastic expression; however, this is not identified by the MLP classifier.

These errors point to the misleading features of tweets which could result in misclassification as either sarcastic or non-sarcastic. A larger training dataset could possibly improve the performance of the classifier.

### 5.4.2 Limitations

The study points to the different ways in which stress is expressed in different domains. However, the scope of the study is limited to specific forms of abusive language. The classifiers were modelled to identify socially offensive language; they do not further classify if the offence is directed at a specific individual, such as in cyber-bullying. Expressions of offence directed at a specific target group, hate speech, was not studied either due to its sparse presence in the dataset. However, this absence of hate speech cannot be realistically assumed in case of other domains. Its presence could be more significant in datasets collected by a wider range of hashtags. In those cases, the classifier would need to be modified accordingly to address these different types of abusive language.

Another limitation is that the tweets are classified with binary labels as offensive or not. More granular labels with the severity of the offence (e.g. light, moderate and heavy) could be assigned to the tweets. It would be interesting to observe the relative presence of such weighted offences in the datasets.

There were challenging cases like character replacements, rare swear words, code-mixed tweets (swearing/socially offensive classification) and indirect/implicit expression (sarcasm classification) which were often misclassified by our classifier models.

In both sarcasm and swearing identification, the model selection was based on a survey of existing methods and features. In the last year, BERT has been established as a highly efficient solution for NLP tasks such as these. Incorporating such more recent state-of-the-art methods could improve the correctness and the classifications and hence modify the insights from high-stress tweet analysis.

## 5.5 Conclusions

In this study, we analysed the linguistic features of the test collection of high-stress Tweets. We investigated how stress gets expressed in tweets; from traditional psychology studies, swearing and sarcasm are ways people give vent to stress. We constructed a Convolutional Neural Network for identifying socially offensive language. A Multilayer Perceptron was constructed for identifying sarcastic tweets. The performances of these systems were compared to and were found better than the baselines in the respective tasks.

Tweets with stress were found to have higher presence of socially offensive language and sarcasm in all the four domains we studied. The percentage of socially offensive tweets were significantly higher in all domains in the high-stress corpus compared to the percentage of sarcastic tweets.

The linguistic patterns of the high-stress tweets were analysed using SAGE. The absence of self-oriented linguistic constructs ('me', 'I', 'mine') and direct references of mental disorders was an important deviation from the observations in comparable studies on high-stress content of Facebook, and we attributed it to our dataset focusing more on short-term stressful scenarios and not on individual declarations of psychological health. Terms related to aggressiveness (politics domain), crime (politics), negative affect terms (personal events) and geographical entities (traffic and airlines) were observed to have high occurrence in high-stress tweets. These linguistic clues could be used as potentially predictive features in classification models for identifying high-stress tweets.

The pivotal contribution of this chapter is the comparative analysis of socially offensive language and sarcasm in high-stress and low-stress tweets. Both have been previously established as ways used by people to vent negative mental states. However, usage in social media in relation to psychological cues in text has not been studied so far. In that light, the significance of this study is two-fold. Firstly, it provides empirical evidence that the high presence of socially offensive language and sarcasm found in traditional psychology studies can be extended to social media texts too. Secondly, within the high-stress corpus, presence of socially offensive language and sarcasm was highly dependent on the stressor. This presents a potential new research direction to explore these linguistic constructs with respect to the factors inducing the mental states.

The significance of this study lies primarily in the existing knowledge gap about presence of socially offensive language and sarcasm in social media texts with expressions of stress. As described above, the relation between socially offensive language and sarcasm with stress has been studied earlier in psychology research. Both have been established as behavioural coping mechanisms in response to psychological stress. However, most of the research works focus on verbal sarcasm and socially offensive language. Our research addressed this gap and analyses the presence of these linguistic features in texts from social media. The resulting insights are useful for researchers in psychology to use social media text features as potential indicators of underlying psychological stress.

# Chapter 6

# TEMPORAL INTENT OF HIGH-STRESS TWEETS

Temporal identification is a way to comprehend a sequence of events and locate them on a scale of time. Intuitive to human understanding, events occur either before, during or after a reference point and thus can be classified into past, present, and future with respect to this chosen point. Temporal intent in text has been used to help to identify the social status and features of users. A constant emphasis on a specific temporal class can give insights into significant psychological features. For example, a focus on negative past events is a likely indicator of depression-prone behaviour.

While the existing research focuses on aggregating temporal information at the user level, domain-level variations of temporal intent are largely unexplored. Also, psychological stress in itself is an indicator and root cause of a wide variety of physiological and mental disorders. This chapter studies the impact of temporal intent on stress measures in social media text and how it varies in different domains for Tweets.

In this study, we analyse the various existing studies in identifying the temporal intent of tweets. Based on these observations, we put forward a classifier to find the temporal intent of a given text as past, present or future. We employ this classifier to learn the temporal intent of tweets having high scores on a stress scale and analyse how the temporal intent varies in with the stress scores and in different domains.

The rest of the chapter is organized as follows: A review of related literature is in Section 6.1. Section 6.2 describes the methods for implementing the temporal classifier. We discuss the results in Section 6.3 and conclude in Section 6.4 with a summary and discussion of scope for future work.

## 6.1 Related Work

Before any connection between expressions of psychological stress and temporal intent can be explored, the idea of temporal intent itself needs to be examined. There are varying views on the concept. Time can be viewed as an objective, measurable concept. At the same time, in psychological studies, it is often treated as individual and subjective. Temporal orientation of a person, the relative significance he or she assigns to the events of the past, present or future is a crucial component of their perspective of self and the world; hence it is termed “the lens through which a person experiences the world” (Begić & Mercer, 2017). It might magnify or distort experiences and hence is vital in clinical or academic settings in any attempt for understanding a person’s lived life. Here, we can see two related but distinct concepts - temporal orientation and temporal

perspective (Mooney et al., 2017). The former denotes whether a person focuses on the events pertaining to the past, present or the future. Temporal perspective further extends this idea to explore whether and how this orientation influences and shapes their behaviour.

Experiences, demographic factors and circumstances can alter time perception of individuals (Witowska, Zajenkowski & Wittmann, 2020). Furthermore, it has been studied in connection with various psychological conditions such as anxiety and depression (Grondin, 2010). Disruptions in regular schedule, such as those experienced during the Covid-19 pandemic also have been shown to influence time perception and induce a paradoxical interpretation in which time spent in quarantine is perceived to be short and yet the beginning is felt to be long back (Grondin, Mendoza-Duran & Rioux, 2020).

In the context of clinical studies, temporal orientation is often depicted as the measure of memory a person can bring to a study. This measure is particularly significant in investigating disorders or conditions which impair thought processes or induce hallucinatory behaviour. Temporal depth, the extent to which a person can think in a given temporal direction, has been found instrumental in his or her ability to follow healthy life choices (Sirois & Pychyl, 2016). Temporal horizon, a related idea, is defined as the extent to which individuals think about their future (Thorstad & Wolff, 2018).

The Zimbardo Time Perspective Inventory (Zimbardo & Boyd, 1999) includes five temporal factors (past-negative, present-hedonistic, future, past-positive and present-fatalistic) and these time perspectives are studied in correlation with personality traits. It established that the students with a predominant future orientation were ambitious with a goal-seeking behaviour. Temporal horizon is found to be an indicator of healthier life choices (Thorstad & Wolff, 2018). Intuitively the farther a person thinks about the future, the less he/she tends to discount future rewards. Extending this to a societal perspective, U.S. states with longer time horizons were found to be more partially liberal and took fewer risks. Both these studies do not investigate the polarity associated with future orientation. Intuitively, excessive future orientation, associated with negative emotions would be counter-productive and detrimental to mental well-being similar to past-negative perspective. Nevertheless, at its core, these studies still put forward temporality as a significant dimension with respect to which other psychological constructs could be studied, which inspired us to investigate the predominance of different temporal orientation in tweets with expressions of psychological stress.

The primary segregation of temporal references into past, present, and future are often combined with sentiment polarity to give a comprehensive perspective about user opinion regarding a topic or event. This combined knowledge has been used to model and analyse public sentiments on Twitter about or around events like general elections or the Grenfell Tower fire disaster (Wang et al., 2017). The temporal orientation

of social media posts has been studied in conjunction with several socio-economic factors of the users, such as income (Hasanuzzaman et al., 2017), personality traits, Attention Deficit Hyperactivity Disorder (Table 6.1). Our study contributes further in this direction by analysing the relative dominance of temporal classes in tweets which are high-stress and low-stress.

**Table 6.1 Existing research on the identification of the temporal classes of social media text**

| Author | Temporal Classes | Purpose | Method | Features |
|---|---|---|---|---|
| Kamila et al., 2019 | Near-past, near-future, far-past, far-future | Correlation between user's focus on temporal distance with demographical and psychological attributes | Bi-LSTM | Temporal keywords, verb POS tags |
| Hasanuzzaman et al., 2017 | Past, present, future with respect to posting dates | Predictive model of income of users from tweets. Finds a correlation between future temporal orientation and income | Temporal Orientation: CNN feature extractor with SVM one vs rest classifiers for each temporal class. Income: Logistic Regression and Gaussian Processes | Tweet vectors extracted by CNN Income: user-level aggregated temporal orientation |
| Kamila, 2018 | Past, present, future tweets | Measures the correlation between the sentiment polarity combined with temporal orientation and five psycho-demographical factors (age, education, | Temporal orientation: biLSTM sentiment classifiers NLTK toolkit Psycho-demographical factors | |

| | | | | |
|---|---|---|---|---|
| | | intelligence, relationship status, and optimism) | Logistic Regression | |
| Park et al., 2015 | Past, present, future from tweets and Facebook posts | Correlation between temporal orientation, age, gender and Big-five factors | Temporal Orientation: Extremely Randomized Trees Psychological factors: self-assessment of participants based on international Personality Item Pool | Tokens from happierfuntokenizer LIWC word categories Message length POS tags Time expressions using SUTime annotator |
| Schwartz et al., 2015 | Past, present, future from Facebook posts | Correlation between temporal orientation and conscientiousness, age, gender, social factors (openness, number of friends, depression, life satisfaction, IQ) | Temporal Orientation: Extremely Randomized Trees | 1-3 token sequences, Message length, time expressions using SUTime, POS tags from Stanford, LIWC lexica |

We see that these studies have two dimensions – firstly, they explore the psychology of time perception and define a specific aspect (e.g. temporal orientation, temporal horizon, temporal perspective) and place it as the centre of the study. Secondly, this temporal orientation is studied in conjunction with a psycho-demographic aspect (e.g. income, anxiety, risk-taking behaviour). Methodologically, various rule-based systems and machine learning classifiers have been employed to detect the temporal classes. In the light of these dimensions, we proceed to define the theoretical, practical and methodological scopes of our present work.

Fundamental to all the varying concepts discussed so far is the idea of the temporal frames (past, present and future) into which human cognitive processes partition experiences. In the present research, we focus on these temporal frames as expressed in the social media text. The interest of our temporal analysis lies in what

is tweeted and not who tweeted it. Thus, we explore how the relative significance given to different temporal frames vary in different domains in high-stress and low-stress tweets. Hence, we use the term temporal intent to avoid confusion with temporal orientation and its usage in psychological research parlance. Temporal intent, as we explore it in this research, is the placement of events within the time continuum, which determines whether the events' occurrence is before, during or after when the tweet is posted. The author's individual orientation into a specific time frame is not considered, unlike the psychological research works previously discussed.

It might be noteworthy here to distinguish between tense of a tweet and its temporal intent. It would be naïve to equate the tense of the verbs present in a tweet to its temporal intent. Reiterating our earlier discussion, what we explore in this research is whether the event/circumstance/topic of the tweet exists prior to/during/after the time of tweeting. The tense of the verbs may or may not correctly represent this intent. For example, in the tweet, "this evening I fly to Alaska to mourn with our family", the tweet's implied temporal intent is future. However, the tense of the verb words in the tweet is present. Because of such indirect expressions, studies exploring the temporal information of social media text employ classifier models to correctly infer the temporal class rather than relying on the tense of the verbs (Table 6.1).

An additional point to note is the inclusion of a fourth, atemporal, class in our experiments. This comes from the observation of a considerable number of tweets in the dataset without a specific temporal class indication (e.g. "And 2.5 hours delay! No co-pilot for a flight to Senegal!"). Some existing research considers not including an 'atemporal' class to indicate tweets as a limitation (Kamila et al., 2019). In our study, we focus on the time of the occurrence of the stressful event/situation/interaction with respect to the time of tweeting, as reported by the author. Separation into three temporal classes (past, present, future) was found to be insufficient for the dataset, with a considerable number of tweets having insufficient temporal cues. This motivated us to include an atemporal class.

