Nuances of Aggression in Social Media Text

Thesis submitted in partial fulfillment
of the requirements for the degree of

Masters of Science
in
Exact Humanities
by Research

by

Arjit Srivastava
201256033
arjit.srivastava@research.iiit.ac.in

International Institute of Information Technology, Hyderabad
Hyderabad - 500 032, INDIA
May, 2021

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Nuances of Aggression in Social Media Text" by Arjit Srivastava, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Dr. Manish Shrivastava
&
Late Prof. Navjyoti Singh

Dedicated to a friend, a philosopher and a guide Prof. Navjyoti Singh

Acknowledgments

First, I would like to thank my current advisor, Dr Manish Shrivastava, who agreed to take me under his supervision at a difficult time and showed me the direction to success. He instilled the belief in me when I needed it the most. Every dialogue with him has been an encouragement and a beam of positivity. Secondly, I want to thank and show my gratitude to Late Prof. Navjyoti Singh who empowered with a research mindset, and the freedom to explore areas of research which intrigued me.

I feel indebted to many people as I write this thesis. However, I am especially thankful to Avijit for keeping me motivated to pursue the research, for giving hope and having the foresight to chime in with the right advice at the right time, almost always. Sarfaraz, Madan and Venumadhav, with their constant positive reinforcement about various pursuits in life, is also a reason for this thesis taking shape. I am glad to have them for enduring me. I would also like to thank Abhinav, Abhirath, Anirudh, Amitha, Anushka, and Sanjana, who kept the ray of hope shining with their timely back-and-forth conversations. These times would be a reminder of when I was in good company.

I would also like to thank the IIIT community for giving me an inspiring environment and loads of opportunities to grow. Last, but not the least, I would finally want to mention my parents and my sister for their constant faith in me. Without them, I would not be here.

Abstract

The advent of social media has immensely increased the number of opinions and arguments voiced on the internet. Social media platforms comprise a significant part of an individual's social interaction. These interactions also generate many opinions on issues where there is a significant division—these virtual interactions, which often result in debates, manifest cases of aggression.

Various online platforms like forums, blogs, and so on help users post comments and reply to other users' comments. Some of these comments can be aggressive, hate speech, lovable, offensive languages etc. With the growing population on social media, interactions over the web have increased and have become aggressive, and related activities like cyberbullying, trolling, hate speech, etc. have also increased manifold across the globe. Thus, aggressive online behaviour incidents have become a significant source of social conflict, potentially resulting in an activity of a criminal nature.

Thus, a fundamental challenge for identifying aggression on social media is to classify it from offensive or vitriolic languages. For the task of Aggression Detection, we used a Hindi-English code-mixed dataset provided for the shared task in the 1st Workshop on Trolling, Aggression and Cyberbullying (TRAC-1). Keeping these ideas in mind, we developed a system to discriminate between Overtly Aggressive, Covertly Aggressive and Non-aggressive content in texts.

While research has been focused mostly on analyzing aggression, stance, and other dimensions of speech in isolation from each other, this work also attempts to gain an extensive and fine-grained understanding of aggression and figurative language use patterns when voicing an opinion. However, this task is daunting since natural language is fraught with ambiguities, and language in social media is boisterous. So, specialized techniques are required to handle issues related to these data streams' unstructured and dynamic nature — it can be further used in various contexts to analyze and gain insights from social behaviours.

Since the users on these social media platforms tend to write in an informal tone in real-time, it is relatively natural to mix languages as they ease communication. This factor could be attributed to these users being informal, being multi-lingual, or non-native language speakers. However, it adds another layer of complexity on top of the dynamic layer of social media data. This thesis explores and develops techniques that can further help us to gain in-depth insights from such data.

We also present a code-mixed dataset in English-Hindi, of opinion on a politico-social issue. We annotate it across multiple dimensions: aggression, hate speech, emotion arousal, and figurative language usage (such as sarcasm/irony, metaphors/similes, puns/word-play) across varied modalities. Like

the one presented, such in-depth datasets are required to analyze the not so apparent forms of verbal aggression displayed on social media and analyze the social dynamics of opinion. The thesis also hopes to understand linguistic patterns better when voicing an opinion and showing aggression. Furthermore, such datasets also facilitate classification models that leverage corpora annotated for auxiliary tasks through transfer learning, joint modelling, and semi-supervised label propagation methods.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *Man is by nature a social animal; an individual who is unsocial naturally and not accidentally is either beneath our notice or more than human. Society is something that precedes the individual. Anyone who either cannot lead the common life or is so self-sufficient as not to need to, and therefore does not partake of society, is either a beast or a god.*
>
> - Aristotle, *Politics (˜384 BC)*

From the first instance, when human societies dawned, to the current contemporary habituation of our digital ecosystem, social functioning and interactions have manifested in varied domains of knowledge and different schools of thoughts. The digital environment has changed this pursuit; which this thesis attempts to empower and advance. Let us start our journey.

1.1 Background

Humans are fundamentally social animals. That is to say, social interactions have always been paramount for humans. However, along the course of human civilisation, the level of sophistication of social interactions and the number of social interactions among humans have advanced to form a complex social need. In current times, we see social media platforms playing this role to perfection, constituting a significant component of an individual's social interaction. Not only just interaction, but these platforms are also considered information dissemination tools to express opinions and robustly share views. As pointed by Newman [7], people rely a lot on these social media tools as their primary source of information and news to connect with the world and get instant updates. Users on these platforms are incredibly dependent on these platforms as their primary source of communication. Therefore, it allows researchers to study various aspects of online human behaviour, including the public stance toward various social and political aspects.

1.2 Social media

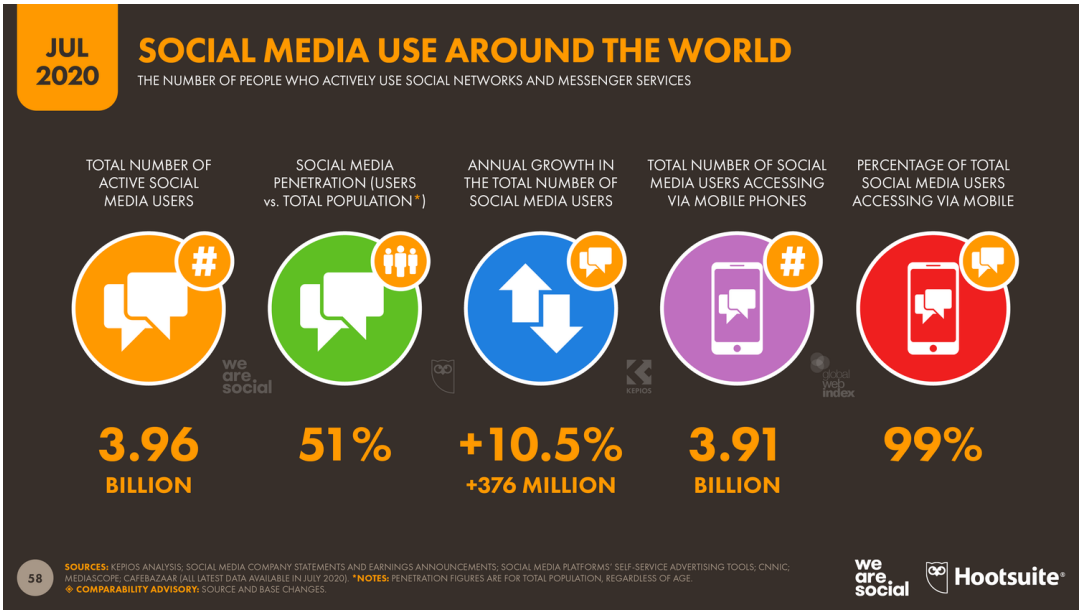

Figure 1.1 Social media usage around the world by July, 2020, from [1].

Figure 1.1 from July 2020, Global Digital Statshot report [1] shows the total number of active social media users. The advent of social media has proliferated communication on the Internet. It has continuously evolved from just being a buzzword to hugely sized media platforms such as Facebook and Twitter, where everyone has the right and power to express their opinions without any inhibitions on the topic of their choice. From July 2019 to July 2020, 376 million new users have been added, translating to almost 12 new users every second.

Since social networks feed off interactions among people, they become more powerful as they grow. Thanks to the Internet, every person with marginal views can see that they are not alone. Moreover, when these people find one another via social media, they can create memes, publications, and entire online worlds that bolster their worldview. When different opinions or ideas come together, some friction is involved, resulting in aggressive exchanges, flame wars, heated debates, hate speech, or cyberbullying.

While social media brings an increased awareness about issues, it also means that people end up misusing the freedom by using offensive language through blogs, posts, comments, or images to target groups and individuals to hurt or defame or insult them. Many depression cases, suicidal attempts, and other mental health issues because of excessive trolling and bullying on social media. So it becomes a moral responsibility to mitigate this type of behaviour.

1.3 Code Mixed content on Social Media

Since users' language on social media platforms is informal and casual, people tend to migrate towards using a hybrid form of communication to voice themselves to establish solidarity and rapport in multilingual discourse. This language interchange involves complex grammar, and the terms "code-switching" and "code-mixing", as used by Lipski [8] are used to describe it. Code-mixing refers to the use of linguistic units from different languages in a single utterance or sentence, whereas code-switching refers to the co-occurrence of speech extracts belonging to two different grammatical systems [9]. As both these phenomena are frequently observed on social media platforms in similar contexts, we use only the code-mixing scenario in this work.

With Hindi being spoken by more than 350 million people globally and English as a bridge language in India, English-Hindi is the most commonly used code-mixed language pair on social media in India. Moreover, mixing multiple languages (code-mixing) has been observed as a widespread phenomenon in social media users from language-dense areas ([10], [11]).

We will demonstrate some instances of Hindi-English code-mixed texts also transliterated in English.

Sentence 1: After so much time, India world cup jeet hi gaya.

Translation 1: After so much time, India managed to win the world cup.

Explanation 1: This sentence contains words in English such as 'After', 'so' and 'time' and words in Hindi such as 'jeet', 'gaya', etc. which are transliterated to English.

Sentence 2: Kitne saal baad we are having so much fun.

Translation 2: After so many years we are having so much fun.

Explanation 2: This sentence contains words in English such as 'we', 'are' and 'fun' and words in Hindi such as 'Kitne', 'saal', etc. which are transliterated to English.

1.4 The Nuances of Aggression

Social media offers us unprecedented opportunities to communicate, but it also comes with unfortunate malicious behaviours. For instance: cyberbullying, racism, hate speech and discrimination are some of the aggressive online behaviours. They manifest in such social media platforms and often have devastating consequences for individual users and society.

Aggression is both explicit and implicit. Explicit through inappropriate postings such as negative feelings and embarrassing photos. It is implicit when it unconsciously hurts online users, for instance: through malicious gossip spreading. The final picture is that social media users are left vulnerable and exposed to many aggression threats. A lot of inter-disciplinary studies, from social psychology to sociolinguistics and computational sciences have focused on cyberaggression. Theories of social

learning, social bonds, and planned behaviour [12] provide the basis of the theoretical formulation of aggressive online behaviour.

The first response in mitigating and managing the aggressive behaviour was to ensure that the user-generated content is moderated and monitored, which was done manually. However, the unmanageable rate at which new data is created on the Internet ensures that these manual methods of weeding out harmful user-generated content are rendered ineffective and impractical. Thus, it becomes imperative that semi-automatic or automatic means of handling such behaviour on the Internet need to come into the picture. However, dealing with this problem in an automated fashion is also not straightforward since it requires us to understand the very abstract notions of the problems we are tackling, the social media platforms we are dealing with, the systems need to be intelligent and nuanced enough to cover as many cases as possible.

Moreover, machine learning has been used to detect such behaviour in online platforms. Even with so much prior work on aggression, as argued by Corcoran et al. [13], online aggression has not been uniformly defined. There are different ways in which we can define online aggression. These could depend on the platform's type, social interactions it facilitates, the aggressor's power over the victim etc. Moreover, on top of that, the systems must be capable of recognising and handling not just the cases of overtly aggressive behaviour, but also covert aggression, too. Similarly, the system would also have to be aware of further such distinctions between aggressive behaviour.

One crucial aspect to keep in mind when we talk about the nuances of aggression is that defining a universal user on Twitter that fits our idea of public perception is inherently tricky. Without considering the socio-economic division, gender, class and location - it is tough to call an opinion a generalised, public opinion. There exists a digital divide, which affects the majority of our population. We should note that expressions of aggression and generally emotion are culture-specific. For example, a female, middle-class doctor who works in Kanpur, Uttar Pradesh, would express her opinion about demonetisation differently from a lower-middle-class, male migrant working for Zomato. Thus, class, gender, location, and many other different permutations and combinations are likely to texture such research in specific ways. However, we also need to understand that all scholars need to make certain assumptions and reduce nuance and variability to be able to develop tools and search for patterns that will help us understand a problem better.

1.5 Motivation

The use of social media networking has not just transformed individuals but has also transformed our communities. Often on social media platforms, we see that a lot of aggressive things are continuously vocalised. The idea of aggression itself is not new, but the way it is now being expressed on social media is a matter of concern. Moreover, this is for many reasons. Some of them being:

- The impact of content articulated aggressively now travels to the offline domain from the online domain very quickly. It has often led to incidents like lynching, riots, and ostracisation of individuals and communities.

- The aggressive content on social media is also now being manipulated to impact individual reputation, communal ties, electoral processes, among other things.

Thus, scholars and researchers need to take the initiative to understand aggression, find methods to identify various sources of aggression, and go one step ahead by strategising ways to control and curb them is the need of the hour. For the very reason, we have initiated this task to identify and understand aggression better. The ambition of common good and contribution to human society through the transformation of novel artificial intelligence approaches into technology platforms delivers actionable insights for societal advancement. The realisation that a deeper understanding of social behaviours through the study of online social data - where we could analyse and generate insights on social behaviours and underlying causalities and make a meaningful contribution to the society was the strongest motivation. It helped explore aggression in social media texts and understand the relationship of various linguistic elements with each other.