The automatic classification of the temporal classes of texts has been explored in some noted shared tasks too. SemEval-2007 introduced the TempEval task (TempEval-1 (Verhagen et al., 2007)). The task constructed a list of events, known as the aEvent Target List. The participants were required to find the temporal relations between 1) time and events in the same sentence (subtask 1) 2) document creation time and events (subtask 2) 3) events of adjoining sentences (subtask 3). TempEval-3 (UzZaman, 2012) elaborated further on this by introducing multilingual (six languages) datasets and more subtasks.

NTCIR-11 Tempora Information Access (Temporalia) included two subtasks – Temporal Query Intent Classification (TQIC) and Temporal Information Retrieval (TIR) (Joho, Jatowt & Blanco, 2014). Temporal orientation is an important aspect of search queries. TQIC required participants to classify queries as past-

related, recency-related, future-related and atemporal. Also, for each given topic and temporal question, TIR asked participants to retrieve the top 100 documents.

Since our work involves the identification of temporal intent of tweets as past, present, future or atemporal, we identify TQIC as the most closely related task and examine the various systems in detail (Table 6.2).

The key task in TQIC was to allocate a query to the most appropriate temporal intent class. The temporality of search queries was expressed either by explicit tense words (e.g. 'who was martin luther') or indirect implication (e.g. 'Yuri Gagarin cause of death'). The queries could also belong to an 'atemporal' class when they did not have a clear temporal intent and the search results were not expected to be related to time (e.g. 'distance from earth to sun'). Most participating teams built classifiers based on n-grams, POS tags and time strings (e.g. "price hike in bangladesh 2008"), variants of time sensitive word dictionaries (with words mapped to a temporal class e.g. 'soon' -> future') and named entity recognition. TempoWordNet, a program to assign temporal classes to WordNet synsets, was also used by a participant but was found to contribute only negligible improvement (Filannino & Nenadic, 2014), the reason being the intrinsic predominance of atemporal class in TempoWordNet as the verbs are represented in infinitive form in WordNet. Different classifier models such as Logistic Regression, SVM, Naïve Bayes and Decision Trees were used.

The task, while similar to our work in the key objective of assigning a temporal intent class to text, had certain distinct differences. Firstly, the dataset consisted of queries in contrast to the tweets dataset in our research. Explicit time strings were extensively used as a feature, whereas in our dataset such mentions of year or a date were sparse. Further, the named entities were not as relevant to our dataset as it was for the queries' dataset. The queries were relatively shorter, 4.2 words on average, compared to 15.6 words on average for the tweets in our dataset. The methods outlined in section 6.2 were chosen based on this comparison.

**Table 6.2 Participating teams in TQIC subtask in Temporalia-11 event**

| Authors | Method | Features |
|---|---|---|
| Hou, et al., 2014 | Rule-based | N-grams, POS tags, named entities, normalized dates, date distance, time-sensitive word dictionary |
| Yu, Kang & Ren, 2014 | Logistic Regression and SVM | Time gap, verb tense, lemmas, and named entities |
| Shah, Shah & Majumder, 2014 | Ensemble classifier with SVM, Naïve Bayes, and Decision Trees | Bag Of Words, Query length, number of verbs using NLTK POS tagger, the difference between the query issue date and temporal expression in the query |
| Burghartz & Berberich, 2014 | Naïve Bayes | POS tags, DMIZ directory info, publication dates |

| | | |
|---|---|---|
| Filannino & Nenadic, 2014 | SVM with polynomial kernel and Random Forests | POS tags, TempoWordNet, comparison with Wiki page Titles |

The participant teams for the TQIC task used rule-based and classification models for identifying the temporal categories of the queries. The rule-based system made use of a time-sensitive word dictionary to map queries to the most likely temporal class. Models such as SVM, Logistic Regression, Naïve Bayes and Decision trees, which were the state of the art at the time for text classification tasks, were also used by various participants. POS tags, date and time information and temporal dictionaries were all found as useful features in the task of finding the temporal intent of queries.

## 6.2 Methods

The dataset consisting of 12000 tweets (described in Section 4.2.3), in which 8871 tweets were identified as stressful tweets and the rest no-stress tweets, was used for our study of temporal intent. 2000 tweets (500 from each domain – traffic, airlines, politics and personal events), were randomly chosen from the 12000 tweets, and were annotated for four different temporal intents (past, present, future and atemporal). The annotation process is described as follows:

### 6.2.1 Dataset Annotation

**Annotation Question**

Out of the four temporal classes – past, present, future and atemporal- which one describes the temporal orientation of the given tweet most appropriately?

**Explanation**

Events were classified into three categories - past, present, and future and atemporal – based on their time of occurrence with respect to the current instant. The annotators were required to mark the perceived temporal orientation of the tweet as past, present, future or atemporal. Tweets without temporal cues, or that the annotators could not reach a consensus about, were put into the fourth category, 'atemporal'.

**Guidelines**

- Mark the tweet as 'past' if the event/circumstance/topic of the tweet was existing earlier but has stopped, at the time of tweeting.
- Mark the tweet as 'present' if the event/circumstance/topic of the tweet is existing/on-going at the time of tweeting.

- Mark the tweet as ‘future’ if the event/circumstance/topic of the tweet has not yet occurred but is anticipated to occur sometime after the time of tweeting.
- Mark the tweet as ‘atemporal’ if none of these can be determined or if the underlying temporal implication cannot be identified.

**Examples:**

13. @ElectionMapsUK He was arrested for Bribery, released on Bail, #UKLabour (past)
14. The EU are extremely worried that the UK are leaving with a nice clean break managed WTO rules #brexit (present)
15. will #uklabour expel advocates of the extreme form of political correctness (future)
16. And 2.5 hours delay! No co-pilot for a flight to Senegal! (atemporal)

### 6.2.2 Identification of temporal intent

For the task of building a classifier to identify the temporal intent of tweets, we explored a number of machine learning classification algorithms with different features. We chose three machine learning classifiers (Naïve Bayes (NB), Support Vector Machines (SVM) and Adaptive Boosting Algorithms (AdaBoost) which were widely used in sentiment analysis and text processing tasks as candidates.

**Naïve Bayes** - Multinomial Naïve Bayes with Laplace smoothing

**Support Vector Machines** – libsvm-based implementation. Linear kernel

**Adaptive Boosting** - DecisionTreeClassifier as base estimator, number of estimators as 100

**Features** As a pre-processing step, the tweets were tokenized using NLTK TweetTokenizer package. URLs and usernames were removed and the following features are extracted from each Tweet.

1. **Tense indicators** The annotators noted the time-related words in each tweet and later aggregated these keywords for each tense. The list of words potentially indicating each sense was expanded by including the synonyms of these keywords from WordNet. The cosine similarity for each tense keyword and the words in lemmas, except the English language stop-words, was found. Words with similarity above the mean value were added to the terms’ list corresponding to that specific Tense. In case, the same word was tagged by annotators as related to more than one temporal class, the number of times it was tagged to each class was counted and the term was allocated to the temporal class with the greatest count. Examples for the tense indicators were as follows:
    a. Past: ***ago, yesterday, before, previous, last***

*b.* Present: ***now, already, today, currently, presently, tonight***

*c.* Future: ***again, soon, later, next, tomorrow, gonna***

The number of tense indicators in a tweet for each tense was taken as a feature.

2. **POS tags** The tokens were POS-tagged using NLTK. For each tense, features were encoded as the ratio of verbs in that tense to the total number of verbs in the Tweet.
3. **Lexicon** Using the 2015 version of LIWC, the percentage of temporal categories past, present, and future were taken as features for the classifiers.
4. **N-grams** Features were encoded as binary scores based on the presence or absence of an n-gram in a Tweet. We considered uni-, bi-, and trigrams. Based on tf-idf, the top 1000 n-grams were chosen for each type of n-grams.

These classifiers discussed in the previous section were implemented using the scikit-learn toolkit and precision, recall and accuracy in temporal intent identification were evaluated by 5-fold cross-validation.

The performance of the different classifiers was evaluated on the manually annotated dataset of 2000 tweets, described in Chapter 5. The dataset consisted of 500 randomly chosen tweets from each of the four domains, politics, traffic, airlines and personal events.

We implemented an ensemble classifier for the identification of temporal intent, based on the solution provided by Andd7 team (Shah, Shah & Majumder, 2014). It is essentially a combination of Naïve-Bayes, SVM and Adaptive Boosting classifiers. If at least two classifiers agree on the temporal intent of the given text, it is accepted as the final decision. If not, the decision of the SVM classifier is taken as the output of the ensemble classifier, based on its better performance (Table 6.3).

**Table 6.3 Performance of SVM, Naïve Byes (NB) and Adaptive Boosting (AdaBoost) classifiers on temporal-orientation classification task with various feature combinations. FC1: n-grams, lexicon, pos tags, and tense**

**indicators. FC2: n-grams, pos tags and tense indicators FC3: pos tags, and n-grams. Performance measured in terms of mean accuracy in 5-fold cross-validation (precision, recall, and F-measure in brackets)**

| Classifier | Features | | |
|---|---|---|---|
| | **FC1** | **FC2** | **FC3** |
| SVM | 0.76(0.71,0.64, 0.67) | 0.72(0.68, 0.67, 0.675) | 0.69(0.64, 0.65, 0.65) |
| Ensemble | **0.79(0.71,0.79,0.74)** | 0.78(0.74,0.72,0.77) | 0.71(0.65,0.76,0.69) |
| NB | 0.77(0.69,0.78,0.71) | 0.75(0.73,0.72,0.67) | 0.72(0.64,0.73,0.67) |
| AdaBoost | 0.76(0.74,0.76,0.75) | 0.76(0.73,0.71,0.74) | 0.71(0.65,0.70,0.67) |

Based on the better performance, the ensemble classifier was chosen to classify the test corpus of 10000 tweets into the four temporal classes.

## 6.3 Results

The distribution of the temporal classes over the combined dataset of 12000 tweets is examined (Table 6.4, Table 6.5, Table 6.6, Table 6.7).

**Table 6.4 Percentage of Tweets in the four temporal classes – Political domain (3000 tweets)**

| Stress score | Temporal Intent | | | |
|---|---|---|---|---|
| | **Past** | **Present** | **Future** | **Atemporal** |
| -1 | 18.2 | 31.2 | 27.1 | 23.5 |
| -2 | 12.3 | 37.3 | 24.0 | 26.4 |
| -3 | 13.5 | 46.1 | 15.0 | 25.4 |
| -4 | 15.4 | 51.9 | 14.2 | 18.5 |
| -5 | 16.3 | 52.4 | 13.3 | 18.0 |

**Table 6.5 Percentage of tweets in the four temporal classes – Airlines domain (3000 tweets)**

| Stress score | Temporal Intent | | | |
|---|---|---|---|---|
| | **Past** | **Present** | **Future** | **Atemporal** |
| -1 | 27.5 | 20.6 | 28.6 | 23.3 |
| -2 | 35.3 | 20.3 | 24.0 | 20.4 |
| -3 | 47.8 | 20.2 | 20.9 | 11.1 |
| -4 | 45.5 | 32.5 | 11.4 | 10.6 |
| -5 | 47.3 | 31.7 | 11.2 | 9.8 |

**Table 6.6 Percentage of tweets in the four temporal classes – traffic domain (3000 tweets)**

| Stress score | Temporal Intent | | | |
|---|---|---|---|---|
| | **Past** | **Present** | **Future** | **Atemporal** |
| -1 | 20.1 | 37.2 | 18.1 | 24.6 |
| -2 | 19.8 | 47.4 | 15.0 | 17.8 |
| -3 | 14.3 | 53.8 | 14.5 | 17.4 |
| -4 | 20.2 | 56.8 | 8.4 | 14.6 |
| -5 | 16.8 | 60.1 | 8.2 | 14.9 |

**Table 6.7 Percentage of tweets in the four temporal classes – personal events (3000 tweets)**

| Stress score | Temporal Intent | | | |
|---|---|---|---|---|
| | **Past** | **Present** | **Future** | **Atemporal** |
| -1 | 26.5 | 28.9 | 26.0 | 18.6 |
| -2 | 32.2 | 34.1 | 22.0 | 11.7 |
| -3 | 32.1 | 47.4 | 10.1 | 10.4 |
| -4 | 37.2 | 38.1 | 16.2 | 8.5 |
| -5 | 48.1 | 24.2 | 18.2 | 9.5 |

Tweets with atemporal intents (those without clear temporal indicators or about which the annotators had unresolved conflicts) had significant presence in no-stress corpus among all domains (airlines -23.2%, politics-23.5%, personal events – 18.6% and traffic- 24.6%). Compared to this, the percentage of atemporal tweets is considerably less in the corpus of tweets with highest stress score (-5) (airlines-9.8%, politics-18.0%, personal events-9.5%, traffic-14.9%).