1.6 Contributions of this Thesis

The main contributions of the thesis are:

- An attempt to understand the notion of aggression in literature not just relevant to Computer Science, but in a sociolinguistics context as well. We also try to understand other linguistic features on social media like stance, hate speech, and emotions.

- Various models for detecting aggression using Machine Learning algorithms and techniques.

- A unified dataset of Hindi-English code-mixed tweets annotated for multiple dimensions namely:

  - Aggression annotated as covert, overt, non-aggressive
  - Stance annotated as favourable, against, neutral
  - Hate Speech annotated as true, false
  - Figurative language use
    - Sarcasm / Irony / Rhetorical Questions annotated as true, false
    - Puns / Word-play annotated as true, false
    - Metaphors / Similes annotated as true, false
  - Emotion arousal rated from 1 to 5

- Translating the above dataset into English for future work in terms of model building and understanding various linguistic concepts on social media texts.
- An attempt at analysing social media opinion on a political issue across varied modalities. This is required since it helps us in:
  - Analysing the not so apparent forms of verbal aggression displayed on social media.
  - Better understanding linguistic patterns when voicing an opinion and displaying aggression.
  - Analysing social dynamics of opinion.
  - Facilitate classification models that leverage corpora annotated for auxiliary tasks through transfer learning, joint modelling as well as semi-supervised label propagation methods.
- A unified dataset which contains the translation of all the code-mixed tweets to English, which can be used for modelling purposes.
- This thesis was motivated by the need to provide a ground-work for analysis of the nuances of opinion on social media concerning aggression and figurative language use.

1.7 Chapter Organisation

The thesis is divided into five chapters. The introductory chapter portrays the broad contours of social media's impact on society and touches on the brief issues about society with much aggressive content being generated on the fly in the world of social media. The second chapter talks about the background work undertaken to understand the context of the problem at hand. We discuss the various thoughts and ideas behind the concept of aggression. We refer these ideas and techniques throughout the thesis. We also visit the contemporary and historical sources consulted to accomplish this work. This chapter provides a review of the literature covering various facets of the research under different headings: Social media, Aggression, Stance, Hate Speech, Emotion, Opinion, Nuances of Opinion. In all the sections, we try to trace the history and nature of the entities. We also touch upon the contemporary ideas from the machine learning domain to tackle the problem of aggression. In the next chapter, we talk about the methodology: this is where we discuss how machine learning can empower us in tackling the problem of aggression at hand. We also share our results using various machine learning approaches on a standard dataset for aggression. In the fourth chapter, we dive deep into a statistical analysis of the dataset prepared for understanding aggression with various other linguistic features at play, with various examples. We also present our conclusions from this analysis. The conclusions and future work chapter lays down possible areas of further research with product suggestions.

Chapter 2

Literature review and reflections

If I have seen further it is only by standing on the shoulders of giants.

- Sir Isaac Newton, *1675*

There is an extensive range of research available about the various topics covered in the research thesis. We have consulted relevant papers and attempted to develop insights and theories based on the various views and opinions. The current chapter of the thesis provides a review of the literature covering various aspects of the research under four different headings, which are:

1. Social Media
    - Challenges
    - Code Mixing
2. Aggression
    - Aggression on social media
3. Socio-linguistic elements on social media
    - Stance
    - Hate Speech
    - Emotion
4. Nuances of opinion
5. Detecting Aggression in Text

WARNING: This chapter contains examples and words that are offensive in nature.

2.1 Social Media

Social media has dramatically changed the creation, transmission, and perception of information. The general availability, affordability, and ease of use have led to an exponential rise in social media platforms being used for social interactions. Although social media platforms' functionalities are diverse and specialised in facilitating the demands of various demographic segments, the primary cut to chase mechanism is engaging in social discussions with other members on the platform.

Platforms like Twitter, which are inherently fast-paced, are used for rapid dissipation of information through various functions like retweeting and sharing ([14], [15], [16] [17]). Sometimes this information may even spread faster than seismic waves during an earthquake as noted by Sakaki et al [18]. Due to the forced character limit, the emotions expressed in tweets are shallow and intense - add to that the complicated nature of the messages on social media, mining and understanding Twitter data becomes a task in itself. Scholars pick Twitter for the aforementioned reasons, so Twitter data was the first choice when we started to think about this analysis.

The research on language analysis in social media has been increasingly focusing on the impact on our daily lives, both personally and professionally. Natural language processing provides a scientific challenge to develop robust methods and algorithms that extract relevant information from a large volume of data from multiple sources and languages in various formats or free forms.

However, the analysis of social media data comes with various challenges, too. Namely: i) Scale of data ii) Time sensitivity iii) Diversity iv) Unstructured Nature of Data.

2.1.1 Challenges

2.1.1.1 Scale of data

As seen from Figure 1.1, we know that hundreds of millions of people globally adopt online social media platforms. The users on these social media platforms also generate large volumes of new data continuously of various types and context. This phenomenon also results in a high-velocity data stream, which requires specialised approaches that can handle such large volumes of data.

2.1.1.2 Time sensitivity

As pointed by Hu and Liu [19], a lot of social media data streams are bursty, that is to say, that they are not distributed uniformly. The reasoning behind these bursts is mostly due to current events of interest disrupting [18], [20] or capturing the attention of [21] a significant number of users on social media. So it becomes evident that social media data is coupled with the time of its publication, and we should be considerate of its publication and time sensitivity for any analytical task, even. So, our algorithms and systems need to be continuously fueled with data streams as new patterns appear over time.

2.1.1.3 Diversity

Social media data contains diverse linguistic patterns, many discussion topics and differences in the expression of emotion. Eisenstein et al. [22] found that there are distinct linguistic patterns which exist on Twitter among groups of similar geographic proximity as well as similar socio-demographics. This diversity in social data acts as noise on patterns that exist in social data. Hence, our algorithms and systems have to first reduce the noise due to diversity by separating diverse social data into coherent groups which can then be used to extract insights.

2.1.1.4 Unstructured nature of data

Since social media data mainly comprises of unstructured text posted by users in online social media platforms, this data is usually different from professional data (for instance formal documents), in brevity, and use of out-of-vocabulary terms primarily.

Baldwin et al. [23] mention that brevity on social media is achieved by relaxing things grammatically and using shortened forms of terms, sometimes due to restrictions imposed by social media platforms (e.g., tweets, previously had a limit of 140 characters, which has now been increased to 280 characters). This creates a challenge for traditional natural language processing techniques.

Hashtags on social media platforms, especially Twitter, are a great example of terms which represent certain events or are very contextual. (e.g., #IPL2020) Social media users frequently coin such terms during social conversations. Brody and Diakopoulos [24] also pointed out that another category of terms is constructed by users, by repeating certain characters of standard terms (e.g., cooooooolll) to emphasise the expressed emotion. Use of such out-of-vocabulary terms is a challenge, as such terms are often not included in references or thesauruses, often used by the natural language processing techniques to derive certain properties of each word. Filtering out such terms also is not useful since they are tightly coupled to the intended meaning of the post. So the techniques and systems have to be significantly extended to capture different aspects of social data.

2.1.2 Code Mixing

As discussed in the previous chapter, Code-mixing is defined as converting one language to another within the same utterance or in the same oral or written text [25]. This phenomenon is widespread in multilingual societies. Since about 40% of the Indian population speaks Hindi and English as the *lingua franca* of the country, naturally EnglishHindi (or sometimes popularly known as Hinglish) ends up becoming a commonly used code-mixed language pair. The similar phenomenon can be seen on social media platforms.

We aim to curate an English-Hindi code-mixed dataset and perform an experiment of relating various linguistic features.

2.2 Aggression

If we were to build a system for aggression detection, it is imperative to understand aggression as a concept. It is a critical exercise to take on the daunting task of tackling aggression on social media.

A lot of research in Natural Language Processing in recent times has been carried out to recognise several related behaviours automatically:

- Offensive or abusive language ([26], [27])
- Insulting or flaming ([28], [29])
- Trolling on the Internet ([30], [31], [32], [33])
- Cyberbullying ([34], [35], [36], [37])

However, excluding exceptions like [38], and [39], there is barely any theoretical insight into the structure and formation of such behaviours.

All of the elements mentioned above are considered undesirable, aggressive, and detrimental for those on the receiving end. However, besides focusing on these behaviours' initiators' intentions, discussions around the syntactic or pragmatic structure are missing. So, Hardaker [38] defined trolling as an activity which with the following purpose: "*to cause disruption and/or to trigger or exacerbate conflict for the purposes of their own amusement.*" While Nitta et al. [34] defined the act of Cyberbullying as "*humiliating and slandering behaviour towards other people.*" Krol [40] talked about flaming as a way "*to offend someone through e-mail, posting, commenting using insults, swearing and hostile, intense language, trolling, etc*". If we go wholly by the understandings mentioned, it is not that difficult to identify the overlap among these phenomena - as we try to classify actual data in one of these categories, the overlap becomes even more prominent.

Many studies have been done to comprehend and identify aggression in various contexts: some of these works target different online platforms like Twitter [41], Wikipedia [42], and ask.fm [43]. Aggressive language detection is traditionally tackled as a regular text classification task. It is often approached with surface features such as token frequencies, text characteristics, linguistic features, and word embeddings ([44], [45], [46]). A lot of work employs statistical machine learning algorithms for the tasks ([45], [42]).

However, rarely tasks have been done in the Indian context and scenario with data relevant to our society. Much research has been carried out on studying these phenomena in isolation and their computational processing; thus making us realise the theoretical gap in understanding these phenomena' interrelationship. As a concept, understanding what aggression is a crucial exercise to take on the daunting task of tackling aggression on social media.

Since aggression is so difficult to define, social psychologists have spent a considerable amount of time trying to determine what should and should not be considered aggression. Social psychologists define aggression as behaviour intended to harm another individual who does not wish to be harmed

[47]. This, on its own, introduces the perception of intent - so, what may look like aggression from one point of view may or may not look that way from another, and the same harmful behaviour may or may not be aggressive depending on its intent. For example, would we consider a dentist who might intentionally give a patient a painful injection of a painkiller as doing an aggressive task, despite the goal being preventing further pain during the procedure?

Social psychologists agree that aggression can be verbal as well as physical. Non-physical aggression includes verbal aggression (yelling, screaming, swearing, and name-calling) and relational or social aggression, which is defined as intentionally harming another person's social relationships, for instance, by gossiping about another person, excluding others from our friendship, or giving others the "*silent treatment*" as mentioned by Crick and Grotpeter [48]. Nonverbal aggression also occurs in sexual, racial, and homophobic jokes and epithets designed to harm individuals.

With the increased number of people being active on social media and voicing their opinions on anything and everything - it was only a matter of time when people were going to recognise that the aggressive behaviour was going to be reflected in the online world, too. Many studies [42], [41] have recently been done to understand and identify aggression in various contexts on social media platforms to encourage other research groups to contribute to aggression identification in these sources.

2.3 Socio-linguistic elements on social media

2.3.1 Stance

Biber and Finegan [49] define stance as expressing the speaker's attitude, standpoint, and judgment toward a proposition. While Du Bois [2] argues that stance-taking "*is a subjective and inter-subjective phenomenon in which the stance-taking process is affected by personal opinions and non-personal factors such as cultural norms*." Taking a stance is not a simple process: it involves understanding cultural, social, and personal entities. Furthermore, McKendrick and Webb [50] thought that the political stance-taking process depends on experiential behaviour. We are still learning about the dynamic, structure, and role of the language and social interaction in understanding a user's stance on a given issue. Suppose we think about the issue of stance from a sociolinguistics point of view, where the main concern is understanding the writer's viewpoint through their text. The core idea is to link stance to multiple factors, the major ones being: linguistic acts, social interactions, and individual identity. Using the linguistic features in detecting a stance is usually associated with adjectives, adverbs, and lexical items as pointed by Jaffe et al[51].

Another interesting point of view about defining stance can be observed in Due Bois's stance triangle shown in Figure 2.1. He mentions that taking a stance, as a process is based on three factors, which are:

- Evaluating objects
- Positioning subject (the self)

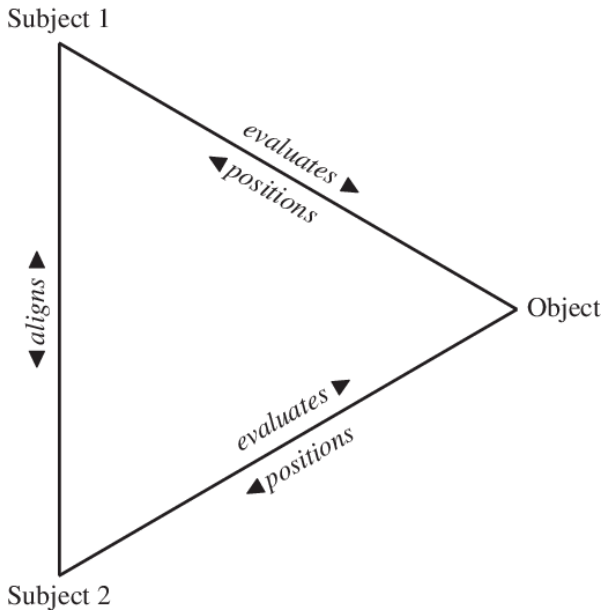

Figure 2.1 The Stance Triangle, adopted from Du Bois [2].

- Aligning with other subjects

For example, "I am with the new legalisation on abortion" has a subject, "self", indicated by the proposition "I" and further, the "with" indicates the favoured position toward the object "abortion."

The last couple of years where social media dominance has been on the rise has seen a rise in Twitter becoming the first choice to study the expressed stance towards various events or topics ([52], [53], [54], [55]). People have been getting more and more reliant on social media platforms as the primary source of their news to connect with the world and get instant updates [56]. Individuals get to explore various aspects of emerging topics, express their points of view, get instant feedback, and explore the public's views. It helps in understanding the public stance toward various social and political aspects.

So, measuring public opinion on social media, particularly on political and social issues, becomes an active use-case to understand and solve the problem of identifying stance. The nature of these issues is usually controversial, wherein people express opposing opinions toward differentiable points. Social issues such as feminism, climate change, abortion, legalisation of marijuana, among others, have been heavily used as target topics for stance detection on social media as pointed out by Mohammad et al. [57]. Similarly, political topics, such as referendums and elections, have always been hot topics used to detect a stance to study public opinion [58].

However, people have a misconception at times and tend to use sentiment and stance interchangeably. For understanding the relationship dynamic between sentiment and stance, several studies have demonstrated that it is insufficient to use the sentiment as the only dependent factor to interpret a user's stance ([59], [60], [61], [62]). This happens mostly because of the complexity of interpreting stances from a given text, as they are not always directly aligned with a given post or a tweet's polarity. Let

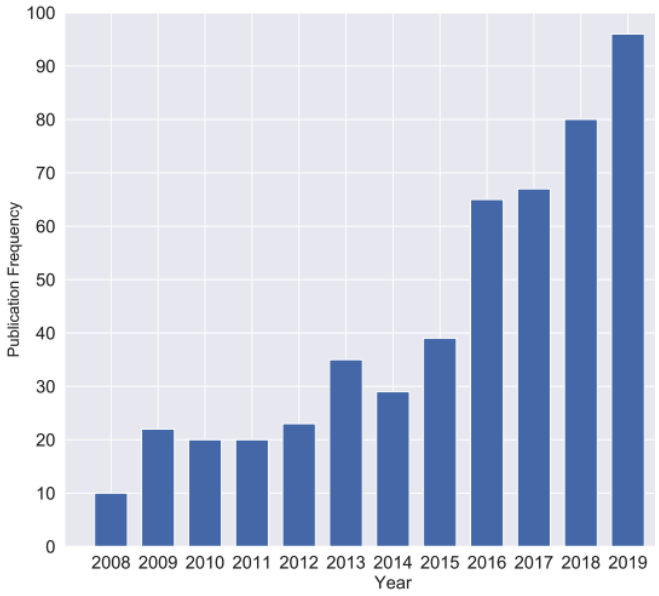

Figure 2.2 Publications per year on stance detection as searched on Web of Science. Keywords used: "stance prediction", "stance detection" and "stance classification", adopted from AlDayel and Magdy [3].

us look at this with some example tweets to illustrate the orthogonal relationship between sentiment polarity and stance.

1. **Tweet:** Let us be honest, THE BIGGEST terror threat in the World is climate change #climate-change #drought #floods

   Target: Climate change is the real concern.

   Sentiment: Negative.

   Stance: In favour of climate change.

   For instance, in the above example, we can see that a negative sentiment can be detected without a supportive stance towards the target.

2. **Tweet:** Abortion does not compute with my thought process. Life is sacred on all levels, after-all. #abortion

   Target: Legalizing abortion.

   Sentiment: Neutral.

   Stance: Against abortion.

Furthermore, a text may not indicate any sentiment and still pose a clear stance toward a target as seen from the above example. This is to mention that both phenomena, that is, sentiment and stance, are orthogonal in nature.

2.3.2 Hate Speech

We do not have a standard international agreed-upon definition of the concept of hate speech, but many definitions exist in parallel. Before we understand the relationship of hate speech with Twitter, let us understand how people look at hate speech.

The European Court of Human Rights adopted a definition on hate speech as "*all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility towards minorities, migrants and people of immigrant origin*" [63].

Nielsen [64] mentions that there is a lack of consensus with regards to understanding what should be the content of the hate speech. He suggests that hate speech can be motivated by all kinds of perceived differences. Leets [65] states that hate speech violates the dignity of an individual, causing distress, humiliation, emotional or psychological pain in the process. Downs and Cowan[66] mention in their research that hate speech is no less than a strong weapon. It harms individuals by terrorising, humiliating, wounding and degrading them. According to Simpson [67]:

Hate speech is a term of art in legal and political theory that is used to refer to verbal conduct – and other symbolic, communicative action – which willfully expresses intense antipathy towards some group or towards an individual on the basis of membership in some group [. . .] Hate speech thus includes things like identity-prejudicial abuse and harassment, certain uses of slurs and epithets, some extremist political and religious speech (e.g. statements to the effect that all Muslims are terrorists, or that gay people are second-class human beings), and certain displays of hate symbols (e.g. swastikas or burning crosses).

There is no exact "standard" international definition of the hate speech concept, but several definitions exist in parallel - all of them equally valid, and relevant in their own context and ways.

However, the reason why Twitter and other social media platforms are criticised for expansion of hate speech are many. Some of them mainly being:

- Given that the Internet allows absolute freedom of expression, on all social media platforms, and the difficulties with censorship add to the woes of hate speech.
- The breadth of the message's reach between various social media platforms; followed by the message's adoption, circulation, and repetition. It fundamentally boils down to the fact, that the number of opportunities for people to see and spread a hate message, or a possible target to come across that message increase.

- There is a particular cloak of anonymity available in the online world, which helps people conceal their real identities and creates a problem of fake accounts spewing venom in some cases.
- A lot of the Internet is still uncontrolled and unregulated. Even though a lot of gatekeeping is being added so that users could be stopped from getting corrupted, it requires a culture change and stricter policies from the social media websites to curb the running issue of hate speech.

Burnap and Williams [68] defined hate speech as responses that include written expressions of hateful and antagonistic sentiment toward a particular race, ethnicity, or religion. They also used a binary classification scheme of hate speech vs non-hate speech, which was also followed by Bohra et al. [69] for their dataset on Hindi-English code-mixed tweets. Malmasi and Zampieri [70] used a 3-way classification scheme between hate speech vs offensive language but not hate speech vs no offensive language. As aggression levels are highly predictive of offensive language but not hate speech category, we used a binary classification speech. An example:

1. **Tweet**: '*ab itni taklif hai to atmadaah kyo nahi kar lete notebandi k khilf. Delhi walo ko bhi mukti milegi tumse*'

   Translation: '*If you have such a huge issue with it, why don't you perform a self-immolation? The people of Delhi would also get freedom from you* '

   In tweet 1, the author refers to Arvind Kejriwal, the leader of opposition party AAP and the Chief Minister of New Delhi (capital of India). The author suggests that Kejriwal should kill himself to free the residents of Delhi. In supporting the decision of Demonetisation, the author of the tweet is making extreme and graphic suggestions towards one of the main opponents of the target issue.

2.3.3 Emotion

Emotions are one of the most crucial aspects of human social life and social behaviour. We define emotions as a complex state of mind that influence the thought process and behaviour in psychology. Not only human social interactions, but emotions also play an integral part in interpersonal communication. Thus, it becomes essential to understand the emotions and the methodologies to capture them to gain invaluable insights about the conversations which happen on social media platforms, all the time. Moreover, those insights also help us figure out the opinions / feelings of individuals towards different topics.

If we were to understand the critical theories of emotion, we could say that they could be categorised into two categories primarily: cognitive and physiological. Cognitive theories mention that emotions occur due to conscious cognitive activities such as evaluating things or people, judgments or mere

thoughts. Lazarus [4] mentions as *Lazarus cognitive theory* that emotions are determined by the cognitive appraisal of an event or stimuli in the environment. It argues that the cognitive process controls the quality and intensity of the emotions. Since the cognitive appraisal process is personalised for people, the same event often yields different emotional responses from different individuals depending on their experience.

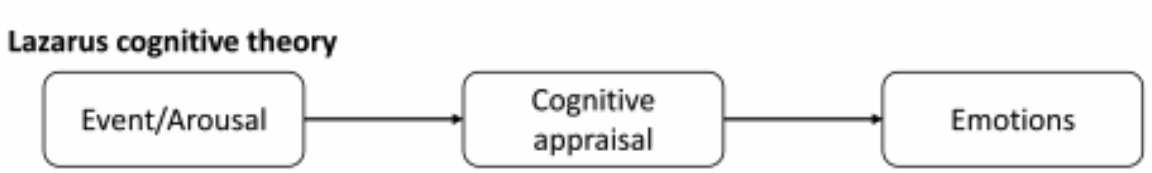

Figure 2.3 Lazarus Cognitive Theory, adopted from Lazarus [4].

If we look at the physiological theories of emotion, all of them suggest that emotions are due to physical changes in the body. For instance, Cannon [71] states that emotions are conscious feelings about bodily changes and nothing more. Schachter and Singe [5] argued in their *Two-factor theory* that a two-factor process generates emotions. Firstly, a physiological event happens. Then, an individual refers to her experience or the immediate environment to find emotional cues to provide an emotion label to that event.

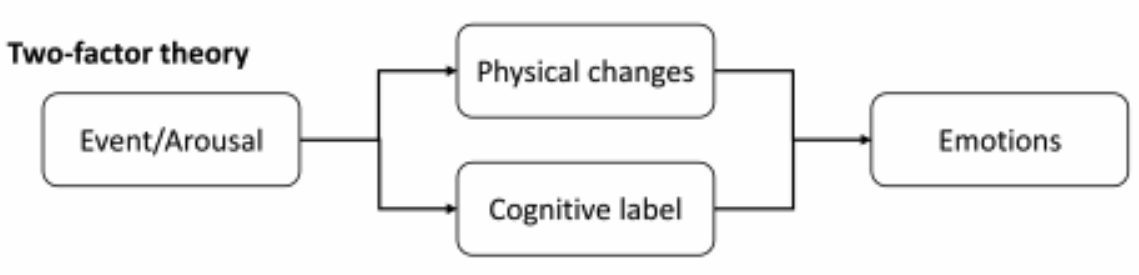

Figure 2.4 Two-Factor Theory, adopted from Schachter and Singer [5].

Further, the next problem the researchers ran into was how to model the emotions with data at hand. Since emotions are a cognitive force at hand, it becomes hard to measure them as an entity. Also, since emotions, in general, are highly personalised to individuals based on their private experiences, the same event could yield a different emotion on an individual. The consensus on modelling emotions have two major approaches: Discrete models, which try to model all emotions into several basic categories: These stem from the hypothesis of a universal set of raw emotions in which researchers showed an array of photos of different people with varied facial expressions and asked people to recognise them. Dimensional models, which try to model emotions as a continuous n-dimensional space: They argue that emotions are often overlapping states generated by a standard neurophysiological system [6]. So, all emotions can be represented in a conceptual continuous n-dimensional space.

An example of a dimensional model, the Circumplex model, Russell [72] states that all emotions originate from two systems: valence (pleasant to unpleasant continuum) and arousal (active to passive continuum) and that every single emotion can be represented as a combination of valence and arousal. So, for example, elated is a pleasant and active emotion, while serene is a pleasant and passive emotion.

Figure 2.5 A graphical representation of the Circumplex model consisting of valence and arousal dimensions. Multiple emotions are placed in the axis system based on their valance and arousal. Image adapted from Posner et al. [6]

For our purpose, while annotating, we decided to restrict our scope to analyse only for emotion arousal level, as emotion valence level is analogous to sentiment.

2.4 Nuances of Opinion

To understand the notion of an opinion, we first need to understand that opinions are beliefs or judgments formed about something that may or may not be based on fact or knowledge. The general notion of an opinion is that it refers to a subjective belief. When people form their personal convictions about certain things, they almost always end up interpreting factual evidence through the filter of our past experiences, feelings, values and emotions.

Opinions are different from facts since facts are verifiable. If we can safely assume that the memory, history or the measuring devices linked to the fact are correct, the truth is beyond argument. An opinion is a judgment which is based on facts. At best, it is an honest attempt to come up with a reasonable conclusion from facts and evidence. But that also means that an opinion is potentially changeable - depending on how the evidence is interpreted.

People often express their opinion on various issues on social media forums like Twitter and often display aggression towards people that support a contradicting belief or towards a particular group of stakeholders on the issue. We can find opinions on almost anything. They range from socio-political issues to movies, to sports to literally anything. In the research community, the data of this kind is known as Big Data. It is characterised by 3V which stands for Volume, Variety and Velocity. Some also refer to it as 5V, i.e. for Value and Veracity, as mentioned by Guellil and Boukhalfa [73].

We need this high-velocity data since it provides deeper insights into public opinion. It also reduces the need for detailed surveys and polls. But on social media networks, users express their opinions in a voluntary fashion. So, despite the inherent selection bias involved, it provides a useful insight into public opinion as seen in existing work ([74], [75]).

2.5 Detecting Aggression in Text

We have seen from the previous sections in this chapter that identifying aggression in social media is closely related to cyberbullying, hate speech, and offensive and abusive language identification. In this section of this chapter, we try to briefly discuss the relevant papers published on the same, which would help us understand aggression in a much better light computationally. In the past, multiple studies have been done which use various computational methods to detect aggression. However, the thing to note here is that majority of the work in this domain has been published in English [76] and a few other languages, but we already know and understand that social media abuse and aggression is independent of language and demographics.

Davidson et al. [77] created a dataset for abusive language identification and categorised the tweets using Naive Bayes, Logistic Regression, Decision Trees, Random Forests and Support Vector Machines (SVM) to classify the tweets in three classes: offensive, hate or neither (neither hate nor offensive).

Some recent works like ([78], [79], [80], [81], [82]) tried to solve issues related to aggression on the Internet. The works by Chatzakou et al. [78] and Chen et al. [79] are focused on Twitter, and standard English text for aggression detection, which is not equally applicable to multilingual cases and for other social media platforms. Chatzakou et al. [78] found improved accuracy after combining user and network-based features with text-based features.

Some researchers of TRAC - 1 shared task ([82], [83], [80], [81]) worked on the challenges mentioned above and achieved limited success. Some participants tried ensemble learning methods with various machine learning classifiers and many deep learning models. Some other group of researchers ([82], [84]) applied data augmentation with the help of machine translation using different languages (Hindi, French, German and Spanish) by preserving the meaning of comments with different wording and found better results for such enlarged dataset.