It can also be observed that the percentage of 'future' temporal class decreases in higher stress score classes of tweets in all domains. The highest presence of 'future' temporal class is in the no-stress corpus in all the four domains.

The corpus with stress scores from -2 to -5 is found to follow domain-specific patterns for the temporal classes. The relative presence of temporal classes does not change significantly with stress scores.

In politics, 'present' is the most dominant temporal intent class. Intuitively, the discussion in political tweets very frequently centers on the current issues.

(44) The most frightening aspect of this is the social credit system. It seems like a super extreme version of FICO

In airlines, tweets stress scores from -2 to -5 have past as the dominant temporal class. It might be because commuters tweet about stressful events which happen during or related to air travel, after its occurrence. An exception would be airlines tweets with delay as the stressor which has the following distribution of temporal classes (34.2% past, 40.1% present, 17.7% future and 8.0% atemporal).

(45) @Delta seems to be more worried about their bottom line than the value of their customers #11hourdelay (present)

For traffic tweets, the majority of tweets across the different stress-score classes belonged to the temporal class of 'present', because people tweet in real-time about ongoing stressful events like congestion and accidents.

(46) police car holds up traffic on busy London street to tell cyclist who is legally allowed to be there what even is that holy s**t (present)

In personal domains, there is a more even distribution between the temporal classes of past and present. In tweets with the highest stress score (-5), past is the dominant temporal class (48.1%) and present is second (24.2%).

(47) Three weeks since my father passed away. I fully haven't adjusted to this #devastated #rip (past)

In summary, we observe the distribution of temporal classes depends on the domain rather than the stress score for high-stress tweets. The temporal classes of tweets follow roughly the same distribution across different stress score categories of the same domain but varies in different domains.

The presence of the four temporal categories were also examined in tweets with different stressors. The results are given in tables 6.8, 6.9, 6.10 and 6.11.

**Table 6.8 Percentage of tweets with different stressors in the four temporal classes– politics (1927 tweets)**

| | Stressor | Past | Present | Future | Atemporal |
|---|---|---|---|---|---|
| Politics | Election | 22.7 | 45.3 | 20.4 | 11.6 |
| | Protest | 28.4 | 46.8 | 15.9 | 8.9 |
| | Violence | 36.5 | 42.4 | 14.0 | 7.1 |
| | Media | 29.0 | 38.5 | 22.2 | 10.3 |
| | Economy | 29.3 | 39.4 | 21.8 | 9.5 |

**Table 6.9 Percentage of tweets with different stressors in the four temporal classes– traffic (2613 tweets)**

| | Stressor | Past | Present | Future | Atemporal |
|---|---|---|---|---|---|
| Airlines | Delay | 34.2 | 40.1 | 17.7 | 8.0 |
| | Luggage | 42.3 | 34.8 | 15.6 | 7.3 |
| | Service | 41.4 | 25.4 | 24.3 | 8.9 |
| | Cost | 39.6 | 32.4 | 18.9 | 9.1 |

**Table 6.10 Percentage of tweets with different stressors in the four temporal classes– airlines (2493 tweets)**

| | Stressor | Past | Present | Future | Atemporal |
|---|---|---|---|---|---|
| Traffic | Accident | 31.5 | 40.6 | 17.5 | 10.4 |
| | People | 26.1 | 49.2 | 16.8 | 7.9 |
| | Delay | 20.3 | 50.9 | 12.8 | 16.0 |
| | Congestion | 25.4 | 51.3 | 8.7 | 14.6 |
| | Violence | 32.8 | 42.8 | 16.1 | 8.3 |
| | Climate | 31.3 | 45.6 | 16.6 | 6.5 |

**Table 6.11 Percentage of tweets with different stressors in the four temporal classes– personal events (1838 tweets)**

| | Stressor | Past | Present | Future | Atemporal |
|---|---|---|---|---|---|
| Personal Events | Death | 41.7 | 32.6 | 16.0 | 9.7 |
| | Relationship | 39.4 | 36.5 | 13.7 | 10.4 |
| | Marriage | 36.8 | 40.3 | 13.4 | 9.5 |
| | Children | 38.2 | 41.9 | 11.6 | 8.3 |
| | School | 32.3 | 37.1 | 18.8 | 11.8 |

## 6.4 Discussion

### 6.4.1 Error Analysis

The confusion matrix for the ensemble classifier is calculated (Table 6.12).

**Table 6.12 Confusion matrix for the ensemble classifier with the feature combination FC1 in the dataset of 2000 tweets from the four domains**

| **Predicted** | **Actual** | | | |
|---|---|---|---|---|
| | **Past** | **Present** | **Future** | **Atemporal** |
| Past | 426 | 8 | 44 | 15 |
| Present | 89 | 507 | 105 | 23 |
| Future | 10 | 102 | 270 | 27 |
| Atemporal | 59 | 4 | 13 | 298 |

The confusion matrix points to a large number of mis-classifications between the present and future temporal classes. There are relatively fewer misclassifications in the atemporal or past classes. Three types of errors were identified.

**Mixed temporal indicators**

(48) Well, that's what they said earlier. It's been 2hrs from the original departure time. We're still here

This was annotated as present; but the ensemble model misclassified it as past which could be because of the past verb forms and tense indicator ('earlier') present in the tweet.

**Ambiguous tense indicators**

(49) Curse you #AmericanAirlines — I am going miss this because no one seems to know when we are leaving Charlotte tonight — UGGHH!!

The annotators marked this as future. But 'tonight' is considered as an indicator of present, according to the list of tense indicators. The tweet was misclassified as present.

**Implied temporal information**

(50) You can always count on @AmericanAir to leave you on the tarmac for an hour, make you miss your connection and then apologize. I can say it was a wonderful experience!!!

This tweet is talking about a past experience, and it is annotated as past by human coders. But there is no direct indicator for the temporal intent of the incident in which the user missed his flight, except for the past tense in the second sentence. The past orientation for the tweet is implicit and it was challenging for the classifier to correctly classify such examples.

### 6.4.2 Limitations

The classification considers the temporal cues in the text. However, the temporal intent of the stressful events could be more subtly represented. A major limitation of the temporal classifier is its dependency on explicit linguistic cues present in the text. This presents two challenges: Firstly, the list of tense indicators was generated from the seed list of words identified as temporal indicators by the human annotators. This seed list was generated from the annotated dataset of 2000 tweets. Using a larger dataset for tense indicators identification could have given more coverage to terms which indicate each temporal class. Secondly, tweets which contain different tenses in the same or adjacent sentences were challenging for human annotators. There are instances of such tweets in the test dataset which were misclassified. The tense indicators list only contains unigrams now; this could be expanded to include bigrams, such as 'this evening' or 'that day' to reflect its indication of specific temporal classes. Instead of assigning a specific temporal category to a tense indicator based on the relative frequency of usage, a probability factor could better represent its correlation with temporal classes (for terms such as 'tonight' which occur in tweets with future or past orientation based on the context). The classification model could also consider the timestamp of the tweet (from the 'created_at' field of the Tweepy search result) for a possibly better classification.

The temporal classes themselves could be more refined to reflect our learning from the study. From the analysis of the error examples and discussion with the annotators, it seems more intuitive to rename the 'present' temporal class as 'recent'. Events that occurred within a specified time window of recency (e.g. within 24 hours) with respect to the time of reference could be assigned to this 'recent' class. In the current model, tweets that talk about the stress about events that occurred before a few minutes are not distinguished

from those which occurred a few days ago. Similarly, the events which are anticipated later the same day and a few months later are both put together in the 'future' class. This could be changed by refining the definition of the temporal classes.

Improving the existing features and adding additional ones, the classification models could likely improve the accuracy of temporal class identification. The distribution of four different classes in the domains could vary according to these modified methods.

It is also interesting to examine how the temporal patterns of Twitter usage could have influenced the findings. Intuitively, tweets tend to focus on very near future or very near past events. However, there is no research evidence to support this because of the lack of studies specifically investigating the temporal intent of tweet content. Hence, we cannot analyse the extent to which the inherent temporal preferences of Twitter usage have influenced our findings in these domains. Some of the methods developed in this thesis may therefore work differently on other types of social media text with different common types of temporality.

## 6.5 Conclusions

Detecting psychological stress from social media can help to systematically identify stressors in real-time, informing systems that may take remedial action. Little is known about the ways in which stress is expressed around different topics, however. This study explores temporal features of 12000 high-stress Tweets from different domains (Politics, Airlines, Traffic, Personal events). We compare the performance of different classifiers and features for identifying the temporal intent of Tweets as past, present, or future. The best performing classifier achieves an accuracy of 79.1% and we further analyse the relative presence of the four temporal classes in Tweets with different stress scores, using this classifier. The results show that the percentages of past, present, and future temporal intent in high-stress tweets vary between domains but are largely the same at different levels of stress.

The percentage of tweets in both future and atemporal classes was significantly lower in the higher stress tweets compared to the no-stress tweets. The reduced presence of atemporal classes in very high-stress tweets might be because they contain more specific temporal indicators in them. In the corpus of tweets with stress expressions, the distribution of temporal classes varied according to the domain. In airlines, the past was the dominant temporal class, which could be due to the tweets being posted after the air travel. In high-stress tweets on politics and traffic, the dominant temporal class was present and in personal events, both past and present had an equally significant presence.

These findings add to the understanding of how psychological stress is expressed in different domains in Twitter. The existing research focuses on the correlation of temporal orientation with a person's ability to cope with psychologically stressful incidents. The specific temporal intent of tweets has not been studied in

high-stress corpus before; hence we address the research gap about the temporal intent of the stressful incidents or circumstances as reported on social media. We provide empirical evidence that high-stress and no-stress corpora follow the distinct distribution of temporal intent classes. This points to temporal intent as a useful dimension to explore while characterizing expressions of psychological stress in social media. Earlier research in the temporal orientation of tweets from users with mental disorders indicated a clear preference for specific temporal classes depending on the mental state. We also found that within the high-stress corpus the distribution varies too depending on the domain of discussion. This inspires including domain information in the analysis of temporal class patterns of the corpus with expressions of specific mental states.

# Chapter 7

# CONCLUSIONS

This study investigated the reasons for stress and attributes of high-stress tweets. Tweets were analysed to detect stress, to find factors associated with stress on Twitter, and to explore the nature of expressions of stress in tweets. Although previous studies have analysed the linguistic characteristics of tweets from users suffering from long-term stress-related disorders, this thesis focused instead on short-term stressful events related to traffic, airlines, politics, and personal events. The presence of socially offensive language and sarcasm was studied in tweets, in connection with psychological stress, for the first time. Previous studies of stress in these domains have been so far limited to traditional methods, such as surveys and interviews.

In terms of data and methods, this study created a corpus of high-stress tweets and applied several NLP techniques to understand the stressors and the linguistic features through which stress is expressed.

Section 7.1 revisits the research questions together with a review of how this study addressed each of them. The contributions from this study are presented in Section 7.2. The limitations and possible directions for future research are given in Section 7.3 and Section 7.4 respectively.

## 7.1 Research Questions

1. *Can word sense disambiguation improve the accuracy of the lexicon-based stress strength measuring scheme, TensiStrength?*

To see if word sense disambiguation changes the accuracy of stress and relaxation scores assigned by the lexicon TensiStrength, we implemented a method unifying word senses and word vector representations (Chapter 3), as originally proposed for sentiment analysis (Chen, Liu & Sun, 2014).

We manually identified 40 ambiguous words in the lexicon of TensiStrength. The lexicon was modified to include separate stress/relaxation scores for different senses of the same word. For each of these words, the different senses were found from the resource WordNet. Using an existing method (Chen, Liu & Sun, 2014), we found the sense vector for each of these senses. For each tweet containing an ambiguous word, the sense with the vector having the highest cosine similarity to the vector representing the tweet was chosen as the correct sense of the word in that context. Then, the stress score corresponding to this sense was chosen from the modified TensiStrength lexicon.