Schmidt and Wiegand [45] and Mishra et al. [85] present recent overviews of related work on the detection of abusive language. Schmidt and Wiegand [45] present a survey on hate speech detection using Natural Language Processing. They mention how supervised learning approaches are predomi-

nantly used for such tasks, and Support vector machines (SVM) and recurrent neural networks are the most widespread. They also talk about how simple surface features like bag-of-words, n-grams etc., are widely used for hate speech detection. Other techniques like word generalisation like word embedding or knowledge-based features like various ontologies are also mentioned. [85] in their survey also gives us an overview of the various datasets that are annotated for abuse, while also mentioning various automated abuse detection methods.

Even within various shared tasks, participants used various machine learning techniques that achieved good performances. This is the case for GermEval [86], SemEval-2019 Task 6 [87] and TRAC [88]. GermEval [86] is a shared task that focuses on detecting offensive language on German tweets. SemEval-2019 Task 6 [87] is a shared task that focused on identifying and classification offensive language in social media, more precisely on English tweets. Finally, TRAC ([88]) is a shared task that focuses on aggression identification considering both English and Hindi languages. The objective is to classify texts into three classes: Non-Aggressive (NAG), Covertly Aggressive (CAG), and Overtly Aggressive (OAG). Facebook posts and comments are provided for training and validation, while, for testing, two different sets, one from Facebook and one from Twitter, were provided. The best performance during the shared task was achieved with deep learning approaches, whether on Facebook test set or Twitter test set. During this shared task, most participants considered classical machine learning methods (e.g. Random Forests) based on features as in ([89], [90], [82]). Transformer based models like BERT [91], a task-agnostic language representation model, consisting of multiple layers of bidirectional transformers have also been considered for the task of aggression identification. The training objective of these models uses a masking technique. Given a sentence, some percent of the input tokens are masked, and the task is to predict these tokens. This technique overcomes unidirectional processing limitation and is also superior to language models that combine right-to-left and left-to-right processing, as mentioned in Peters et al [92]. BERT has also been attempted in various shared tasks on offensive language detection or hate speech ([93], [94]). The results demonstrated from TRAC-1 that it was notoriously hard to distinguish between overt and covert aggression in social media, especially in a code-mixed dataset. This then became the key motivating factor for us to attempt this shared task.

Chapter 3

Modeling Aggression

No event can be judged outside of the era and the circumstances in which it took place.

- Fidel Castro, *1953*

3.1 Background

Due to the recent emergence of various social media platforms like Twitter, Facebook, and Reddit, and social networking tools, the availability of user-generated content on the Internet has increased manifolds. It results in a positive exchange of ideas, but it also leads to extensive dissemination of aggressive and harmful content on the Internet.

These incidents not only cause mental and psychological agony to the users on the Internet, but it also forces people to delete or deactivate their social media accounts, and in some extreme cases, take measures of self-harm [95]. As Culpeper [96] pointed out, social media aggression is targeted to a particular person or group to damage the identity and lower their status and prestige. Therefore, it is crucial that preventive measures need to be taken to cope with abusive behaviour aggression online.

In this chapter, we present the different machine learning models we built on the shared task - Shared Task on Aggression Identification organised as part of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC - 1) at COLING 2018[1], to distinguish different levels of aggression over multiple datasets from Facebook and other social media that cover both English and Hindi texts. The number of teams which participated in this task was greater than 100. The best system had obtained a weighted F-score of **0.64** for both Hindi and English on the Facebook test sets. While the best scores on Twitter, were **0.60** and **0.50** for English and Hindi, respectively. We share the details of these best systems later in the chapter. Furthermore, we shared our findings and perceived standing among the submissions.

[1] `https://sites.google.com/view/trac1/shared-task`

3.2 Dataset

For the task mentioned above, all the teams were provided with a dataset which was annotated on aggression. This contained *15,000* Facebook Posts and Comments - in both, English and Hindi for training and validation. The Hindi dataset had posts both in Roman and Devanagari script. For testing purposes, two different sets - one from Facebook and another from Twitter (this had *3,000* comments) - were provided. The data was annotated with three levels of aggression:

- **Overtly Aggressive** (OAG): this represents human behaviour which is intended to harm another verbally, physically and psychologically. It can be summarised as follows:
  - Aggression shown openly with verbal attack directly pointed towards any group or individuals.
  - Attack commenced using abusive language, words or name-calling or comparing in a derogatory manner.
  - By supporting false attacks or supporting others' hateful comments.
  - Sometimes, these texts also contain indirect references, for which context is needed.
- **Covertly Aggressive** (CAG): behaviour where aggression is generally hidden. It usually contains sarcastic, negative emotions due to its indirect nature. It can be summarised as follows:
  - By using metaphorical words to attack an individual, nation, religion.
  - Praising someone by criticising group irrespective of being right or wrong.
  - Sometimes, these texts also contain direct references.
- **Non-Aggressive** (NAG): Most of the times, the statements which fall in this category lack the intention to be aggressive. They are mostly used while referring to the correct facts or supporting individuals or groups on social issues.

Class	Sentence
Overtly Aggressive	1. we support his speech... RSS is dvding the country, they do not want to stay peace
	2. We want to get rid of all u Indians.....why don't u hear our loud cries
	3. The United States Government of Mr. Donald Trump will do nothing
Covertly Aggressive	1. Modi ji, why are you giving all the Pain & no Real Gain
	2. udhav has shown bjp its place,bravo shivsena
	3.Reservation is like another form of terrorism
Non-Aggressive	1. Sorry sir I forgot.
	2. When is work on NH-8 getting completed? Particluarly Hero Honda Chowk??
	3. I will upgrade from my 220 CC to 400CC

Table 3.1: Some examples from the dataset.

Figure 3.1 Devanagari Examples of Overtly Aggressive.

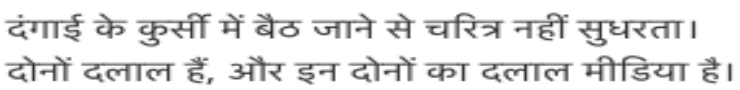

Figure 3.2 Devanagari Examples of Covertly Aggressive.

Table 3.1 contains some examples of Overtly Aggressive, Covertly Aggressive, and Non-aggressive posts, from the dataset [88].

We also saw Devanagari content for all three categories. Examples for the same are in Figures 3.1, 3.2, 3.3. In Table 3.2 and Table 3.3, we see the details of the dataset. we decided to find the split between the three categories for the entire dataset. Furthermore, in Table 3.4 and Table 3.5, we see the split of the data (training and testing) among the three mentioned categories.

3.2.1 Issues with the dataset

- **The annotation issue**: While working on this dataset, several instances of *supposedly* incorrect annotation were found. Some of the annotations looked highly implausible. This is when there is agreement on the fact that aggression is a highly subjective phenomenon, and different annotators may have different judgments about the same comment. Thus, the dataset needs further scrutiny and validation.

- **The language issue**: We observed in the English dataset, that some statements contained code-mixed Hindi-English data. Some statements also contained data from other languages like German. They had to be explicitly handled and needed to be filtered out.

3.3 Data Cleaning

As a general trend, the data from all the social media channels is usually noisy and contains a lot of grammatical and syntactical errors. It also contains ad-hoc spellings making it challenging to analyse. Thus, the first step to tackle this problem was to clean and prepare the data as much as possible, before

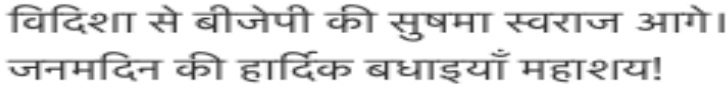

Figure 3.3 Devanagari Examples of Non-Aggressive.

English Data	
Train	11999
Dev	3001
Test (Twitter)	916
Test (Facebook)	1257

Table 3.2: Statistics for English Data

Hindi Data	
Train	12000
Dev	3001
Test (Twitter)	970
Test (Facebook)	1194

Table 3.3: Statistics for Hindi Data

feeding it to our systems. We applied two stages of cleaning. In the first stage, we used the ekphrasis toolkit. The toolkit was created as a part of the text processing pipeline for DataStories team's submission for SemEval-2017 Task 4 (English), Sentiment Analysis in Twitter [97].

This helped us in doing the following:

- Replacing \n with a space.
- Normalising URLs, emails, percentages, money, phone, user, time, date, and number with tags like "url".
- Unpacking contractions (can't - can not)
- Unpacking hashtags (#ShutDownJNU - Shut down JNU)
- De-emojizing (Changing the thumbs up emoji to a textual form - :thumbs_up:)
- Spell correction: You can replace a misspelt word, with the most probable candidate word.
- Lower-casing
- Drop Punctuation
- Social Tokenization: A custom text tokeniser which was geared towards social networks (Facebook, Twitter, etc.). This is curated so that it understands complex emoticons that are regularly used on social media platforms, emojis and other unstructured expressions like timezones, dates, times and more.

Class	Train	Dev	Test (Facebook)	Test (Twitter)
OAG	2708	711	144	361
CAG	4240	1057	142	413
NAG	5051	1233	630	483

Table 3.4: English Data Split

Class	Train	Dev	Test (Facebook)	Test (Twitter)
OAG	4856	1217	362	459
CAG	4869	1246	413	381
NAG	2275	538	195	354

Table 3.5: Hindi Data Split

For the English dataset, using WordNet Lemmatization, which is essentially the process of clubbing together various inflected forms of a word. The idea is to analyse these as a single item. It is similar to stemming, but the crucial difference is that it brings context to the words. Precisely why it links words with similar meaning to one word.

Note: The ekphrasis toolkit works only on English text. So, for doing the things mentioned above in Hindi, we implemented custom RegEx patterns for our in-house processing to replicate the behaviour of the ekphrasis toolkit for our purpose.

The following examples show the distinction from the original text to the final cleaned text.

1. **Original Text 1**: Well said sonu..you have courage to stand against dadagiri of Muslims

   Basic Cleaned Text 1: well said sonu you have courage to stand against dadagiri of muslims

   Language Cleaned Text 1: well said sonu you have courage to stand against dadagiri of muslim

2. **Original Text 2**: Most of Private Banks ATM's Like HDFC, ICICI etc are out of cash. Only Public sector bank's ATM working

   Basic Cleaned Text 1: most of private banks atm s like hdfc icici etc are out of cash only public sector bank s atm working

   Language Cleaned Text 1: most of private bank atm s like hdfc icici etc are out of cash only public sector bank s atm working

3. **Original Text 3**: Wondering why Educated Ambassador is struggling to pay through Credit/Debit at a Decent Restaurant! Cant imagine that diplomat of a Developed nation is not having a Card and he needs Cash only for Dinner.

   Basic Cleaned Text 3: wondering why educated ambassador is struggling to pay through credit debit at a decent restaurant cant imagine that diplomat of a developed nation is not having a card and he needs cash only for dinner

   Language Cleaned Text 3: wondering why educated ambassador is struggling to pay through credit debit at a decent restaurant cant imagine that diplomat of a developed nation is not having a card and he need cash only for dinner

4. **Original Text 4**: How does inflation react to all the after shocks of this demon...?

   Basic Cleaned Text 4: how does inflation react to all the after shocks of this demon

   Language Cleaned Text 4: how does inflation react to all the after shock of this demon

3.3.1 Social Tokenization

Social tokenisation is a difficult problem to tackle because one needs to avoid splitting words or expressions that should be kept as one token, that is, intact. This becomes even more important when dealing with text from social networks where the text is a lot more "creative" with elements like emoticons and hashtags, and so on. The ekphrasis tokeniser can identify almost all emoticons, emojis and many complex expressions. Especially for tasks where aggression is involved, many expressions play a decisive role in identifying the text's sentiment. Expressions like these are:

- Censored words such as f**k, c**t, s**t
- Words with emphasis, such as a *GREAT* time, I don't *think* so
- ASCII Emoticons
- Dash-separated words, such as over-consumption, anti-american, mind-blowing

Moreover, ekphrasis can identify information-bearing expressions. This helps in keeping preserve / extract them as one token (IR). Expressions like these are:

- Dates, such as Mar 13th, December 26, 1992, December 26-2016, 10/13/92, 13 December 2016, April 24, 1973, 11.15.16, November 24th 2016, January 21st.
- Times, such as 5:43 pm, 11:13 AM, 2:44 pm, 5:30.
- Currencies such as $213M, $73.000,

- URLs, such as http://www.cs.unipi.gr, https://www.youtube.com/watch?v=yjfNXi2EtA
- Phone numbers.

3.4 Methodology

3.4.1 Models

We decided to explore classical machine learning models for our problem of identifying aggression for the TRAC-1 dataset. The reasoning behind choosing traditional machine learning models was that the TRAC-1 dataset for both Hindi and English had content available in the order of thousands only, and limited quantity. So, the selection of models meant looking at traditional models, since the amount of data compared to other deep learning models required is not available to us. Deep Learning models are built on word embeddings, and since the dataset had a significant amount of code-mixed content, we were more inclined towards traditional methods. This necessarily forced us to ensure that our data cleaning and processing would be of the highest standard.

Some of the models which we considered and explored for which we share our results are the following:

- Naive Bayes
  - Multinomial Naive Bayes
  - Bernoulli Naive Bayes
- Linear SVM

Naive Bayes

Naive Bayes algorithms are supervised learning algorithms based on applying the Bayes' theorem with a "naive" assumption - the conditional independence between every given pair of features given the value of the class variable. The Bayes' theorem states the following relationship, given class variable y and dependent feature vector x_1 through x_n:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)} \tag{3.1}$$

Using the naive conditional independence assumption that:

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y), \tag{3.2}$$

for all i, this relationship ends up getting simplified to:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)} \tag{3.3}$$

Since $P(x_1, \ldots, x_n)$ is constant given the input, we can use the following classification rule:

$$\begin{aligned} P(y \mid x_1, \ldots, x_n) &\propto P(y) \prod_{i=1}^{n} P(x_i \mid y) \\ &\Downarrow \\ \hat{y} &= \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y), \end{aligned} \tag{3.4}$$

and Maximum A Posteriori can be used for estimating $P(y)$ and $P(x_i \mid y)$. The former then ends up being the relative frequency of class in the training data.

Every naive Bayes classifier differs mainly by its assumptions regarding the distribution of $P(x_i \mid y)$.

As Zhang [98] mentioned, that even though Naive Bayes classifiers have over-simplified assumptions, they work well in many real-world situations. Some of them being spam filtering and document classification. They also do not require a tremendous amount of training data to estimate the necessary parameters.

Another advantage of Naive Bayes classifiers and learners is that they are swift compared to more sophisticated machine learning methods. We can also take care of the problems which stem from the curse of dimensionality since the decoupling of the class conditional feature distributions essentially means that each distribution can be independently estimated as a one-dimensional distribution.

*Bernoulli Naive Bayes Bernoulli Naive Bayes implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions. It means that there may be many features, but every one of them is assumed to be a binary-valued (Bernoulli, boolean) variable. So, this particular class requires samples to be represented as binary-valued feature vectors.

The decision rule for Bernoulli naive Bayes is based on:

$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i) \tag{3.5}$$

Which differs from multinomial naive Bayes' rule. It explicitly castigates the non-occurrence of a feature i that is an indicator for class y; while in the multinomial variant, we would simply ignore a non-occurring feature.

In the case of a text classification problem, word occurrence vectors may be used to train and use this classifier. Bernoulli Naive Bayes might perform better on some datasets, especially those with shorter documents.

Multinomial Naive Bayes

Multinomial Naive Bayes implements the naive Bayes algorithm for multinomially distributed data. The distribution is parametrized by vectors $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$ for each class y where n is the number of features (in text classification, the size of the vocabulary) and θ_{yi} is the probability $P(x_i \mid y)$ of feature i appearing in a sample belonging to class y.

The parameters θ_y is estimated by a smoothed version of maximum likelihood, i.e. relative frequency counting:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \tag{3.6}$$

where $N_{yi} = \sum_{x \in T} x_i$ is the number of times feature i appears in a sample of class y in the training set T, and $N_y = \sum_{i=1}^{n} N_{yi}$ is the total count of all features for class y.

The smoothing priors $\alpha \geq 0$ accounts for features not present in the learning samples and prevents zero probabilities in further computations. Setting $\alpha = 1$ is called Laplace smoothing, while $\alpha < 1$ is called Lidstone smoothing.

SVM

Support Vector Machines are supervised machine learning classification models which discriminate data by constructing a hyperplane. In other words, SVM constructs a hyperplane in multidimensional space to separate different classes. It generates optimal hyperplane in an iterative manner, which is used to minimise an error. The core idea of SVM is to find a maximum marginal hyperplane that best divides the dataset into classes. Figure 3.4 shows how blue nodes are separated from red ones by a hyperplane.

A feature set of size n is used to represent the data set in n-dimensional space. Accordingly, for data represented in the $n-$dimensional space, we are looking for a hyperplane in $n-1$ dimension, which can classify the data. This is called a linear classifier whereby that hyperplane is chosen whose distance from both classes' nearest data point is maximised. Such a hyperplane is called maximum margin hyperplane, and the training data points that are closest to the hyperplane are called support vectors.

We can write any hyperplane as the set of points $\vec{x}$ satisfying the following:

$$\vec{w}.\vec{x} - b = 0 \tag{3.7}$$

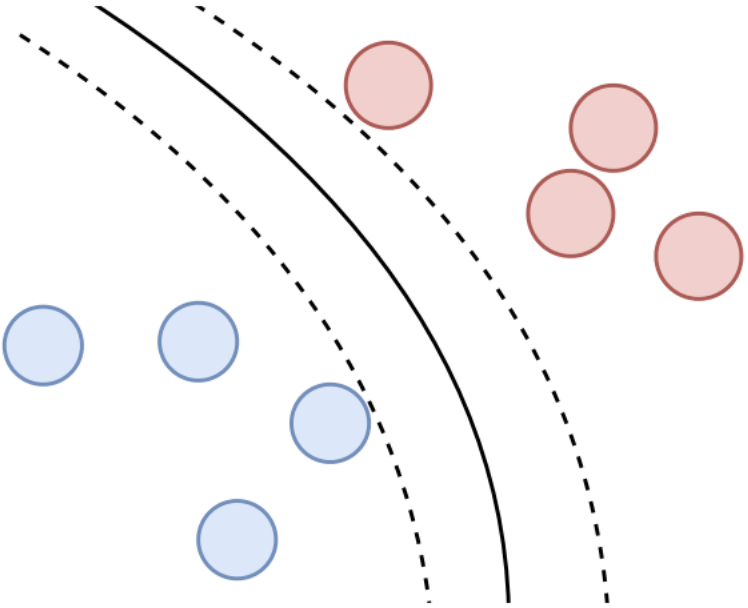

Figure 3.4 The SVM classifier outputs a line (solid) separating the red and blue circles present in two dimensional space. The functional margin shown in dotted line is the largest distance of the classifier to the nearest training data point. A larger margin implies a more robust classifier.

where $\vec{w}$ is a normal vector to the hyperplane. If the training data is linearly separable, we can select two separating hyperplanes so that the distance between the two data classes is maximum. The following equations can represent these two hyperplanes:

$$\vec{w}.\vec{x} - b = 1 \tag{3.8}$$

and

$$\vec{w}.\vec{x} - b = -1 \tag{3.9}$$

Anything on or above the boundary given by Equation 3.8 belongs to one class, represented by label 1, and anything on or above the boundary given by Equation 3.9 belongs to the other class, represented by label -1.

The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Still effective in cases where the number of dimensions is greater than the number of samples.

- Different Kernel functions can be specified for the decision function - making them versatile.

The disadvantages of support vector machines include:

- If the number of features is much greater than the number of samples, avoiding over-fitting in choosing Kernel functions and regularisation term is crucial.
- They do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

3.4.2 Modeling

The core idea we were operating on was that instead of using complicated features, we wanted to rely on our understanding of the dataset and our cleaning prowess. Thus, the feature we considered for our models was a bag of n-grams. Essentially, bag of words and bag of characters. This then meant that we had different models with two combinations: one for character, and word.

- Bernoulli Naive Bayes with bag of words and character
- Multinomial Naive Bayes with bag of words and character
- Linear SVM with with bag of words and character

We used SciKit Learn [99] for our modeling purposes. Further, the range *max_df* and *ngram_range* of params selected, for each of these models was:

- Word:
  - max_df: [0.5, 0.7, 1.0]
  - ngram_range: [(1, 1), (1, 2)]
- Char:
  - max_df: [0.5, 0.7, 1.0]
  - ngram_range: [(3, 6)]

Where n_gram range denotes the lower and the upper boundary of the range of n-values for different n-grams to be extracted. All values of n such such that $min_n <= n <= max_n$ will be used. While max_df is used for removing terms that appear too frequently, also known as "corpus-specific stop words". So, for instance, when the value of max_df is 0.50, it means that the model would ignore terms that appear in more than 50% of the text. Then, in the next step where we used Hyper-parameter tuning to optimise the hyper-parameters of our model. For finding out the best params, we trained our models on the training data and then validated with the validation dataset. We essentially used a train-test split,

and we found the best results on the validation set and then retrained the entire model on the best params on train and validation.

The best params for all the models can be seen in Tables 3.6, 3.7 and 3.8. The validation scores for all the models can be seen in 3.9. The F1 scores for all the models on the English dataset can be seen in in 3.10. The final results were computed on held out test set.

Params	**Word**	**Char**
clf__alpha	0.1	0.1
vect__max_df	0.5	0.5
vect__ngram_range	(1, 1)	(3, 6)

Table 3.6: Bernoulli Naive Bayes' best params for English

MultinomialNB	**Word**	**Char**
clf__alpha	0.01	0.01
tfidf__use_idf	False	False
vect__max_df	0.5	0.7
vect__ngram_range	(1, 2)	(3, 6)

Table 3.7: Multinomial Naive Bayes' best params for English

LinearSVM	**Word**	**Char**
clf__alpha	0.0001	0.0001
clf__max_iter	20	50
vect__max_df	0.5	0.5
vect__ngram_range	(1, 2)	(3, 6)

Table 3.8: Linear SVM's best params for English

Similarly, for Hindi, the best params for each of the models are shared in Tables 3.11, 3.12, and 3.13. We can see that depending on the language, the choices and params are also updated - that is to say, different language and different datasets pose a different problem. Further, the validation scores for all the models on the Hindi dataset can be seen in Table 3.14. The F1 scores for all the models on the Hindi dataset can be seen in Table 3.15.

Model	Value Score
BernoulliNB - Word	0.548
BernoulliNB - Char	0.542
MultinomialNB - Word	0.566
MultinomialNB - Char	0.591
LinearSVM - Word	0.576
LinearSVM - Char	0.584

Table 3.9: Validation Scores for Models on English dataset

Model	Features	Facebook Data	Twitter Data
BernoulliNB	Word	0.576	0.579
MultinomialNB	Word	0.563	0.570
LinearSVM	Word	0.628	0.583
BernoulliNB	Char	0.567	0.591
MultinomialNB	Char	0.580	0.581
LinearSVM	Char	0.623	0.598

Table 3.10: F1 Scores for Models on English Dataset

3.5 Conclusions

After carefully running various experiments with these models and cleaning our data, we found out that for the English dataset:

- Linear SVM outperforms both the Naive Bayes models on both types of features (word, char) for in-domain Facebook data testing.
- For out-of-domain Twitter testing, SVM and Bernoulli Naive Bayes are primarily comparable.
- Char and Word features are comparable for in-domain (Facebook) data, but character models are better for generalising for out-of-domain (Twitter).

Similarly, For Hindi:

- For in-domain Facebook testing, SVM is better when using char features but comparable to Multinomial NB when using word features.
- For out-of-domain Twitter testing, SVM is significantly better.
- Character models are significantly better than word models for both in-domain and out-of-domain data.

Params	Word	Char
clf__alpha	0.1	0.01
vect__max_df	0.5	0.5
vect__ngram_range	(1, 2)	(3, 6)

Table 3.11: BernoulliNB's best params - Hindi

MultinomialNB	Word	Char
clf__alpha	0.01	0.1
tfidf__use_idf	False	True
vect__max_df	0.5	0.5
vect__ngram_range	(1, 2)	(3, 6)

Table 3.12: MultinomialNB's best params - Hindi

We also noticed that due to our superior understanding of the dataset and its preprocessing and cleaning prowess, we saw that our results were comparable to the best results seen in the TRAC-1 submissions.

For English, the in-domain Facebook data, we managed a score of **0.628** using Linear SVM on Word NGram. In comparison, the shared task best score was **0.6425** by Aroyehun and Gelbukh [84] who used LSTM and Data Augmentation through translation. For the out-of-domain Twitter data, our best score was **0.598** using Linear SVM on Character N-Gram, while the shared task's best was **0.595**, achieved by Raiyani et al. [83] using a Multi-layer Perceptron on Bag of Words.

For Hindi, the in-domain Facebook data, we managed a score of **0.641** using Linear SVM on Character NGram. In contrast, the shared task best score was **0.6292** by Samghabadi et al. [81] who used Logistic regression models taking word ngram and character ngram. For the out-of-domain Twitter data, our best score was **0.490** using Linear SVM on Character N-Gram, while the shared task's best was **0.5** by Modha et al. [80] who used CNN on Fast-Text Embeddings.

Our system managed to beat state of the art in the 2018 dataset. Our final takeaway from this was that despite using no transliteration techniques, translation, or data augmentation, ensemble leaning, or neural embedding models, our models are best for the Hindi dataset (both in-domain/out-of-domain). They are also the best on English out-of-domain Twitter (comparable on the in-domain Facebook dataset). Thus, we concluded that the key to success is a clear understanding of the dataset, a simple model with great preprocessing.

The comparison can be seen in Table 3.16 for English and 3.17 for Hindi.

In spite of the fact that the models performed well on the dataset, we are aware that these do not represent the state of research of NLP. These results can be improved further by adding our cleaning

LinearSVM	Word	Char
clf__alpha	0.0001	0.0001
clf__max_iter	50	50
vect__max_df	0.7	1.0
vect__ngram_range	(1, 2)	(3, 6)

Table 3.13: Linear SVM's best params - Hindi

Model	Value Score
BernoulliNB - Word	0.562
BernoulliNB - Char	0.557
MultinomialNB - Word	0.579
MultinomialNB - Char	0.578
LinearSVM - Word	0.597
LinearSVM - Char	0.631

Table 3.14: Value Scores for Models - Hindi

prowess with neural embedding models such as word2vec [100], Glove [101] for word embeddings - FastText [102] for character based word representations.

3.6 Future Work

Some avenues to improve the research would be:

- Language specific cleaning would have enhanced the results for Hindi
- Data augmentation (transliteration of CodeMixed into one single script)
- Morphological analysis

Model	Features	Facebook Data	Twitter Data
BernoulliNB	Word	0.575	0.355
MultinomialNB	Word	0.595	0.385
LinearSVM	Word	0.595	0.457
BernoulliNB	Char	0.561	0.381
MultinomialNB	Char	0.602	0.419
LinearSVM	Char	0.641	0.490

Table 3.15: F1 Scores for Models on Hindi Dataset

	English	
Models	**Facebook In-Domain**	**Twitter Out-Domain**
Our Best	0.628	**0.598**
Shared Task Best	**0.6425**	0.595

Table 3.16: Comparison on English dataset

	Hindi	
Models	**Facebook In-Domain**	**Twitter Out-Domain**
Our Best	**0.641**	0.490
Shared Task Best	0.6292	**0.5**

Table 3.17: Comparison on Hindi dataset

- We are using bag of words features, previously mentioned avenues become more important when using the neural embedding approaches such as word2vec, etc. (so that we can have embedding in the same space)
- Exploring BERT [91]

However, we also saw that multilingual BERT models multilingual models might not be great for codemixed data as pointed by Pires [103]. Exploring how a lot of the transformer based multilingual models perform well on noisy social media platforms data would be an intriguing task.