A dataset of 1000 tweets with ambiguous words was collected from Twitter. A gold standard for stress and relaxation scores was created by averaging the scores given by three independent human coders. The sense disambiguation method was implemented with vectors trained by the Twitter Word2Vec model released as

part of the ACL W-NUT task (Godin et al., 2015). Compared to the original TensiStrength and machine learning classifiers (SVM, Naïve Bayes, Decision Trees with Adaptive Boosting, Logistic Regression) with n-gram features, the new method gave better performance (Table 3.6, Table 3.7).

*2. Can stressors be identified from social media posts?*

Existing research on stress expressions in social media posts has focused on long-term stressful events and their impact on the mental health of people. For these, tweets were collected and analysed to see how the linguistic features vary before and after the onset of a psychological disorder. Whilst other stress-inducing domains, such as travel and politics, have been investigated offline by psychologists, they have been ignored so far in social media stress-related research.

This thesis reports a new framework to find the stressors from a given tweet. There were two challenges in this – creating a list of the common stressors in a given domain and finding the stressor in each tweet, from this list.

Common stressors have been identified from the existing literature on each domain (Table 7.1).

**Table 7.1 Stressors for different domains from existing literature**

| Domain | Stressors | References |
|---|---|---|
| Air-travel | Fellow passengers<br>Lack of trust in airlines/airports<br>Events during air-travel<br>Take-off/landing | Bricker, 2005<br>McIntosh et al., 2006 |
| Travel | Work-related travel<br>(travel arrangements, hotel preferences, unhealthy lifestyle, concerns about the destination, work/personal issues)<br>Recreation travel | Defrank, Konopaske & Ivancevich, 2000<br>Chen, 2017 |
| Personal events | Work (getting tired, retirement, change to a different line of work)<br>living conditions (change in residence, change in living conditions, revision of personal habits)<br>Recreation (change in recreation, vacation)<br>Marital life (marital separation, reconciliation, divorce, marriage)<br>Death (of a spouse, of a close family member) | Holmes & Rahe, 1967 |
| Politics | Candidate losing in election, partisan politics, referendums | Hughes, 2019<br>Roche & Jacobson, 2019 |

However, these studies rely on evidence from questionnaires, which might not match with the reasons expressed on the social web, and so new methods were needed to extract stressors from social media posts. For this, we employed a combination of LDA topic modelling and k-means clustering to construct a potential stressors list for each of the four domains included in the study – traffic, airlines, personal events, and politics.

We compared the performance of LDA and LSA methods of topic modelling using UMass and UCI metrics which measure the coherence of word pairs in the topic modelling results. Based on this, LDA was used to find the topics for each of the high-stress datasets. A k-means based clustering using the Word2Vec vector representation was then implemented to form clusters of these topics. The number of clusters in each domain was obtained from silhouette analysis. Each cluster then represented a stressor. The stressors are summarized for the four domains (Table 7.2).

**Table 7.2 Stressors for different domains, extracted from a corpus of 5539 tweets with high stress scores (-3, -4, or -5)**

| Domain | Reasons |
|---|---|
| Politics | Election, protest, violence, media, economy |
| Personal events | Death, relationships, marriage, children, school |
| Traffic | Climate, violence, accident, delay, congestion, people, campaign |
| Airlines | Luggage, service, delay, cost |

The topic-modelling-based analysis of tweets found different stressors compared to those discussed in traditional psychological surveys and studies ( Table 7.1, Table 7.2).

Overall, the questionnaires focused on personal responses to stressful events whereas a domain-based approach brought light to the events themselves. For example, a favourite candidate losing an election is stated as a stressor in a survey (Smith, Hibbing & Hibbing, 2019) whereas, in our study, elections are a stressor in political tweets. Surveys reported mundane activities related to air-travel (take-off and landing) as stressors. In our study, the reasons found were about the performance of airlines (luggage, service, delay, or cost) and the associated issues, rather than personal perceptions or experiences. The reasons in tweets were the ones that are far more likely to impact a wider audience. This could arise from the fundamental difference in the data sources these studies looked at – surveys look for personal responses whereas, in tweets, people proactively share their experiences which they perceive to have a probability of resonating with the experience of the social media community.

After finding the potential reasons, the next task was to map a given tweet to one of these stressors. Three-word vector-based methods were introduced for this.

1. Maximum word similarity method
2. Context vector similarity method
3. Cluster vector method

The cluster vector method performed better than the other two methods and three machine learning baseline methods (SVM, decision trees with adaptive boost, and Logistic Regression). Tweets with indirect expressions were found to be difficult to classify into the correct stressor category, however. There were also problematic tweets with multiple stressors. The methods could be modified to accommodate these.

3. *How prevalent are swearing/socially offensive language and sarcastic language in social media posts with expressions of stress?*

Chapter 5 investigates this question. Based on the insights from a literature review, we implemented the following to identify swearing/socially offensive language and sarcasm from our high-stress corpus.
Swearing: a CNN was implemented using Keras, based on the architecture proposed in Kim (2014). The CNN had an embedding layer, a convolutional layer, one maxpool layer, a fully connected layer, and one softmax layer with dropout. Initially, the performance of this CNN was compared to a Logistic Regression and SVM classifiers. The CNN was found to perform better in terms of F-measure on the offensEval-2018 dataset and the high-stress dataset consisting of 2000 tweets.

Sarcasm: A MultiLayer Perceptron model based on NIHRIO solution (Vu et al., 2018) for SemEval-2018 task 3. The MLP consisted of an input layer, two hidden layers with ReLU activation, and a softmax output layer. Similar to the previous experiment, this MLP implementation was compared to SVM and Logistic Regression on SemEval-2018 task 3 dataset and the high-stress dataset consisting of 2000 tweets.

The CNN and MLP were used to identify swearing and sarcasm in a high-stress dataset of 8871 tweets.

Overall, high-stress tweets had a greater percentage of swearing and sarcasm in all four domains compared to no-stress tweets (with a stress score of -1) (Table 5.10).

The key findings from this comparative analysis were:

- High-stress corpus had a greater percentage of swearing/socially offensive tweets compared to the no-stress corpus in all domains

- The high-stress corpus had a greater percentage of sarcastic tweets compared to the no-stress corpus in all domains
- In all domains in the high-stress and low-stress corpus, the percentage of socially offensive tweets is higher compared to sarcastic tweets

There were specific stressors in each domain which co-occurred with socially offensive language and sarcasm in comparison to other stressors (Table 7.3).

**Table 7.3 stressors with the highest percentages of offensive and sarcastic tweets in different domains (number of tweets = 8871).**

| **Domain** | **Stressor with the highest percentage of offensive tweets (percentage in brackets)** | **Stressor with the highest percentage of sarcastic tweets (percentage in brackets)** |
|---|---|---|
| Airlines | Cancel(23.2) delay(22.9) | Cancel(15.2) delay(14.4) |
| Personal events | Marriage(25.9) | Marriage(18.4) |
| Politics | Economy(28.7) | Economy(17.1) |
| Traffic | People(22.6) | Congestion(14.6) |

In all domains, except traffic, the same stressors hold the highest percentage of both offensive and sarcastic tweets.

4. *What is the distribution of the four categories of temporal intent (past, present, future, and atemporal) in the social media posts with expressions of stress?*

The temporal intent in high-stress tweets was studied in Chapter 6. Section 6.1 summarizes the relevant previous work about the identification of temporal intent and its correlation with various socio-demographic factors.

We considered three machine learning classifiers (SVM, Naïve Bayes, and AdaBoost) for the task of temporal intent identification. An ensemble classifier was constructed based on voting from these three classifiers. On a dataset of 2000 tweets, this ensemble classifier performed with an accuracy of 0.76.

The percentage distribution of high-stress tweets across temporal intent categories was found to vary between domains. A higher percentage of tweets was found in the present category in politics and traffic. In airline-related tweets, the highest percentage of tweets was in the past category. In tweets belonging to personal events, present was the dominant temporal category but tweets with past as the temporal intent also occupied a significant percentage.

## 7.2 Contributions

The pursuit of the thesis research goals resulted in a set of novel contributions to research.

### 7.2.1 An Improved Lexical Stress-Detection Method for Social Web Posts

The upgraded version of TensiStrength produced in this thesis is more accurate than the original one. In the absence of competitors, it offers state-of-the-art performance for the task of stress and relaxation strength detection. The improved lexicon can possibly improve the performance of hybrid models used for user-level stress analysis similar to the one used in an analysis of the language of stress in the Facebook users in the USA (Guntuku et al., 2018).

### 7.2.2 Two-Stage Stressor Detection Method

The second contribution is the framework to identify stressors for a given domain. This framework first finds the possible stressors given a corpus of high-stress tweets, using LDA topic modelling and k-means clustering. the second stage maps tweets to one of the reasons found from the first step. This is the first algorithm to extract stressors from tweets belonging to multiple domains. In the present study, the identification of stressors is demonstrated for four domains – airlines, personal events, politics, and traffic. This can be easily adapted for other domains, though the exact performance in new domains is subject to verification.

### 7.2.3 Greater Prevalence of Swearing, Sarcasm in High-Stress Tweets than low-stress tweets

Although the presence of swearing, sarcasm, and the dominant temporal category have been studied in relation to psychological stress and mental disorders before, this thesis reports their prevalence in high-stress social media posts for the first time. Whilst the results cover only a few domains, the methods can be extended to others. The presence of swearing and sarcasm was found to be greater in the high-stress corpus compared to the no-stress corpus in all the four domains studied.

### 7.2.4 Temporal Intents

The relative dominance of temporal categories was found to vary in high-stress corpus depending on the domain. Though temporal intent has been studied in relation to some other psycho-demographic factors, this is the first time the relative presence of different temporal intent categories in high-stress and no-stress tweets are studied.

### **7.2.5 Overall Contributions**

Collectively, the results of these experiments contribute towards a better understanding of how psychological stress is expressed in social media. This understanding is relevant for the current research directions in

psychology and, in a broader sense, social sciences. The novelty of our research lies primarily in the following directions we pursued.

Short-term vs long-term stressors: Within the sparse literature exploring psychological stress in social media, we observe the almost exclusive focus on long-term stressors such as divorce, death, illness, etc. Our research while considering such stressors in personal lives (and politics, where the stressors are predominantly long-term) also studies short-term stressors from domains such as traffic and air travel. We detected comparable high-stress levels in datasets from all domains. Distinct stressors could be detected from all four domains. Though these stressors were by nature long-term in personal events and politics, and short-term in traffic and airlines, all domains had a higher presence of socially offensive language and sarcasm in the high-stress corpus with respect to the low-stress corpus. Such common findings point to the necessity of extending psychological research into stress eliciting from short-term events.

Domain-based vs individual: Another novel aspect was the domain-based understanding of psychological stress. So far, psycholinguistic analysis of tweets focused on individual mental states. In our study, we construct a collective picture of how psychological stress is expressed in tweets belonging to certain domains. We identify domain-specific patterns in the analysed tweets. This is specifically relevant for the social sciences perspective – e.g. how often swearing is used in the high-stress political tweets about elections.

It is also worthwhile to examine the contributions of this study compared to traditional surveys. Identifying expressions of psychological stress is the first step in understanding and alleviating it. One can observe specific benefits of analysing social media responses. Together with the accessibility and affordances of Twitter, online conversations consist of spontaneous inputs and opinions compared to the guided prompts of surveys. Furthermore, surveys seek explicit responses which the participants, especially when it is about the psychological states, might have reservations in sharing. As for tweets, studies extract implicit textual clues. Such a screening of text overcomes a significant limitation of surveys – the dependence on the explicit projection of self-theories of people. Finally, and especially with the new Twitter Academic API (January 2021), it is possible to access much larger scale stress-related information from Twitter than from surveys, making it possible to investigate a wider range of stress-related topics (e.g., stress due to a Covid-19 diagnosis).

## 7.3 Limitations

### 7.3.1 Data Sources

One of the challenges was the lack of annotated social media data for stress scores and other stress features. As described in Chapter 3, we collected tweets based on hashtag searches. Thus, the annotated data is limited to a single source, Twitter, and a relatively simplistic data collection method. We relied on hashtags

to collect tweets belonging to different domains. The hashtags were sufficient to create a corpus consisting of high-stress and low-stress tweets about the concerned topics, however, the number of hashtags is relatively small.