We did not proceed in improving out models further since they were already beating the state of the art, as the thesis focuses on understanding nuances of aggression more, to improve our results even more, we wish to diversify our knowledge of aggression rather than modeling aggression.

All the mentioned datasets, models, and scripts are publicly available at : https://github.com/arjitsrivastava/MultidimensionalViewOpinionMining

Chapter 4

Aggression and opinion

I alone cannot change the world, but I can cast a stone across the waters to create many ripples.

- Mother Teresa, *1973*

The advent of social media has immensely proliferated the number of opinions and arguments voiced on the internet. These virtual debates often present cases of aggression. After building models for aggression on the dataset released by [104], we wanted to explore further the nuances which data sets involving opinions could present. While a lot of research has been done on analyzing aggression and stance in isolation from each other, an attempt to understand patterns of aggression and figurative language use when voicing opinion was attempted in this chapter. A Hindi-English code-mixed dataset of opinion on the politico-social issue of *2016 India banknote demonetization*, which was annotated across multiple dimensions such as aggression, hate speech, emotion arousal and figurative language usage (such as sarcasm/irony, metaphors/similes, puns/word-play) is presented.

4.1 Introduction

Social media and online forums encourage users to share their thoughts with the world, resulting in a vast resource of opinion-rich data. This phenomenon has garnered a lot of attention from the research community. Since it allows for analyzing the interactions between users as well as their usage of informal language in depth.

The author of any given opinion may be in favour, against or neutral towards the issue at hand. The target issue analyzed for this purpose is *2016 Indian banknote demonetization*. On 8 November 2016, the Government of India announced the demonetization of all ₹500 and ₹1,000 banknotes of the Mahatma Gandhi Series. It also announced that new banknotes of ₹500 and ₹2,000 banknotes would be circulated in exchange for the demonetized banknotes. However, this decision received mixed reactions from the people of India, with many people questioning its effectiveness.

People often express their opinion on socio-political issues on social media forums like Twitter by displaying aggression towards people that support a contradicting belief or towards a particular group of stakeholders on the issue. Given below are some example tweets from our dataset on the target issue of demonetization in India. **The reader is warned on the strongly-worded and derogatory nature of these tweets.**

1. **Tweet**: '*ye AAPtards aise behave kar rahe hain jaise Modi ji ne Notebandi nahi inki Nassbandi kara di ho*'

   Translation: '*These AAPtards are behaving as if its not demonetisation but castration for them.*'

   Glossary: "*AAP*": Opposition political party, "AAPtards": slang term for supporters of AAP (inspired by the English slang "Libtards"), "Notebandi": demonetisation of higher currency notes, "Nassbandi": castration

2. **Tweet**: '*Aam admi se jyada politician ko dikate ho rhi h notebandi se aisa kyun?????*'

   Translation: '*Politicians seem to be more affected by demonetisation compared to the common man. Why is so ?*'

3. **Tweet**: '*tera kejri mar jaye sala suar to modi ji vaise he ek din ke liye notebandi vapis le lege*'

   Translation: '*If your leader Kejri, a stupid pig, dies then Modi ji would take demonetisation back for a day.*'

   Glossary: "*Kejri*": refering to Arvind Kejriwal (leader of opposition party AAP), "*Modi ji*": Honorific refering to Narendra Modi (Prime Minister of India)

4. **Tweet**: '*what if .. Modi Ji says Mitron ,,, kal raat ko zyada ho gayi thi ,,. Kuch nahi badla he.. #NoteBandi*'

   Translation: '*What if Modi says that he had too much to drink last night and nothing has really changed. #Demonetisation.*'

This is an attempt at analyzing social media opinion on a political issue across varied modalities. More in-depth datasets like the one presented here are required for:

- Analyzing the not so apparent forms of verbal aggression displayed on social media.
- Understanding linguistic patterns when voicing an opinion and displaying aggression in a better manner.
- Analyzing social dynamics of opinion.
- Facilitate classification models that leverage corpora annotated for auxiliary tasks through transfer learning, joint modelling as well as semi-supervised label propagation methods.

4.2 Data Statistics and Analysis

[105] had collected 3500 code-mixed Hindi-English tweets. These tweets were collected using the Twitter Scraper API, where filtering by the keywords "*notebandi*" and "*demonetization*" over 6 months after Demonetisation was implemented was done. These tweets were then further annotated for stance (*favourable, against and neutral*). This dataset has 964 tweets in favor, 647 tweets against and 1934 tweets that have no stance towards the target.

Table 4.1 presents the tweet level average statistics on the corpus. The dataset tweets contain majorly Hindi language tokens (written in the Roman script instead of Devanagari). A total of 119 tweets had discernible code-mixing (3 or more English words). As our tweets were sampled from the dataset by [105], who had referred to their dataset as code-mixed, we continue to refer it that way. Subsequent model building on this corpus would benefit from special handling for token-level spelling differences that come with Devanagari to Latin script switching for Hindi.

Avg. # tokens	21.1
Avg. # tokens (EN)	1.0
Avg. # tokens (HI)	16.9
Avg. # tokens (Rest)	3.2

Table 4.1: Tweet Level Statistics

We randomly sampled **1001** tweets from this dataset and annotated these sampled tweets. The reason for picking up a lower number of tweets to annotate was to ensure an honest attempt at a high quality of annotation across multiple dimensions.

4.3 Annotation

In our sampled dataset, the dimensions which we annotated for are: (3 domain expert annotators for each dimension).

- **Aggression**: Overt vs Covert vs Neutral
- **Hate Speech**: True vs False
- **Sarcasm / Irony / Rhetorical Question**: True vs False
- **Metaphor / Simile**: True vs False
- **Pun / Word-play**: True vs False
- **Emotion Arousal**: 5 point ordinal scale

The final label on each binary classification dimension was taken as the majority label from choices of 3 annotators. For aggression classification, which was a multi-class classification, adjudication was provided for cases where no simple majority could be reached. For emotion arousal levels, scores from individual annotators were averaged for the final emotion arousal level score.

The original dataset for stance was also re-annotated, for it had favourable or against tags only on tweets that displayed outright support or disapproval respectively. It was found that the majority of opinion was displayed through attacking/supporting other opinions on the issue, i.e. examples of indirect or implied support/disapproval. For example, if we look at the tweet below:

1. **Tweet**: '*Notebandi k khilaf kyu ho...? Kaale dhan m share holder ho kya @ArvindKejriwal*'

   Translation: '*Why are you against demonetisation ? Are you a shareholder in black money @ArvindKejriwal*'

   Glossary: "*kale dhan*": black money, "*Arvind Kejriwal*": Leader of opposition political party AAP, "*Notebandi*": demonetisation

   Tweet 1 was originally classified as a neutral stance. On further thought, the realization that cases like above can be confidently annotated as favourable to the issue (i.e. favourable to demonetization). The author rhetorically and sarcastically questions the opinion, intentions and reasons of those against the issue (in this case leader of opposition party). This tweet is also an example of what was considered as covert aggression.

For aggression annotation, we follow the guidelines by [88] who had presented a detailed typology of aggression on Hindi-English code-mixed data. We only annotate for aggression level. While they had

additional layers based on discursive role (attack, defend, abet) and discursive effect (physical threat, sexual aggression, gendered aggression, racial aggression, communal aggression, casteist aggression, political aggression, geographical aggression, general non-threatening aggression, curse). The definitions for three aggression levels along with examples from our dataset are:

Covertly-Aggressive (C) Contains text which is an indirect attack and is often packaged as (insincere) polite expressions (through the use of conventionalized polite structures) such as satire, rhetorical questions, etc.

2. **Tweet**: '*Notebandi ka niyam : khata nahi hai to khulwao. Aam aadmi : khulwa to lun. Par bhai bank main ghusun Kasey ?*'

   Translation: '*Rule of Demonetisation: If you do not have an account then open one. Common man: I will open but let me know how to enter the bank first?*'

   Disapproval of demonetization through sarcastic reference to long queues in front of banks due to high demand for exchange of demonetized currency.

Overtly-Aggressive (O) Contains texts in which aggression is overtly expressed either through the use of specific kind of lexical items, syntactic structures or lexical features.

3. **Tweet**: '*Ye Notebandi Atankbaadiyo aur Bharashtachaariyo ki NAKEBANDI hai. Sare Rashtrabhakta is nakebandi ke sath aur samarthan me aye.*'

   Translation: '*Demonetisation is a barricading of terrorists and corrupt. All the nationalists should support this barricading.*'

Non-Aggressive (NAG) Refers to texts which are not lying in the above two categories.

4. **Tweet**: '*kya Aam aadmi ke liye NoteBandi ka Faisla Shi hai?*'

   Translation: '*Is the decision of demonetisation in the favour of common man?*'

Prior works regarding sarcasm and irony detection on social media data like Reddit [106], and Twitter [107] have shown that context is essential in understanding sarcasm. Therefore, most social media datasets of sarcasm are self-annotated, i.e. hashtag specific Twitter scraping like #sarcasm and #not-sarcasm. As we are re-annotating a previously scraped dataset which was not self-annotated through

specific hashtags, we rely on the domain knowledge of context expert annotators on the Indian socio-political scenario and focus issue of demonetization. This, however, is not a drawback because in a dataset like ours. Since it is rich with strongly opinionated tweets, annotating sarcasm is reasonably straightforward.

In the current scope of the research, rhetorical questions are thought of as functioning, similar to sarcasm and irony. It is understood that there are a lot of subtle grained linguistic differences between sarcasm, irony and rhetorical questions which exist. However, for our purpose, we have clubbed them into a single category of figurative language. Similarly, puns and word-play are merged into a single category of figurative language as well, and the annotation guidelines were based on the SemEval 2017 task of detecting English puns [108]. Rhyming usage of 'Notebandi' (demonetization) with 'Nasbandi' (castration) as shown in the earlier examples, was the most common word-play seen. A third figurative language category of metaphors (and occasionally similes) can also be observed in our corpus. Metaphor identification has been typically treated as a token level, or phrase-level tagging task [109]. To be consistent with other figurative language categories used in this work, metaphors were annotated at the tweet level which was also the annotation level for SemEval 2015 task on figurative language in Twitter data [110]. The following tweet is an example of metaphor usage:

5. **Tweet**: *'kabhi kabhi sher ka shikar karne ke liye bhed (aam janta) ko chara banana padta hai.' '#notebandi'*

   Translation: '*Sometimes sheep need to be sacrificed in order to to hunt lions #Demonetisation*.'

   In tweet 9, 'sheep' is a metaphor for some members of common public and 'lions' is a metaphor for large scale corruption.

Burnap and Williams [68] defined hate speech as responses that include written expressions of hateful and antagonistic sentiment toward a particular race, ethnicity, or religion. They used a binary classification scheme of hate speech vs non-hate speech, which was also followed by Bohra et al. [69] for their dataset on Hindi-English code-mixed tweets. Malmasi and Zampieri [70] used a three-way classification scheme between hate speech vs offensive language but not hate speech vs no offensive language. As aggression levels are highly predictive of offensive language but not of hate speech category, we used a binary classification speech. However, annotators faced difficulty in differentiating over a personal attack full of hatred than a community being targeted. An example:

6. **Tweet**: '*ab itni taklif hai to atmadaah kyo nahi kar lete notebandi k khilf. Delhi walo ko bhi mukti milegi tumse*'

   Translation: '*If you have such a huge issue with it, why don't you perform self-immolation?*

The people of Delhi would also get freedom from you'

In tweet 10, the author is referring to Arvind Kejriwal, who is the leader of opposition party AAP and also the Chief Minister of New Delhi (capital of India). The author suggests that Kejriwal should kill himself to free the residents of Delhi. In the process of supporting the decision of Demonetisation, the author of the tweet is making extreme and graphic suggestions towards one of the main opponents of target issue.

Emotion classification in text is widely understood as lying across two orthogonal dimensions - valence (polarity of emotion) and arousal (intensity of emotion) russell1999emotion. Despite that, many works on emotion classification in text have generally used directly annotated six emotion categories (happy, sad, anger, fear, disgust, surprise) instead of first annotating arousal and valence separately before mapping them into emotion categories. The scope was restricted for our research to analyze only for emotion arousal level as emotion valence level is analogous to sentiment. For emotion arousal level, Bradley and Lang [111] averaged annotations on a 9 point scale and Mohammad [112] used a Best-Worst scale to obtain fine-grained scores. Similar to the SemEval 2017 task [113] for sentiment analysis on Twitter, we use a 5-point ordinal scale (Very Low, Low, Neutral, High, Very High) for emotion arousal level.

Task	Category	# Tweets
Stance	Favour	583
	Against	180
	Neutral	238
Aggression	Overt	140
	Covert	264
	None	597
Hate Speech	True	29
Figurative Language	Sarcasm / Irony / Rhetorical Ques.	163
	Word-play / Pun	140
	Metaphor / Simile	189

Table 4.2: Distribution of annotations across corpus

Table 4.2 has the corpus wide statistics across various phenomena annotated. There is a significant skew towards favourable stance in the corpus. To accommodate for this imbalance, subsequent analysis of phenomena concerning stance contain marginal class percentage statistics, for example, percentage of sarcastic tweets in favour of the issue concerning the total number of tweets favourable to the issue.

Another point to note is the insufficient number of hate speech instances. This could be attributed to the stringent guideline that only directed abusive attacks on specific groups/communities are to be regarded as hate speech. Annotations with looser guidelines, where individual offensive language against individuals are also considered hate speech, would correlate highly with overt aggression category. Since we annotated on tweets regarding a polarizing legislation, it was expected that a fair amount would display aggression (either covert or overt). The same observation is evident from the statistics.

4.3.1 Annotation Agreement

Task	Fleiss's kappa
Stance	0.84
Aggression	0.62
Hate Speech	0.47
Sarcasm / Irony / Rhetorical Questions	0.61
Puns / Word-play	0.72
Metaphors / Similes	0.65

Table 4.3: Fleiss's kappa score on multiple annotations across dimensions

We used Fleiss's kappa to measure inter-annotator agreement on categorical annotation tasks, and the results are given in table 4.3. Due to the explicit polarizing nature of the issue at hand, annotations for stance were of very high correlation. Hate speech annotations had the worst kappa score and can be attributed to what constitutes an abusive personal attack. For figurative language use, the annotations for puns and word-play were of higher correlation as can be expected due to the apparentness in surface forms. Annotations for sarcasm/irony / rhetorical questions while still being of a high agreement had lower agreement rate than both metaphors/similes as well as word-play. This can be attributed to the more significant subjective nature of sarcasm as well as it being a more context-dependent phenomenon than metaphor or word-play.

Spearman's Rank Correlation Emotion Arousal		
Annotator	2	3
1	0.655	0.652
2		0.64

Table 4.4: Spearman correlation on emotion arousal annotations across annotator pairs

Table 4.4 gives the Spearman's rank correlation coefficient across three annotators for emotional arousal, which has been rated on an ordinal scale of 1 to 5. Although annotating for emotion is a reasonably tricky task and annotating for only the arousal dimension even more so. However, we achieve a decent average correlation of 0.65 which can be attributed to the fact that these tweets were sampled for a polarizing issue which had apparent emotional states (high arousal emotions like anger as well as low arousal emotions like sadness). For each pair of annotators, the results of emotional arousal agreement were statistically significant with p-values $<<<$ 0.005.

4.4 Opinion Specific Analysis

Stance	**Marginal Class % of Hate Speech**
Favour	2.92%
Against	2.2%
Neutral	3.36%

Table 4.5: Distribution of hate speech

Table 4.5 presents the statistics of hate speech across stance classes. An anomalous observation is the higher marginal percentage of hate speech evidence for *neutral* stance. This could be attributed to the poorer understanding of what constitutes hate speech. Additionally, upon investigating, we found tweets similar to the one given below. Though the tweet does not take a definitive stance on the issue at hand (demonetization), it is an abusive personal attack at an individual as well as a group.

11. **Tweet**: '*MR. RAVISH VYAPARI IMAANDAR HAI.KANOON KA SANMAAN KARTSHAI. PAR MEDIA NEWS AUR TV SAB SAALE CHOR AUE HARAMKHOR HAI. NOTEBANDI*'

    Translation: '*Mr. Ravish, businessmen are honest and respect the law. But media, news and TV (personalities) are thieves and bastards.*'

    Glosses: "*Ravish*": Refering to news anchor Ravish Kumar

    Tweet 11, defends integrity of businessmen while attacking and name calling news personalities.

Table 4.6 gives the distribution of aggression categories (*covert / overt / non*) across stance. It is interesting to note the comparisons for overt vs covert aggression when in favour (majority population stance in this sample) as opposed to against (minority population in this sample) on the issue. Although covert aggression evidence is always more than overt aggression evidence across stance categories, the difference is much lesser for *favourable* stance samples. It is not difficult to hypothesize that holding a

Stance	Aggression	Marginal Class % Aggression
Against	Overt	8.3%
	Covert	40%
	None	**51.7%**
Favour	Overt	17.8%
	Covert	23.8%
	None	**58.3%**
Neutral	Overt	8.8%
	Covert	22.3%
	None	**68.9%**

Table 4.6: Distribution of aggression across stance

majority stance on issues will lead to open bullying in many cases. Users in the minority tend to be more covert to avoid being bullied by the majority group, possibly. However, validating this social hypothesis based on the analysis of multiple issues is beyond our current scope.

Stance	Sarcasm / Irony / Rhetorical Question	
	Raw Count	Marginal Class %
Against	45	**25%**
Favour	76	13%
Neutral	42	17.6%

Table 4.7: Distribution of sarcasm, irony and rhetorical questions

Stance	Pun / Word-play	
	Raw Count	Marginal Class %
Against	30	**16.7%**
Favour	84	14.4%
Neutral	26	10.9%

Table 4.8: Distribution of pun and word-play

Tables 4.7, 4.8 and 4.9 present the distributions of figurative language use across stance classes. It is evident from the data of *against* issue category, the usage of all types of figurative language is consistently high. It should also be noted that evidence for sarcasm is especially stronger in *against* issue opinion (minority stance in this dataset). Keeping in mind the observations on covert aggression when voicing minority stance, it can be noted that covert aggression is expressed through figurative language like sarcasm and puns. Metaphors are not as disguised as sarcasm and puns, and we see that it does

Stance	Metaphor / Simile	
	Raw Count	**Marginal Class %**
Against	35	19.4%
Favour	123	**21.1%**
Neutral	31	13%

Table 4.9: Distribution of metaphor and simile

not follow the same pattern concerning stance. The scope of this work is limited to a single issue, and it would be interesting to note if these trends are observed across datasets. A dataset of annotations of multiple issues would allow for hypothesis testing to validate these trends.

Stance	Marginal Class %					Emotional Arousal
	1-2	**2-3**	**3-4**	**4-5**	**5**	**Class Avg.**
Favour	3.6%	23.67%	49.91%	18.01%	4.8%	3.26
Against	13.3%	26.67%	47.78%	10.56%	1.67%	2.91
Neutral	10.9%	36.97%	44.12%	6.72%	1.3%	2.83

Table 4.10: Marginal distribution of emotional arousal

Finally in table 4.10, statistics for emotion arousal are presented across stance classes. Opposed to prior analyzed phenomena (hate speech, aggression and figurative language use), the data for emotion arousal is ordinal on a 1 to 5 scale. The average emotion arousal for *favourable* stance (majority class) is much more than that in *against* stance (minority class). Similarly, looking at the very high arousal state bucket of *5* emotion arousal (when all three annotators gave a 5 rating), the percentage for majority stance (favourable) is three times than that for minority stance (against). These findings are in line with the observations for other phenomena like overt aggression and figurative language use in the majority stance. The higher percentage of lowest arousal state tweets when against the issue must also be noted. These lowest arousal tweets correspond to emotions like depression and sadness.

4.5 Translation of data

We further translated our data to English so that researchers could take advantage of the same, and draw comparisons with code-mixed approaches. We used simple normalisation techniques when translating the dataset. We ensured the following:

- "notebandi" was converted to demonetisation including "#notebandi"

- Hashtags in the middle of the tweet, are removed and a maximum limit for them is kept. Hashtags at the beginning or the end of a tweet are not removed.
- Hyperlinks are removed from all tweets.
- Consistency with translating "nasbandhi, nasbandi, nashbandi" or its variations to vasectomy.

Such a unified dataset should assist on other linguistic patterns
All the mentioned datasets (including this one) are publicly available on the author's Github page.

4.6 Conclusion and Future Work

This research was motivated by the need to provide a ground-work for analysis of the nuances of opinion on social media with respect to aggression and figurative language use. The observed correlations are encouraging and call for a deeper analysis of these social dynamics. Testing for statistical significance along with corpus-linguistic analysis of informative words for each category was beyond our current scope. The first aim would be to create similar corpora on wide variety of issues (not limited to political debate) to evaluate the consistency of these trends and determine significance of our findings.

Though the scope was limited to corpus creation and analysis of interactions across phenomena, the larger goal is to allow for better classification systems on social media data. An immediate goal is to build baseline models and analyze their performance on the different phenomena annotated in this corpus. It would be interesting to compare performance of models that directly model a single dimension with those models that have cascaded or joint modeling on multiple dimensions. Another avenue to explore is semi-supervised label propagation utilizing both larger corpora on a single dimension such as sarcasm as well as this corpus containing multi-dimensional annotations. Having a single corpus of annotations across dimensions has allowed the possibility to explore transfer learning strategies in classification.

Author's Github page. `https://github.com/arjitsrivastava/MultidimensionalViewOpinionMining`

Chapter 5

Conclusions

In this thesis, we first presented our motivation for pursuing and understanding a concept like aggression. We touched upon the impact of social media on society. Furthermore, issues about society with much aggressive content being generated. We further talked about the background work undertaken to understand the problem at hand. We discussed the various thoughts and ideas behind the concept of aggression.

We then explored the TRAC-1 dataset for making machine learning models, for identifying aggression in the given datasets. We shared our results for the same, comparable to the best models from the shared task for most tasks and were better than state of the art for other cases. For now, our datasets experiments were limited as the majority work dealt with code-mixed social media data. Understanding code-mixed datasets and their relationship with word embeddings and experiments to improve the word embeddings for the same will also help bring many code-mixed informal text classification tasks to the state-of-the-art positions of similar monolingual tasks. This, in the long run, this will attract researchers and bring new research ideas to this field.

Further, we dived deeper into a sociopolitical topic on social media to understand aggression and its nuances when the matter of putting forth your opinion comes into the picture. We also saw exciting comparisons for overt vs covert aggression when favouring (majority population's stance in this sample) instead of against (minority population in this sample) on the issue. We hypothesised that holding a majority stance on issues will lead to open bullying in many cases. Users in the minority tend to be more covert to avoid being bullied by the majority group, possibly. We also present our conclusions from this analysis. Such an analysis would be beneficial for identifying the dynamics of society further. Future consolidated works in other dimensions, etc. would help in future research with product suggestions.

Related Publications

1. Srivastava, **Arjit**, et al. "A Multi-Dimensional View of Aggression when voicing Opinion." Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying. 2020.

2. Srivastava, **Arjit**, et al. "Understanding aggression through the lens of transformers." To be submitted.

Secondary Publications

1. **Arjit Srivastava** and Navjyoti Singh, A dataset for Indian Legal Judgments: ICJ. "ICAIL 2017: XVI International Conference on AI and Law." Link: `https://www.andrew.cmu.edu/user/mgrabmai/asail2017/`

2. Minocha, Akshay, Navjyoti Singh, and **Arjit Srivastava**. "Finding relevant Indian judgments using dispersion of citation network." Proceedings of the 24th International Conference on World Wide Web. 2015.

3. Akhtar, Syed Sarfaraz, Arihant Gupta, Avijit Vajpayee, **Arjit Srivastava**, and Manish Shrivastava. "Word similarity datasets for indian languages: Annotation and baseline systems." In Proceedings of the 11th Linguistic Annotation Workshop, pp. 91-94. 2017.

4. Akhtar, Syed Sarfaraz, Arihant Gupta, Avijit Vajpayee, **Arjit Srivastava**, and Manish Shrivastava. "Unsupervised morphological expansion of small datasets for improving word embeddings." arXiv preprint arXiv:1711.05678 (2017).

5. Akhtar, Syed Sarfaraz, Arihant Gupta, Avijit Vajpayee, **Arjit Srivastava**, Madan Gopal Jhawar, and Manish Shrivastava. "An unsupervised approach for mapping between vector spaces." arXiv preprint arXiv:1711.05680 (2017).

6. Gupta, Arihant, Syed Sarfaraz Akhtar, Avijit Vajpayee, **Arjit Srivastava**, Madan Gopal Jhanwar, and Manish Shrivastava. "Exploiting morphological regularities in distributional word representations." In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 292-297. 2017.

Bibliography

[1] K. S., "July 2020, global digital statshot report.," 2020.

[2] J. W. Du Bois, "The stance triangle. stancetaking in discourse: Subjectivity, evaluation, interaction, ed. by robert englebretson, 139-182," 2007.

[3] A. AlDayel and W. Magdy, "Stance detection on social media: State of the art and trends," *arXiv preprint arXiv:2006.03644*, 2020.

[4] R. S. Lazarus, "Progress on a cognitive-motivational-relational theory of emotion.," *American psychologist*, vol. 46, no. 8, p. 819, 1991.

[5] S. Schachter and J. Singer, "Cognitive, social, and physiological determinants of emotional state.," *Psychological review*, vol. 69, no. 5, p. 379, 1962.

[6] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 3, p. 715, 2005.

[7] N. Newman, "Mainstream media and the distribution of news in the age of social discovery,"

[8] J. Lipski, "Code-switching and the problem of bilingual competence," *Aspects of bilingualism*, vol. 250, p. 264, 1978.

[9] J. J. Gumperz, *Discourse strategies*, vol. 1. Cambridge University Press, 1982.

[10] N. Isharyanti, "Mónica stella cárdenas-claros," 2009.

[11] L. A. Shafie and S. Nayan, "Languages, code-switching practice and primary functions of facebook among university students," *Study in English Language Teaching*, vol. 1, no. 1, pp. 187–199, 2013.

[12] J. Lee and Y. Lee, "A holistic model of computer abuse within organizations," *Information management & computer security*, 2002.

[13] L. Corcoran, C. M. Guckin, and G. Prentice, "Cyberbullying or cyber aggression?: A review of existing definitions of cyber-based peer-to-peer aggression," *Societies*, vol. 5, no. 2, pp. 245–255, 2015.

[14] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web*, pp. 591–600, 2010.

[15] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in *2010 43rd Hawaii international conference on system sciences*, pp. 1–10, IEEE, 2010.

[16] G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce, *et al.*, "The arab spring— the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions," *International journal of communication*, vol. 5, p. 31, 2011.

[17] K. Lovejoy, R. D. Waters, and G. D. Saxton, "Engaging stakeholders through twitter: How nonprofit organizations are getting more out of 140 characters or less," *Public relations review*, vol. 38, no. 2, pp. 313–318, 2012.

[18] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, pp. 851–860, 2010.