The data is also limited to English tweets and the results could be different in other languages. Another limitation is the relatively small size of the annotated corpus. It would be interesting to develop a larger corpus and see if the models developed in chapters 5 and 6 perform better on it.

The stress scores and reasons were subjective in nature and hence the annotation process deserved special attention. We decided against using crowdsourcing for annotations and instead chose three human coders with near-native English language skills. Clear annotation instructions and examples were provided and further questions from the annotations were clarified on an individual basis. While we believe this resulted in the high quality of annotations (illustrated by Krippendorff's alpha measures), a crowd-sourcing approach could have been used to create a larger annotated dataset.

Our method for word sense disambiguation for TensiStrength performance was tested on a dataset of 1000 tweets selected based on the occurrence of ambiguous affect words. The generalizability of the method's performance has yet to be verified using larger datasets. Similarly, the machine learning models used for comparison used a rather simple feature selection based exclusively on n-grams. Furthermore, TensiStrength with WSD was slower in processing compared to the original TensiStrength software. This increased processing speed must be carefully weighed with the potential improvement in performance to choose the best-suiting method for a given dataset.

Due to the broad nature of the hashtags used for dataset collection, the stressors extracted by our method were general in scope. The applicability of our methods in more focused domains (e.g. BREXIT tweets) has to be verified.

Although our models for sarcasm/swearing/temporal intent detection performed comparably to the methods in existing research, as found in the error analysis, there were cases like rare/code-mixed swear words, misleading sarcastic expressions, and ambiguous temporal indicators which were challenging to handle correctly. A larger dataset for training might reduce these errors.

### 7.3.2 Domains and Methods

The choice of domains was based on the insights from some studies in traditional psychology (as discussed in Chapter 1) and the domain analysis of the dataset collected for the WSD experiment. While politics, personal events, and travel are domains with stressful events, this thesis does not cover other similar domains. For example, intuitively, work-related tweets could contain high-stress expressions. A potential alternative will be

to topically model a large, high-stress corpus and thus identify specific domains which frequently induce psychological stress. Reddit forums like r/stress could be considered as an auto-labelled high-stress corpus for this purpose

The stressor identification method for different domains was the first of its kind; however, it finds tweets with indirect expressions and/or multiple stressors, challenging. The method does not consider the case in which a tweet does not belong to any of the given stressor categories.

The timeline of the research influenced the choice of some methods for the analysis. For example, in 2019, several NLP research studies employed BERT with excellent results. Since our experiments were completed by that time, we didn't use BERT for identifying offensive and sarcastic language which could have possibly resulted in better classifications.

Over the last few years, instead of disambiguating specific ambiguous words in text, constructing a contextual representation of text has been established as an alternate approach. Embeddings with contextual information created by models such as BERT and ELMO have surpassed state-of-the-art methods in NLP tasks. In the light of these recent breakthroughs, context information from such embeddings could be an alternate solution to address the polysemy of stress/relaxation terms of the TensiStrength lexicon. As noted in chapters 5 and 6, such contextual representations of tweets could also improve the classification and thus our understanding of socially offensive language, sarcasm, and temporal classes in the high-stress corpus.

## 7.4 Directions for Future Research

A few possible directions for future research can be thought of, based on the limitations and findings of the current study. During the extraction of stressors from tweets, we observed that incorporating k-means clustering significantly improved the coherence of topic clusters retrieved from LDA topic modelling. However, this is only an initial result and we intend to explore this possibility further with other datasets and clustering methods as a future research direction.

For stressor identification, in all the domains we used a Word2Vec model trained on 400 million tweets (Godin et al., 2015). It would be interesting to use Word2Vec models trained on domain-specific datasets and to note if it influences the performance of stressor identification methods significantly.

Other sources like Facebook and Reddit have been shown to be useful in the analysis of psychological stress. Posts in subreddits like r/stress could be particularly interesting because they have conversations dedicated to psychological stress and have no limitation on length. A possible future study would be to see if the methods for finding stressors are equally effective and the prevalence of swearing and sarcastic language is similarly distributed in social media data from these sources, compared to Twitter.

Our methods could also be extended to such datasets as a future research direction. The methods to find the stressor were evaluated on relatively broad topic areas like UK politics and London traffic. They may perform differently on more fine-grained domains (e.g., UK political tweets about BREXIT, traffic tweets collected during a major sports event). Also, instead of predicting a single stressor, the methods can be extended to predict single, multiple, or no stressors.

Since BERT has been demonstrated to give superior performance in several NLP tasks (e.g. Table 5.2), it would be interesting to apply it to the identification of swearing, sarcasm, and temporal categories for possibly more accurate classifications. It might reduce the errors found from the analysis of the existing classification models.

Similar to swearing and sarcasm, the co-occurrence of emotion words (negative emotions like fear, anger, sadness) in high-stress tweets could also be studied in future research. Stress is shown to be closely related to the negative emotions of fear and anxiety, but not directly linked to disgust (Fredrikson & Furmark, 2006). It would be interesting to examine how these negative emotions co-occur with stress in the corpora we developed.

The methods to identify stressors would be potentially helpful in finding and remedying stress inducers in multiple domains. As an example, for service-oriented industries like airlines or hotels, it would be beneficial to know which stressors are jeopardizing their customers' travel/stay experiences. It would be useful for city planners to understand and take measures to reduce stressors from commuter tweets. The insights into the expressions of stress may also be helpful for traditional psychology studies and social media analysis.

# References


Alfina, I., Sigmawaty, D., Nurhidayati, F., & Hidayanto, A. (2017). Utilizing Hashtags for Sentiment Analysis of Tweets in The Political Domain. Proceedings of ICMLC.

Allport, G. W. (1961). Pattern and growth in personality. Holt, Reinhart & Winston.

Alvarez-Melis, D., & Saveski, M. (2016). Topic Modeling in Twitter: Aggregating Tweets by Conversations. Proceedings of the International Conference on Web and Social Media.

Ao, S. (2018). Sentiment Analysis Based on Financial Tweets and Market Information. International Conference on Audio, Language and Image Processing (ICALIP), (pp. 321-326).

Assefa, B. (2014). KUNLPLab:Sentiment Analysis on Twitter Data. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), (pp. 391-394).

Azzouza, N., Aklii Astouati, K., & Ibrahim, R. (2020). TwitterBERT: Framework for Twitter Sentiment Analysis Based on Pre-trained Language Model Representations. In Emerging Trends in Intelligent Computing and Informatics.

Baali, M., & Ghneim, N. (2019). Emotion analysis of Arabic tweets using deep learning approach. J Big Data **6,** 89 https://doi.org/10.1186/s40537-019-0252

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010. Valletta, Malta.

Badeaa, M., Gomaa, W., & Haggag, M. (2017). Twitter Messages Sentiment Analysis Model based on Deep and Machine Learning. European Journal of Scientific Research, 93-101.

Badjatiya, P. &. (2017). Deep Learning for Hate Speech Detection in Tweets Pages 759–760. Proceedings of the 26th International Conference on World Wide Web Companion, (pp. 759-760).

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. arXiv.

Balahur, A., & Jacquet, G. (2015). Sentiment analysis meets social media – Challenges and solutions of the field in view of the current information sharing context. Journal of Information Processing and Management.

Bansal, B., & Srivastava, S. (2018). On predicting elections with hybrid topic based sentiment analysis of tweets. Proceedia Computer Science, 346-353.

Bayrak, F. & Alper, S. (2021). A Tale of Two Hashtags: Differences in Moral Content of Pro- and Anti-Government Tweets in Turkey. European Journal of Social Psychology. 10.1002/ejsp.2763.

Baziotis, C., Nikolaos, A., Papalampidi, P., Kolovou, A., Paraskevopoulos, G., Ellinas, N., & Potamianos, A. (2018). NTUA-SLP at SemEval-2018 Task 3: Tracking Ironic Tweets using Ensembles of Word and Character Level Attentive RNNs. Proceedings of The 12th International Workshop on Semantic Evaluation SemEval-2018, (pp. 613-621).

Begić, I., & Mercer, S., (2017) Looking back, looking forward, living in the moment: understanding the individual temporal perspectives of secondary school EFL learners, Innovation in Language Learning and Teaching, 11:3, 267-281, DOI: 10.1080/17501229.2017.1317261

Beigi, G., Hu, X., Maciejewski, R., & Liu, H. (2015). An overview of sentiment analysis in social media and its applications in disaster relief. In W. P. Shyi-Ming Chen, Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence.

Beiser, M., & Wickrama, K. (2004). Trauma, time and mental health: A study of temporal reintegration and Depressive Disorder among Southeast Asian refugees. Psychological medicine, 899-910. doi:10.1017/S0033291703001703

Bellot, P., Hamdan, H., & Béchet, F. (2013). Experiments with DBpedia, WordNet and SentiWordNet as resources for sentiment analysis in micro-blogging. Second Joint Conference on Lexical and Computational Semantics - Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 993-1022.

Boldyreva, E. (2018). Cambridge Analytica: Ethics And Online Manipulation With Decision-Making Process. Proceedings of the 18th PCSF 2018 - Professional Culture of the Specialist of the Future. doi:10.15405/epsbs.2018.12.02.10

Bowes, A., & Katz, A. (2011). When Sarcasm Stings. Discourse Processes, 215-236. doi:10.1080/0163853X.2010.532757

Boyd, R., & Pennebaker, J. (2017). Language-based personality: A new approach to personality in a digital world. Current Opinion in Behavioral Sciences. 18. 63-68. 10.1016/j.cobeha.2017.07.017.

Bricker, J. (2005). Development and evaluation of the Air Travel Stress Scale. Journal of Counseling Psychology, 615-628.

Buechel, S., Buffone, A., Slaff, B., Ungar, L., & Sedoc, J. (2018). Modeling empathy and distress. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics. doi:10.18653/v1/D18-1507

Burghartz, R., & Berberich, K. (2014). MPI-INF at the NTCIR-11 Temporal Query Classification Task. Proceedings of the 11th NTCIR Conference. Tokyo, Japan.

Cachola, I., Holgate, E., Preotiuc-Pietro, D., & Li, J. (2018). Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media. Proceedings of the 27th International Conference on Computational Linguistics, (pp. 2927-2938). Santa Fe.

Cai, Y., Cai, H., & Wan, X. (2019). Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. *ACL*.

Calabrese, K. (2000). Interpersonal Conflict and Sarcasm in the Workplace. Genetic, social, and general psychology monographs, 459-494.

Caselli, T., Basile, V., Mitrovic, J., & Granitzer, M. (2020). HateBERT: Retraining BERT for Abusive Language Detection in English. *ArXiv, abs/2010.12472*.

Chang, A., & Manning, C. (2012). SUTime: A library for recognizing and normalizing time expressions. Proceedings of the 8th International Conference on Language Resources and Evaluation, (pp. 3735-3740).

Chen, H. (2017). Travel well, road warriors: Assessing business travelers' stressors. Tourism Management Perspectives. doi:10.1016/j.tmp.2016.12.005

Chen, X., Liu, Z., & Sun, M. (2014). A Unified Model for Word Sense Representation and Disambiguation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), (pp. 1025-1035).

Chen, Y., Yuan, J., You, Q., & Luo, J. (2018). Twitter Sentiment Analysis via Bi-sense Emoji Embedding and Attention-based LSTM. arXiv.

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. International Conference on and 2012 International Confernece on Social Computing (SocialCom).

Choi, J. (2009). Using Heart Rate Monitors to Detect Mental Stress. International Workshop on Wearable and Implantable Body Sensor Networks (pp. 219-223). BSN 2009.

Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. Proceedings of the Conference on Human Factors in Computing Systems, (pp. 3267-3276). doi:10.1145/2470654

Chun, J., Lee, J., Kim, J., & Lee, S. (2020). An international systematic review of cyberbullying measurements. *Comput. Hum. Behav., 113*, 106485.

Chung, C., & Pennebaker, J. (2007). The Psychological Functions of Function Words. Social communication.

Cohen, S., Janicki-Deverts, D., & Miller, G. (2007). Psychological stress and disease. The Journal of the American Medical Association, 1685-1687.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. Journal of Health and Social Behaviour, 24(4), 385-396. doi:10.2307/2136404

Cohen, S., Kessler, R. C., & Gordon, L. U. (1995). Measuring stress: A guide for health and social scientists. Oxford University Press.

Colston, H. L., & Gibbs, R. W., Jr. (2007). A brief history of irony. In R. W. Gibbs, Jr. & H. L. Colston (Eds.), *Irony in language and thought: A cognitive science reader* (pp. 3–21). Lawrence Erlbaum Associates Publishers.

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying Mental Health Signals in Twitter. Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, (pp. 51-60). doi:10.3115/v1/W14-3207

Coppersmith, G., Harman, C., & Dredze, M. (2014). Measuring post traumatic stress disorder in twitter. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM, (pp. 579-582).

Cowen, A., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proceedings of the National Academy of Sciences of the United States of America. 114. 10.1073/pnas.1702247114.

Crotts, J., Zehrer, A. (2012). An Exploratory Study of Vacation Stress. Tourism Analysis. 17. 547-552. 10.3727/108354212X13473157390920.

Dabiri, S., & Heaslip, K. (2019). Developing a Twitter-based traffic event detection model using deep learning architectures. *Expert systems with applications*, *118*, 425-439.

Deerwester, S. (1988). Improving Information Retrieval with Latent Semantic Indexing. Proceedings of the 51st Annual Meeting of the American Society for Information Science 25, (pp. 36-40).

Defrank, R., Konopaske, R., & Ivancevich, J. (2000). Executive travel stress: Perils of the road warrior. Academy of Management Perspectives, 58-71.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT.

Dews, S., Kaplan, J., & Winner, E. (1995). Why not say it directly? The social functions of irony. *Discourse Processes, 19*, 347-367.

Dickinson, T., Fernández, M., Thomas, L., Mulholland, P., Briggs, P., & Alani, H. (2016). Identifying Important Life Events from Twitter Using Semantic and Syntactic Patterns. Proceedings of the WWW/Internet Conference (pp. 143-150). IADIS Press.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. Annual Review of Psychology, 417-440.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate Speech Detection with Comment Embeddings. (pp. 29-30). Proceedings of the 24th International Conference of the World Wide Web. doi:10.1145/2740908.2742760

Domingo, C., Gonzalez-Ferrero, T., & Gonzalez-Dios, I. (2021). What is on Social Media that is not in WordNet? A Preliminary Analysis on the TwitterAAE Corpus. *GWC*.

Dong, X., Li, C., & Choi, J.D. (2020). Transformer-based Context-aware Sarcasm Detection in Conversation Threads from Social Media. *ArXiv, abs/2005.11424*.

Dos Santos, C., Melnyk, I., & Padhi, I. (2018). Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer. arXiv.

Dynel, M. (2017). The Irony of Irony: Irony Based on Truthfulness. Journal of Corpus Pragmatics, 3-36. doi:10.1007/s41701-016-0003-6

Eichstaedt, J., Smith, R., Merchant, R., Ungar, L., Crutchley, P., Preotiuc-Pietro, D., Schwartz, H. (2018). Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences. doi:10.1073/pnas.1802331115

Eisenstein, J., Ahmed, A., & Xing, E. (2011). Sparse Additive Generative Models of Text. Proceedings of the 28 th International Conference on Machine Learning. Bellevue, WA, USA.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). Emotion in the human face: Guidelines for research and an integration of findings. Pergamon Press.

Fan, L., Yu, H., & Yin, Z. (2020). Stigmatization in social media: Documenting and analyzing hate speech for COVID-19 on Twitter. Proceedings of the Association for Information Science and Technology. Association for Information Science and Technology, *57*(1), e313. https://doi.org/10.1002/pra2.313

Filannino, M., & Nenadic, G. (2014). Using machine learning to predict temporal orientation of search engines' queries in the Temporalia challenge. Proceedings of NTCIR-11, EVIA 2014 (NII Testbeds and Community for Information access Research).

Fraisse, P. (1984). Perception and Estimation of Time. Annual Review of Psychology, 1-37.

Fredrikson, M., & Furmark, T. (2006). Brain Mechanisms In Stress and Negative Affect. In B. B. Arnetz, & R. Ekman, Stress in Health and Disease. doi:10.1002/3527609156.ch14

Gajarla, V., & Gupta, A. (2015). Emotion Detection and Sentiment Analysis of Images. Georgia Institute of Technology .

Gamallo, P., & Garcia, M. (2014). Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), (pp. 171-175).

Gamon, M., Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting Depression via Social Media. Proceedings of Association for the Advancement of Artificial Intelligence.

Gayo-Avello, D. (2012). No, you cannot predict elections with Twitter. IEEE Internet Computing, 91-94.

Giachanou, A., Ríssola, E. A., Ghanem, B., Crestani, F., & Rosso, P. (2020). The Role of Personality and Linguistic Patterns in Discriminating Between Fake News Spreaders and Fact Checkers. Natural Language Processing and Information Systems: 25th International Conference on Applications of Natural Language to Information Systems, NLDB 2020, Saarbrücken, Germany, June 24–26, 2020, Proceedings, 12089, 181–192. https://doi.org/10.1007/978-3-030-51310-8_17

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Smith, N. (2011). Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In Proceedings of The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, (pp. 42-47). Portland, Oregon.

Giora, R. (1998). Irony: Grade Salience and Indirect Negation. Journal of Metaphor and Symbol , 83-101.

Godin, F., Vandersmissen, B., De Neve, W., & Van de Walle, R. (2015). Multimedia Lab @ ACL W-NUT NER shared task: Named entity. Proceedings of ACL - IJCNLP, (pp. 146-150).

Goel, A., Gautam, J., & Kumar, S. (2016). Real time sentiment analysis of tweets using Naive Bayes. 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), (pp. 257-261).

Goldstein, D., & Kopin, I. (2007). Evolution of concepts of stress. Stress, 109-120.

Gomathi, K., Ahmed, S., Sreedharan, J., & Aman, H. (2012). Psychological Health of First-Year Health Professional Students in a Medical University in the United Arab Emirates. Sultan Qaboos University medical journal. 12. 206-213. 10.12816/0003114.

González-Ibáñez, R. &. (2011). Identifying sarcasm in twitter: A closer look. Proceedings of The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.

Grice, H. P. (1975). Logic and Conversation. In P. C. Morgan, Syntax and Semantics, Vol. 3, Speech Acts (pp. 41-58). New York: Academic Press.

Grondin, S. (2010). Timing and time perception: A review of recent behavioral and neuroscience findings and theoretical directions. *Attention, Perception, & Psychophysics, 72*, 561-582.

Grondin, S., Mendoza-Duran, E., & Rioux, P. A. (2020). Pandemic, Quarantine, and Psychological Time. *Frontiers in psychology*, *11*, 581036. https://doi.org/10.3389/fpsyg.2020.581036

Gruda, J., & Hasan, S. (2019) Feeling anxious? Perceiving anxiety in tweets using machine learning. Computers in Human Behavior, 98. pp. 245-255. ISSN 0747-5632

Guntuku, S. C., Buffone, A., Jaidka, K., Eichstaedt, J., & Ungar, L. (2018). Understanding and Measuring Psychological Stress using Social Media. Proceedings of the International Conference on Web and Social Media.

Hallmann, K., Kunneman, F., Liebrecht, C., Van den Bosch, A., & van mulken, M. (2016). Sarcastic Soulmates, Intimacy and irony markers in social media messaging. Linguistic Issues in Language Technology, 1-23.

Hasanuzzaman, M., Kamila, S., Kaur, M., & Ekbal, A. (2017). Temporal Orientation of Tweets for Predicting Income of Users. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), (pp. 659-665).

Hazarika, D., Poria, S., Gorantla, S., Cambria, E., Zimmermann, R., & Mihalcea, R. (2018). CASCADE: Contextual Sarcasm Detection in Online Discussion Forums. *ArXiv, abs/1805.06413*.

Holgate, E., Cachola, I., & Preotiuc-Pietro, D., & Li, J. (2018). Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions. 4405-4414. 10.18653/v1/D18-1471.

Holman, E., & Silver, R. C. (1998). Getting "stuck" in the past: temporal orientation and coping with trauma. Journal of personality and social psychology, 1146-1163.

Holmes, T. H., & Rahe, R. H. (1967). The Social Readjustment Rating Scale. Journal of Psychosomatic Research, 213-218.

Hoover, J., Portillo-Wightman, G. J., Yeh, L., Havaldar, S., Mostafazadeh D., A., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., Moreno, G., Park, C., Chang, T., Chin, J., Leong, C., Leung, J., Mirinjian, A., Dehghani, M. (2019). Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment. 10.31234/osf.io/w4f72.

Hou, Y., Tan, C., Xu, J., Pan, Y., Chen, Q., & Wang, X. (2014). HITSZ-ICRC at NTCIR-11 Temporalia Task. Proceedings of NTCIR-11.

Huang, L., Gino, F., & Galinsky, A. (2015). The highest form of intelligence: Sarcasm increases creativity for both expressers and recipients. Organizational Behavior and Human Decision Processes. doi:10.1016/j.obhdp.2015.07.001

Hughes, B. (2019). The Psychology of Brexit: From Psychodrama to Behavioural Science. Palgrave Macmillan.

Hulpus, I., Prangnawarat, N., & Hayes, C. (2015). Path-Based Semantic Relatedness on Linked Data and Its Use to Word and Entity Disambiguation. International Semantic Web Conference, (pp. 442-457).

Hung, C., & Chen, S. (2016). Word sense disambiguation based sentiment lexicons for sentiment classification. Knowledge Based Systems, 224-232.

Hutto, C., & Gilbert, E. (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media. Ann Arbor, MI.

Janssen, S., & Rubin, D. (2011). Age Effects in Cultural Life Scripts. Applied Cognitive Psychology, 291-298.

Jay, T., & Janschewitz, K. (2008). The pragmatics of swearing. Journal of Politeness Research-language Behaviour Culture, 267-288.

Jia, J. (2018). Mental Health Computing via Harvesting Social Media Data. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18).

Jia, J., Lin, H., Guo, Q., Xue, Y., Huang, J., Cai, L., & Feng, L. (2014). Psychological stress detection from cross-media microblog data using Deep Sparse Neural Network. Proceedings of the 2014 IEEE International Conference on Multimedia and Expo (ICME), (pp. 1-6). doi:10.1109/ICME.2014.6890213

Joho, H., Jatowt, A., & Blanco, R. (2014). NTCIR temporalia: a test collection for temporal information access research. WWW '14 Companion: Proceedings of the 23rd International Conference on World Wide Web, (pp. 845-850).

Jonsson, E., & Stolee, J. (2016). An Evaluation of Topic Modelling Techniques for Twitter.

Jorgensen, J. (1996). The functions of sarcastic irony in speech. Journal of Pragmatics, 613-634. doi:10.1016/0378-2166(95)00067-4

Joshi, A., Tripathi, V., Bhattacharyya, P., Carman, M., Singh, M., Saraswati, J., & Shukla, R. (2016). How Challenging is Sarcasm versus Irony Classification ? : An Analysis From Human and Computational Perspectives. Proceedings of Australasian Language Technology Association Workshop, (pp. 123-127).

Kamila, S. (2018). Fine-Grained Temporal Orientation and its Relationship with Psycho-Demographic Correlates. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).

Kamila. S., Hasanuzzaman, M., Ekbal, A., & Bhattacharyya, P., (2019) Resolution of grammatical tense into actual time, and its application in Time Perspective study in the tweet space. PLOS ONE 14(2): e0211872. https://doi.org/10.1371/journal.pone.0211872

Karanatsiou, D., Sermpezis, P., Gruda, J., Kafetsios, K., & Dimitriadis, I., & V, Athena. (2020). My tweets bring all the traits to the yard: Predicting personality and relational traits in Online Social Networks.

Karoui, J., Benamara, F., Moriceau, V., Aussenac-Gilles, N., & Belguith, L. (2015). Towards a Contextual Pragmatic Model to Detect Irony in Tweets. 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), (pp. 644-650). doi:10.3115/v1/P15-2106

Kenyon-Dean, K., Ahmed, E., Fujimoto, S., Georges-Filteau, J., Glasz, C., Kaur, B., . . . Ruths, D. (2018). Sentiment Analysis: It's Complicated! Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), (pp. 1886-1895). doi:10.18653/v1/N18-1171

Kshirsagar, R., Cukuvac, T., McKeown, K., & McGregor, S. (2018). Predictive Embeddings for Hate Speech Detection on Twitter. 26-32. 10.18653/v1/W18-5104.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. doi:10.3115/v1/D14-1181

Kircaburun, K., & Griffiths, M. (2018). Instagram addiction and the Big Five of personality: The mediating role of self-liking. Journal of behavioral addictions. doi:10.1556/2006.7.2018.15

Kittel, J., Kornitzer, M., & Dramaix, M. (1986). Evaluation of Type A personality. Journal of Postgraduate Medicine, 62(730), 781-783. doi:10.1136/pgmj.62.730.781

Kolchinski, Y.A., & Potts, C. (2018). Representing Social Media Users for Sarcasm Detection. *ArXiv, abs/1808.08470*.