[19] X. Hu and H. Liu, "Text analytics in social media," in *Mining text data*, pp. 385–414, Springer, 2012.

[20] X. Zhou and L. Chen, "Event detection over twitter social media streams," *The VLDB journal*, vol. 23, no. 3, pp. 381–400, 2014.

[21] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on twitter," 2011.

[22] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "Diffusion of lexical change in social media," *PloS one*, vol. 9, no. 11, p. e113114, 2014.

[23] T. Baldwin, P. Cook, M. Lui, A. MacKinlay, and L. Wang, "How noisy social media text, how diffrnt social media sources?," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 356–364, 2013.

[24] S. Brody and N. Diakopoulos, "Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 562–570, 2011.

[25] J. W. Y. Ho *et al.*, "Code-mixing: Linguistic form and socio-cultural meaning," *The International Journal of Language Society and Culture*, vol. 21, no. 7, pp. 1–8, 2007.

[26] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pp. 71–80, IEEE, 2012.

[27] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th international conference on world wide web*, pp. 145–153, 2016.

[28] S. Sax, "Flame wars: Automatic insult detection," 2016.

[29] A. Bansal, S. M. Sharma, K. Kumar, A. Aggarwal, S. Goyal, K. Choudhary, K. Chawla, K. Jain, M. Bhasin, *et al.*, "Classification of flames in computer mediated communications," *arXiv preprint arXiv:1202.0617*, 2012.

[30] E. Cambria, P. Chandra, A. Sharma, and A. Hussain, "Do not feel the trolls," *ISWC, Shanghai*, 2010.

[31] L. G. M. de la Vega and V. Ng, "Modeling trolling in social media conversations," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[32] T. Mihaylov, G. Georgiev, and P. Nakov, "Finding opinion manipulation trolls in news community forums," in *Proceedings of the nineteenth conference on computational natural language learning*, pp. 310–314, 2015.

[33] S. Kumar, F. Spezzano, and V. Subrahmanian, "Accurately detecting trolls in slashdot zoo via decluttering," in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pp. 188–195, IEEE, 2014.

[34] T. Nitta, F. Masui, M. Ptaszynski, Y. Kimura, R. Rzepka, and K. Araki, "Detecting cyberbullying entries on informal school websites based on category relevance maximization," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 579–586, 2013.

[35] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste, "Detection and fine-grained classification of cyberbullying events," in *International Conference Recent Advances in Natural Language Processing (RANLP)*, pp. 672–680, 2015.

[36] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 3, pp. 1–30, 2012.

[37] M. Dadvar, D. Trieschnigg, and F. de Jong, "Experts and machines against bullies: A hybrid approach to detect cyberbullies," in *Canadian Conference on Artificial Intelligence*, pp. 275–281, Springer, 2014.

[38] C. Hardaker, "Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions," *Journal of politeness research*, vol. 6, no. 2, pp. 215–242, 2010.

[39] C. Hardaker, ""uh.... not to be nitpicky, but... the past tense of drag is dragged, not drug.": An overview of trolling strategies," *Journal of Language Aggression and Conflict*, vol. 1, no. 1, pp. 58–86, 2013.

[40] E. Krol, *The whole Internet: user's guide & catalog*. O'Reilly & Associates, Inc., 1994.

[41] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *Proceedings of the NAACL student research workshop*, pp. 88–93, 2016.

[42] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in *Proceedings of the 26th International Conference on World Wide Web*, pp. 1391–1399, 2017.

[43] N. S. Samghabadi, S. Maharjan, A. Sprague, R. Diaz-Sprague, and T. Solorio, "Detecting nastiness in social media," in *Proceedings of the First Workshop on Abusive Language Online*, pp. 63–72, 2017.

[44] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 759–760, 2017.

[45] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proceedings of the Fifth International workshop on natural language processing for social media*, pp. 1–10, 2017.

[46] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network," in *European semantic web conference*, pp. 745–760, Springer, 2018.

[47] R. A. Baron and D. R. Richardson, *Human aggression*. Springer Science & Business Media, 2004.

[48] N. R. Crick and J. K. Grotpeter, "Relational aggression, gender, and social-psychological adjustment," *Child development*, vol. 66, no. 3, pp. 710–722, 1995.

[49] D. Biber and E. Finegan, "Adverbial stance types in english," *Discourse processes*, vol. 11, no. 1, pp. 1–34, 1988.

[50] D. McKendrick and S. A. Webb, "Taking a political stance in social work," *Critical and Radical Social Work*, vol. 2, no. 3, pp. 357–369, 2014.

[51] A. Jaffe *et al.*, *Stance: sociolinguistic perspectives*. OUP USA, 2009.

[52] S. Djemili, J. Longhi, C. Marinica, D. Kotzinos, and G.-E. Sarfati, "What does twitter have to say about ideology?," in *NLP 4 CMC: Natural Language Processing for Computer-Mediated Communication/Social Media-Pre-conference workshop at Konvens 2014*, vol. 1, pp. http–www, Universitätsverlag Hildesheim, 2014.

[53] O. Fraisier, G. Cabanac, Y. Pitarch, R. Besançon, and M. Boughanem, "Stance classification through proximity-based community detection," in *Proceedings of the 29th on Hypertext and Social Media*, pp. 220–228, 2018.

[54] M. Lai, D. I. H. Farías, V. Patti, and P. Rosso, "Friends and enemies of clinton and trump: using context for detecting stance in political tweets," in *Mexican International Conference on Artificial Intelligence*, pp. 155–168, Springer, 2016.

[55] W. Magdy, K. Darwish, N. Abokhodair, A. Rahimi, and T. Baldwin, "# isisisnotislam or# deportallmuslims? predicting unspoken views," in *Proceedings of the 8th ACM Conference on Web Science*, pp. 95–106, 2016.

[56] N. Newman, "Mainstream media and the distribution of news in the age of social media," 2011.

[57] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "Semeval-2016 task 6: Detecting stance in tweets," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 31–41, 2016.

[58] A. Murakami and R. Raymond, "Support or oppose? classifying positions in online debates from reply activities and opinion expressions," in *Coling 2010: Posters*, pp. 869–875, 2010.

[59] S. Somasundaran and J. Wiebe, "Recognizing stances in online debates," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 226–234, 2009.

[60] P. Sobhani, S. Mohammad, and S. Kiritchenko, "Detecting stance in tweets and analyzing its interaction with sentiment," in *Proceedings of the fifth joint conference on lexical and computational semantics*, pp. 159–169, 2016.

[61] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and sentiment in tweets," *ACM Trans. Internet Technol.*, vol. 17, June 2017.

[62] A. Aldayel and W. Magdy, "Assessing sentiment of the expressed stance on social media," in *International Conference on Social Informatics*, pp. 277–286, Springer, 2019.

[63] T. McGonagle *et al.*, "The council of europe against online hate speech: Conundrums and challenges," in *Expert paper. Belgrade: Council of Europe Conference of Ministers responsible for Media and Information Society*, 2013.

[64] L. B. Nielsen, "Subtle, pervasive, harmful: Racist and sexist remarks in public as hate speech," *Journal of Social Issues*, vol. 58, no. 2, pp. 265–280, 2002.

[65] L. Leets, "Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speech," *Journal of social issues*, vol. 58, no. 2, pp. 341–361, 2002.

[66] D. M. Downs and G. Cowan, "Predicting the importance of freedom of speech and the perceived harm of hate speech," *Journal of applied social psychology*, vol. 42, no. 6, pp. 1353–1375, 2012.

[67] R. M. Simpson, "Dignity, harm, and hate speech," *Law and Philosophy*, vol. 32, no. 6, pp. 701–728, 2013.

[68] P. Burnap and M. L. Williams, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," *Policy & Internet*, vol. 7, no. 2, pp. 223–242, 2015.

[69] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava, "A dataset of hindi-english code-mixed social media text for hate speech detection," in *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pp. 36–41, 2018.

[70] S. Malmasi and M. Zampieri, "Detecting hate speech in social media," *arXiv preprint arXiv:1712.06427*, 2017.

[71] W. B. Cannon, "The james-lange theory of emotions: A critical examination and an alternative theory," *The American journal of psychology*, vol. 39, no. 1/4, pp. 106–124, 1927.

[72] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[73] I. Guellil and K. Boukhalfa, "Social big data mining: A survey focused on opinion mining and sentiments analysis," in *2015 12th international symposium on programming and systems (ISPS)*, pp. 1–10, IEEE, 2015.

[74] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," *Tepper School of Business*, p. 559, 2010.

[75] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Fourth international AAAI conference on weblogs and social media*, Citeseer, 2010.

[76] L. P. Del Bosque and S. E. Garza, "Aggressive text detection for cyberbullying," in *Mexican International Conference on Artificial Intelligence*, pp. 221–232, Springer, 2014.

[77] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *arXiv preprint arXiv:1703.04009*, 2017.

[78] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on twitter," in *Proceedings of the 2017 ACM on web science conference*, pp. 13–22, 2017.

[79] J. Chen, S. Yan, and K.-C. Wong, "Verbal aggression detection on twitter comments: Convolutional neural network for short-text sentiment analysis," *Neural Computing and Applications*, pp. 1–10, 2018.

[80] S. Modha, P. Majumder, and T. Mandl, "Filtering aggression from the multilingual social media feed," in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pp. 199–207, 2018.

[81] N. S. Samghabadi, D. Mave, S. Kar, and T. Solorio, "Ritual-uh at trac 2018 shared task: aggression identification," *arXiv preprint arXiv:1807.11712*, 2018.

[82] J. Risch and R. Krestel, "Aggression identification using deep learning and data augmentation," in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pp. 150–158, 2018.

[83] K. Raiyani, T. Gonçalves, P. Quaresma, and V. B. Nogueira, "Fully connected neural network with advance preprocessor to identify aggression over facebook and twitter," in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pp. 28–41, 2018.

[84] S. T. Aroyehun and A. Gelbukh, "Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling," in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pp. 90–97, 2018.

[85] P. Mishra, H. Yannakoudakis, and E. Shutova, "Tackling online abuse: A survey of automated abuse detection methods," *arXiv preprint arXiv:1908.06024*, 2019.

[86] J. M. Struß, M. Siegel, J. Ruppenhofer, M. Wiegand, M. Klenner, *et al.*, "Overview of germeval task 2, 2019 shared task on the identification of offensive language," 2019.

[87] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 75–86, 2019.

[88] R. Kumar, A. N. Reganti, A. Bhatia, and T. Maheshwari, "Aggression-annotated corpus of hindi-english code-mixed data," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[89] F. Ramiandrisoa and J. Mothe, "Irit at trac 2018," in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pp. 19–27, 2018.

[90] I. Arroyo-Fernández, D. Forest, J.-M. Torres-Moreno, M. Carrasco-Ruiz, T. Legeleux, and K. Joannette, "Cyberbullying detection task: the ebsi-lia-unam system (elu) at coling'18 trac-1," in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pp. 140–149, 2018.

[91] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[92] M. E. Peters, M. Neumann, L. Zettlemoyer, and W.-t. Yih, "Dissecting contextual word embeddings: Architecture and representation," *arXiv preprint arXiv:1808.08949*, 2018.

[93] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A bert-based transfer learning approach for hate speech detection in online social media," in *International Conference on Complex Networks and Their Applications*, pp. 928–940, Springer, 2019.

[94] A. Nikolov and V. Radivchev, "Nikolov-radivchev at semeval-2019 task 6: Offensive tweet classification with bert and ensembles," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 691–695, 2019.

[95] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," *Archives of suicide research*, vol. 14, no. 3, pp. 206–221, 2010.

[96] J. Culpeper, "Impoliteness: Using language to cause offence," *Impoliteness: Using Language to Cause Offence*, pp. 1–292, 01 2011.

[97] C. Baziotis, N. Pelekis, and C. Doulkeridis, "Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, (Vancouver, Canada), pp. 747–754, Association for Computational Linguistics, August 2017.

[98] H. Zhang, "Exploring conditions for the optimality of naive bayes," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 19, no. 02, pp. 183–198, 2005.

[99] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[100] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[101] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[102] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[103] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?," *arXiv preprint arXiv:1906.01502*, 2019.

[104] S. Madisetty and M. S. Desarkar, "Aggression detection in social media using deep neural networks," in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pp. 120–127, 2018.

[105] S. Swami, A. Khandelwal, V. Singh, S. S. Akhtar, and M. Shrivastava, "An english-hindi code-mixed corpus: Stance annotation and baseline system," *arXiv preprint arXiv:1805.11868*, 2018.

[106] B. C. Wallace, L. Kertz, E. Charniak, *et al.*, "Humans require context to infer ironic intent (so computers probably do, too)," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 512–516, 2014.

[107] D. Bamman and N. A. Smith, "Contextualized sarcasm detection on twitter," in *Ninth International AAAI Conference on Web and Social Media*, 2015.

[108] T. Miller, C. Hempelmann, and I. Gurevych, "SemEval-2017 task 7: Detection and interpretation of English puns," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, (Vancouver, Canada), pp. 58–68, Association for Computational Linguistics, Aug. 2017.

[109] E. Shutova and S. Teufel, "Metaphor corpus annotated for source-target domain mappings.," in *LREC*, vol. 2, pp. 2–2, 2010.

[110] A. Ghosh, G. Li, T. Veale, P. Rosso, E. Shutova, J. Barnden, and A. Reyes, "Semeval-2015 task 11: Sentiment analysis of figurative language in twitter," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 470–478, 2015.

[111] M. M. Bradley and P. J. Lang, "Affective norms for english words (anew): Instruction manual and affective ratings," tech. rep., Technical report C-1, the center for research in psychophysiology . . . , 1999.

[112] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 174–184, 2018.

[113] S. Rosenthal, N. Farra, and P. Nakov, "Semeval-2017 task 4: Sentiment analysis in twitter," *arXiv preprint arXiv:1912.00741*, 2019.