Kolmes, K. (2012). Social media in the future of professional psychology. *Professional Psychology: Research and Practice, 43*(6), 606–612. https://doi.org/10.1037/a0028678

Kotikalapudi, R., & Chellappan, S. (2012). Associating Internet usage with depressive behavior among college. IEEE Technol SocMag, 31, 73-80.

Krippendorff, K. (2004). Content analysis: An Introduction to its Methodology. Thousand Oack, CA.

Kulkarni, V., Kern, M., Stillwell, D., Kosinski, M., Matz, S., Ungar, L., Schwartz, H. (2017). Latent Human Traits in the Language of Social Media: An Open-Vocabulary Approach. PLOS ONE, 13(11). doi:10.1371/journal.pone.0201703

Kuss, D., & Griffiths, M. (2011). Online Social Networking and Addiction—A Review of the Psychological Literature. International Journal of Environmental Research and Public Health. doi:10.3390/ijerph8093528

Kuss, D., & Griffiths, M. (2012). Online gaming addiction in children and adolescents:A review of empirical research. Journal of Behavioural Addictions, 1-20. doi:10.1556/JBA.1.2012.1.1

Lazarus, R. (1993). From psychological stress to the emotions: a history of changing outlooks. Annual review of psychology, 44, 1-21 .

Lee, P. S., Sohn, J. N., Lee, Y. M., Park, E. Y., & Park, J. S. (2005). *Taehan Kanho Hakhoe chi*, *35*(1), 195–205. https://doi.org/10.4040/jkan.2005.35.1.195

Leis, A., Ronzano, F., Mayer, M., Furlong, L., & Sanz, F. (2019). Detecting Signs of Depression in Tweets in Spanish: Behavioral and Linguistic Analysis. Journal of Medical Internet Research. 21. e14199. 10.2196/14199.

Lemmens, J., Burtenshaw, B., Lotfi, E., Markov, I., & Daelemans, W. (2020). Sarcasm Detection Using an Ensemble Approach. 264-269. 10.18653/v1/2020.figlang-1.36.

Lewis, P., Newburn, T., Taylor, M., Mcgillivray, C., Greenhill, A., Frayman, H., & Procter, R. (2011). Reading the Riots: Investigating England's Summer of Disorder. The London School of Economics and Political Science and The Guardian, London, UK.

Li, J., Ritter, A., Cardie, C., & Hovy, E. (2014). Major Life Event Extraction from Twitter based on Congratulations/Condolences Speech Acts. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), (pp. 1997-2007).

Li, Y., & Fleyeh, H. (2018). Twitter Sentiment Analysis of New IKEA Stores Using Machine Learning. International Conference on Computer and Applications.

Liebrecht, C., Kunneman, F., & Van den Bosch, A. (2013). The perfect solution for detecting sarcasm in tweets #not. Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, (pp. 29-37).

Lin, H., Jia, J., Huang, J., Zhou, E., Fu, J., Liu, Y., & Luan, H. (2016). Moodee: An Intelligent Mobile Companion for Sensing Your Stress from Your Social Media Postings. AAAI. Proceedings of the AAAI 2016.

Lin, H., Jia, J., Guo, Q., Xue, Y., Li, Q., Huang, J., Feng, L. (2014). User-level psychological stress detection from social media using deep neural network. Proceedings of the 22nd ACM international conference on Multimedia, (pp. 507-516).

Lin, Jia, J., Nie, L., Shen, G., & Chua, T. S., (2016). What does social media say about your stress? Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, (pp. 3775-3781).

Lin, Jia, J., Qiu, J., Zhang, Y., Shen, G., Xie, L., Chua, T.S., (2017). Detecting Stress Based on Social Interactions in Social Networks. IEEE Transactions on Knowledge and Data Engineering. doi:10.1109/TKDE.2017.2686382

Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

Liu, L., Preotiuc-Pietro, D., Samani, Z., Moghaddam, M., & Ungar, L. (2016). Analyzing Personality through Social Media Profile Picture Choice. ICWSM.

Liu, P., Li, W., & Zou, L. (2019). NULI at SemEval-2019 Task 6: Transfer learning for offensive language detection using bidirectional transformers. Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval).

Lloyd, S. (1982). Least squares quantization in PCM. IEEE Transactions Information Theory, 129-137.

Maas, A., Daly, R., Pham, P., Huang, D., Ng, A., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. Proceedings of The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, (pp. 142-150). Portland, Oregon.

Mahata, D., Zhang, H., Uppal, K., Kumar, Y., Shah, R. R., & Laiba, S. S. (2019). MIDAS at SemEval-2019 Task 6: Identifying offensive posts. Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval).

Malmasi, S., & Zampieri, M. (2017). Detecting Hate Speech in Social Media. Proceedings of RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning, (pp. 467-472).

Maynard, D., & Greenwood, M. (2014). Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. Proceedings of LREC, (pp. 4238-4243).

McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., & & Spiro, E. S. (2017). Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. . Sociological methods & research, 390-421.

McCrae, R., & John, O. (1992). An Introduction to the Five-Factor Model and Its Applications. Journal of Personality, 175-215. doi:10.1111/j.1467-6494.1992.tb00970.x

McIntosh, I., Swanson, V., Power, K., Raeside, F., & Dempster, C. (2006). Anxiety and Health Problems Related to Air Travel. Journal of Travel Medicine, 198-204.

Mehdad, Y., & Tetreault, J. (2016). Do Characters Abuse More Than Words? Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, (pp. 299-303). doi:10.18653/v1/W16-3638

Mehl, M., & Pennebaker, J. (2003). The Sounds of Social Life: A Psychometric Analysis of Students' Daily Social Environments and Natural Conversations. Journal of personality and social psychology, 857-870. doi:10.1037/0022-3514.84.4.857

Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013). Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling. The 36th Annual ACM SIGIR Conference, (pp. 889-892).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neutral information processing systems, 3111-3119.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1991). Introduction to WordNet: An On-line Lexical Database. International Journal of Lexicography .

Mimno, D., Wallach, H., Talley, E., Leenders, M., & Mccallum, A. (2011). Optimizing Semantic Coherence in Topic Models. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, (pp. 262-272). Edinburgh, UK.

Monika, R., Deivalakshmi, S., & Janet, B. (2019). Sentiment Analysis of US Airlines Tweets Using LSTM/RNN. Proceedings of the 2019 IEEE 9th International Conference on Advanced Computing (IACC), (pp. 91-95).

Mooney, A., Earl, J. K., Mooney, C. H., & Bateman, H. (2017). Using Balanced Time Perspective to Explain Well-Being and Planning in Retirement. *Frontiers in psychology*, *8*, 1781. https://doi.org/10.3389/fpsyg.2017.01781

Moreno, M., Jelenchick, L., Egan, K., Cox, E., Young, H., Gannon, K., & Becker, T. (2011). Feeling bad on Facebook: Depression disclosure of students on a social networking site. Depression and anxiety, 28(6), 447-455.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. *COMPLEX NETWORKS*.

Mrva-Montoya, A. (2012). Social Media: New Editing Tools or Weapons of Mass Distraction? The Journal of Electronic Publishing.

Myers, I. (1962). The Myers-Briggs Type Indicator: Manual. Paolo Alto, California, USA: Consulting Psychologists Press. doi:DOI:10.1037/14404-000

Naskar, D., Mokeddem, S. A., Rebollo, M., & Onaindia, E. (2016). Sentiment Analysis in Social Networks through Topic Modeling. . Proceeings of the Language Resources and Evaluation Conference (LREC) 2016.

Navigli, R. (2009). Word Sense Disambiguation: A Survey. ACM Computing Surveys.

Navigli, R., & Ponzetto, S. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 217-250.

Newman, D., Noh, Y., Talley, E., Karimi, S., & Baldwin, T. (2010). Evaluating topic models for digital libraries. Proceedings of the ACM International Conference on Digital Libraries, (pp. 215-224). doi:10.1145/1816123.1816156

Nguyen, T., & Shirai, K. (2015). Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction . Proceedings of The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP).

Nikolov, A., & Radivchev, V. (2019). Nikolov-Radivchev at SemEval-2019 Task 6: Offensive tweet classification with BERT and ensembles. Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval).

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive Language Detection in Online User Content. Proceedings of the World Wide Web Conference - WWW 2016, (pp. 145-153).

Nuttin, J., & Lens, W. (1985). Future time perspective and motivation: Theory and research method. Leuven University Press & Erlbaum.

Olson, M. (1965). The logic of collective action : public goods and the theory of groups. Cambridge, Massachusetts: Harvard University Press.

Oscar, N., Fox, P., Croucher, R., Wernick, R., Keune, J., & Hooker, K. (2017). Machine Learning, Sentiment Analysis, and Tweets: An Examination of Alzheimer's Disease Stigma on Twitter. The Journals of Gerontology Series B Psychological Sciences and Social Sciences.

Ozalp, S., Williams, M., Burnap, P., Liu, H., & Mostafa, M. (2019). Antisemitism on Twitter: Collective Efficacy and the Role of Community Organisations in Challenging Online Hate Speech. 10.1177/2056305120916850.

Pal, A., & Saha, D. (2015). Word sense disambiguation: a survey. International Journal of Control Theory and Computer Modeling.

Park, G., Schwartz, H., Sap, M., Kern, M., Weingarten, E., Eichstaedt, J., Seligman, M. (2015). Living in the Past, Present, and Future: Measuring Temporal Orientation With Language. Journal of Personality .

Park, J., & Fung, P. (2017). One-step and Two-step Classification for Abusive Language Detection on Twitter. Proceedings of the First Workshop on Abusive Language Online, (pp. 41-45).

Patwa, P., Aguilar, G., Kar, Sudipta, Pandey, S., PYKL, S., Chakraborty, T. S. (2020). SemEval-2020 Sentimix Task 9: Overview of SENTIment Analysis of Code-MIXed Tweets. Proceedings of the 14th International Workshop on Semantic Evaluation-2020. Association for Computational Linguistics.

Pelicon, A., Martinc, M., & Novak, P. K. (2019). Embeddia at SemEval-2019 Task 6: Detecting hate with neural network and transfer learning. Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval).

Peng, R., Barreto, A., Gao, Y., & Adjouadi, M. (2011). Affective assessment of computer users based on processing the pupil diameter signal. Annual International Conference of the IEEE Engineering in Medicine and Biology Society, (pp. 2594-2597). Boston.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology, *77*(6), 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representations. Proceedings of the conference on empirical methods in natural language processing, (pp. 1532-1543).

Pinker, S. (2007). The stuff of thought: Language as a window into human nature. New York: Penguin.

Pitsilis, G., Ramampiaro, H., & Langseth, H. (2018). Effective hate-speech detection in Twitter data using recurrent neural networks. Journal of Applied Intelligence.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Theories of emotion*, 1, 3--31.

Poria, S., Cambria, E., Hazarika, D., Vij, P. (2016). A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. Proceedings of COLING.

Rajadesingan, A., Zafarani, R., & Liu, H. (2015). Sarcasm Detection on Twitter: A Behavioral Modeling Approach. Proceedings of the 8th ACM International Conference on Web Search and Data Mining, (pp. 97-106). doi:10.1145/2684822.2685316

Rangwani, H., Kulshreshtha, D., & Singh, A. (2018). NLPRL-IITBHU at SemEval-2018 Task 3: Combining Linguistic Features and Emoji pre-trained CNN for Irony Detection in Tweets. Proceedings of The 12th International Workshop on Semantic Evaluation SemEval-2018, (pp. 638-642). doi:10.18653/v1/S18-1104.

Rentoumi, V., Giannakopoulos, G., Karkaletsis, V., & Vouros, G. (2009). Sentiment Analysis of Figurative Language using a Word Sense Disambiguation Approach.

Reuter, J., Pereira-Martins, J., & Kalita, J. (2016). Segmenting twitter hashtags. International Journal on Natural Language Computing, (pp. 5:23–36)

Reyes, A., Rosso, P., & Veale, T. (2013). A multidimensional approach for detecting irony in Twitter. Language Resources and Evaluation.

Riker, W., & Ordeshook, P. (1968). A Theory of the Calculus of Voting. American Political Science Review, 25-42. doi:10.1017/S000305540011562X

Riloff, E., Qadir, A., Surve, P., Silva, L., Gilbert, N., & Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. Proceedings of EMNLP, (pp. 704-714).

Risch, J., & Krestel, R. (2020). Bagging BERT Models for Robust Aggression Identification. *TRAC*.

Roche, M., & Jacobson, N. (2019). Elections Have Consequences for Student Mental Health: An Accidental Daily Diary Study. Psychological Reports. doi:10.1177/0033294118767365

Rohanian, O., Taslimipoor, S., Evans, R., & Mitkov, R. (2018). WLV at SemEval-2018 Task 3: Dissecting Tweets in Search of Irony. Proceedings of The 12th International Workshop on Semantic Evaluation SemEval-2018, (pp. 553-559). doi:10.18653/v1/S18-1090

Roubinov, D. S., Hagan, M. J., & Luecken, L. J. (2012). If at first you don't succeed: the neuroendocrine impact of using a range of strategies during social conflict. Journal of Anxiety, stress, and coping, 397-410. doi:10.1080/10615806.2011.6134

Rousseeuw, P. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics, 53-65. doi:10.1016/0377-0427(87)90125-7

Salas Zarate, M., Medina, J., Lagos-Ortiz, K., Luna Aveiga, H., Rodríguez-García, M., & Valencia-García, R. (2017). Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach. Journal of Computational and Mathematical Methods in Medicine, 1-9.

Savage, B., Lujan, H., Thipparthi, R., & Dicarlo, S. (2017). Humor, laughter, learning, and health! A brief review. Advances in Physiology Education, 341-347. doi:10.1152/advan.00030.2017

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media (pp. 1-10). Valencia, Spain: Assocaition for Computational Linguistics.

Schmitt, N. (2002). An Introduction to Applied Linguistics. London: Arnold Publishers.

Schwartz, H., Park, G., Sap, M., Weingarten, E., Eichstaedt, J., Kern, M., . . . Ungar, L. (2015). Extracting Human Temporal Orientation in Facebook Language. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, CO.

Segalla, M., Catalin, C., Travel, C., Rouzies, D., & Lebunetel, V. (2012). Global Business Travel Builds Sales and Stress.

Selye, H. (1956). The Stress of Life. McGraw Hill.

Sha, H., Hasan, M., Mohler, G., & Brantingham, P. (2020). Dynamic topic modeling of the COVID-19 Twitter narrative among U.S. governors and cabinet executives. *ArXiv, abs/2004.11692*.

Shah, A., Shah, D., & Majumder, P. (2014). Andd7 @ NTCIR-11 Temporal Information Access Task. Proceedings of the 11th NTCIR Conference. Tokyo, Japan.

Shi, Y., Ruiz, N., Taib, R., Choi, E., & Chen, F. (2007). Galvanic skin response (GSR) as an index of cognitive load. Conference on Human Factors in Computing Systems - Proceedings, (pp. 2651-2656). doi:10.1145/1240866.1241057

Sirois, F. M., & Pychyl, T. A. (Eds.). (2016). *Procrastination, health, and well-being.* Elsevier Academic Press.

Smailović, J., Grčar, M., Lavrac, N., & Žnidaršič, M. (2013). Predictive sentiment analysis of tweets: A stock market application. In Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data.

Smith, K., Hibbing, M., & Hibbing, J. (2019). Friends, relatives, sanity, and health: The costs of politics. PLOS ONE 14(9).

Sokolova, M., Huang, K., Matwin, S., Ramisch, J., Sazonova, V., Black, R., Sambuli, N. (2016). Topic Modelling and Event Identification from Twitter Textual Data. ArXiv.

Souri, A., Rahmani, A., & Hosseinpour, S. (2018). Personality classification based on profiles of social networks' users and the five-factor model of personality. Human-centric Computing and Information Sciences, 8-24. doi:10.1186/s13673-018-0147-4

Stephens, R., & Umland, C. (2011). Swearing as a Response to Pain-Effect of Daily Swearing Frequency. The journal of pain : official journal of the American Pain Society, 1274-1288.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring Topic Coherence over many models and many topics. Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). The general inquirer: A computer approach to content analysis. Cambridge, MA: The MIT Press.

Strötgen, J., & Gertz, M. (2010). HeidelTime: High quality rule-based extraction and normalization of temporal expressions. Proceedings of the 5th International Workshop on Semantic Evaluation, (pp. 321-324).

Strötgen, J., Bögel, T., Zell, J., Armiti, A., Canh, T., & Gertz, M. (2014). Extending HeidelTime for Temporal Expressions Referring to Historic Dates. Proceedings of the Language Resources and Evaluation Conference.

Strötgen, J., Zell, J., & Gertz, M. (2013). HeidelTime: Tuning english and developing Spanish resources for TempEval-3. Proceedings of the 7th International Workshop on Semantic Evaluation, (pp. 15-19).

Subramanian, J., Sridharan, V., Shu, K., & Liu, H. (2019). Exploiting Emojis for Sarcasm Detection.

Sumanth, C., & Inkpen, D. (2015). How much does word sense disambiguation help in sentiment analysis of micropost data? Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, (pp. 115-121).

Sun, C., Huang, L., & Qiu, X. (2019). Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. 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).

Tabari, N., & Hadzikadic, M. (2019). Context sensitive sentiment analysis of financial tweets: A new dictionary. In Intelligent Methods and Big Data in Industrial Applications.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 24-54.

Thelwall, M. (2008). Fk yea I swear: Cursing and gender in MySpace. Corpora, 83-107. doi:10.3366/E1749503208000087

Thelwall, M. (2017). TensiStrength: Stress and relaxation magnitude detection for social media texts. Information Processing and Management, 106-121.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 2544-2558.

Thorstad, R., & Wolff, P. (2018, February 20). A big data analysis of the relationship between future thinking and decision-making. Proceedings of the National Academy of Sciences of the United States of America, pp. 1740-1748.

Tungthamthiti, P., Shirai, K., & Mohd, M. (2014). Recognition of Sarcasms in Tweets Based on Concept Level Sentiment Analysis and Supervised Learning Approaches. Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing, (pp. 404-413).

Tutaysalgir, E., Karagoz, P., & Toroslu, I. (2019). Clustering based personality prediction on turkish tweets. 825-828. 10.1145/3341161.3343513.

Unsvåg, E., & Gambäck, B. (2018). The Effects of User Features on Twitter Hate Speech Detection. 75-85. 10.18653/v1/W18-5110.

UzZaman, N., Llorens, H., Allen, J., Derczynski, L., Verhagen, M., & Pustejovsky, J. (2012). TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations. arXiv.

Van Hee, C., Lefever, E., & Hoste, V. (2018). Exploring the fine-grained analysis and automatic detection of irony on Twitter. Language Resources and Evaluation. doi:10.1007/s10579-018-9414-2

Vayansky, I., Kumar, S., & Li, Z. (2019). An Evaluation of Geotagged Twitter Data during Hurricane Irma Using Sentiment Analysis and Topic Modeling for Disaster Resilience. *2019 IEEE International Symposium on Technology and Society (ISTAS)*, 1-6.

Verhagen, M., Gaizauskas, R., Schildre, F., Hepple, M., Katz, G., & Pustejovsky, J. (2007). SemEval-2007 task 15: TempEval temporal relation identification. SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations, (pp. 75-80).

Vingerhoets, A., Bylsma, L., & Vlam, C. (2013). Swearing: A Biopsychosocial Perspective. Psychological Topics, 287-304.

Vu, T., Nguyen, D. Q., Vu, X.-S., Nguyen, D., Catt, M., & Trenell, M. (2018). NIHRIO at SemEval-2018 Task 3: A Simple and Accurate Neural Network Model for Irony Detection in Twitter. Proceedings of the 12th International Workshop on Semantic Evaluation SemEval-2018.

Wajnryb, R. (2005). *Expletive deleted: A good look at bad language*. New York: Free Press.

Wang, B., Liakata, M., Tsakalidis, A., Kolaitis, S., Papadopoulos, S., Apostolidis, L., Kompatsiaris, I. (2017). TOTEMSS: Topic-based, Temporal Sentiment Summarisation for Twitter. Proceedings of the International Joint Conference on Natural Language Processing. Taipei, Taiwan.

Wang, P. Y. (2013). #Irony or #Sarcasm - A Quantitative and Qualitative Study Based on Twitter. 27th Pacific Asia Conference on Language, Information, and Computation, (pp. 349-356).

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. (2014). Cursing in English on Twitter. Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW.

Wang, Y. (2015). Understanding Personality through Social Media. Journal of Psychology.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. Proceedings of the Second Workshop on Language in Social Media, (pp. 19-26).

Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. Proceedings of the NAACL Student Research Workshop, (pp. 88-93). doi:10.18653/v1/N16-2013

Wiegand, M., Ruppenhofer, J., Schmidt, A., & Greenberg, C. (2018). Inducing a Lexicon of Abusive Words – a Feature-Based Approach. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), (pp. 1046-1056). doi:10.18653/v1/N18-1095

Williams, M., & Burnap, P. (2015). Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data. British Journal of Criminology.

Witowska, J., Zajenkowski, M., & Wittmann, M. (2020). Integration of balanced time perspective and time perception: The role of executive control and neuroticism. *Personality and Individual Differences, 163*, 110061.

Wolpert, D.H., & Macready, W.G. (1995). No Free Lunch Theorems for Search, Technical Report SFI-TR-95-02-010 (Santa Fe Institute)

Wong, S., Teh, P., & Cheng, C. (2020). How Different Genders Use Profanity on Twitter?. 10.1145/3388142.3388145.

Wu, C., Wu, F., Wu, S., Liu, J., Yuan, Z., & Huang, Y. (2018). THU NGN at SemEval-2018 Task 3: Tweet Irony Detection with Densely Connected LSTM and Multi-task Learning. Proceedings of the International Workshop on Semantic Evaluation, SemEval-2018. New Orleans, LA, USA.

Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., & Zhu, T. (2020). Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. *PLoS ONE, 15*.

Yang, X., Macdonald, C., & Ounis, I. (2018). Using word embeddings in twitter election classification. *Information Retrieval Journal*, *21*(2), 183-207.

Yen, A., Huang, H., & Chen, H. (2018). Detecting Personal Life Events from Twitter by Multi-Task LSTM. *Companion Proceedings of the The Web Conference 2018*.

Yu, H., Kang, X., & Ren, F. N. (2014). TUTA1 at the NTCIR-11 Temporalia Task. Proceedings of the NTCIR-11 Conference. Tokyo, Japan.

Yu, D., Xu, D., Wang, D., & Ni, Z. (2019). Hierarchical Topic Modeling of Twitter Data for Online Analytical Processing. *IEEE Access, 7*, 12373-12385.

Yuan, Y., & Zhou, Y. (2015). Twitter Sentiment Analysis with Recursive Neural Networks.

Zamani, M., Buffone, A., & Schwartz, H. (2018). Predicting Human Trustfulness from Facebook Language.

Zampieri, M., & Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the Type and Target of Offensive Posts in Social Media. 1415-1420. 10.18653/v1/N19-1144.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). *ArXiv, abs/1903.08983*.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, Ç. (2020). SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020).

Zhao, R., Zhou, A., & Mao, K. (2016, January). Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking* (pp. 1-6).

Zhang, M., Zhang, Y., & Fu, G. (2016). Tweet Sarcasm Detection Using Deep Neural Network. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, (pp. 2449-2460).

Zhang, Z., Wu, Y., Zhao, H., Li, Z., Zhang, S., Zhou, X., & Zhou, X. (2020). Semantics-aware BERT for Language Understanding. Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020).

Zhou, T., Hu, G., & Wang, L. (2019). Psychological Disorder Identifying Method Based on Emotion Perception over Social Networks. International Journal of Environmental Research and Public Health. doi:10.3390/ijerph16060953

Zhu, J., Rosset, S., Zou, H., & Hastie, T. (2006). Multi-class AdaBoost. Statistics and its interface. doi:10.4310/SII.2009.v2.n3.a8

Zhu, J., Tian, Z., & Kubler, S. (2019). UM-IU@LING at SemEval-2019 Task 6: Identifying offensive tweets using BERT and SVMs. Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval).

Zimbardo, P., & Boyd, J. (1999). Putting Time in Perspective: A Valid, Reliable Individual-Differences Metric. Journal of Personality and Social Psychology, 1271-1288. doi:10.1037/0022-3514.77.6.1271