HostileNet: Multi-Label Hostile Post Detection in Hindi

by

Mohit Bhardwaj
MT19014

Under the Supervision of
Dr. Tanmoy Chakraborty,
Dr. Md. Shad Akhtar

Indraprastha Institute of Information Technology Delhi

July, 2021

HostileNet: Multi-Label Hostile Post Detection in Hindi

by

Mohit Bhardwaj
MT19014

Submitted
in partial fulfillment of the requirements for the degree of
Master of Technology in Computer Science & Engineering
(Specialization in AI)

to

Indraprastha Institute of Information Technology Delhi
July, 2021

Certificate

This is to certify that the thesis titled "**HostileNet: Multi-Label Hostile Post Detection in Hindi**" being submitted by **Mohit Bhardwaj** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under our supervision. In our opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree. The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

July, 2021

Dr. Tanmoy Chakraborty
Department of Computer Science & Engineering
Indraprastha Institute of Information Technology Delhi
New Delhi 110020

Dr. Md. Shad Akhtar
Department of Computer Science & Engineering
Indraprastha Institute of Information Technology Delhi
New Delhi 110020

Acknowledgements

I would like to express my sincere gratitude and indebtedness to Dr. Tanmoy Chakraborty for his exemplary guidance, supervision and constant encouragement throughout. I am grateful and remain indebted to Dr. Md Shad Akhtar for his constant support and his insightful comments and suggestions at every stage of the research project. I would also like to express my gratitude to Mr. Manjot Bedi and Ms. Megha Sundriyal for their continuous support and guidance. I would also like to thank members of LCS2 lab for being a constant source of motivation. Finally, I would like to thank my supportive family and friends who encouraged me and kept me motivated throughout the thesis.

Abstract

The swift escalation in hostile content on the web and specifically on Online Social Media (OSM) has lately become a matter of concern that we must tackle. The situation worsens to a whole different level with recent events such as COVID-19 pandemic, BLM, and #MeToo movements. Even though the existing systems address the problem of hostile post detection in one or more dimensions, e.g., hate, fake, etc., there has not been sufficient studies that address multiple hostile dimensions in a unified system. Moreover, a significant majority of the existing systems devour the English language, and research in regional languages (e.g., Hindi, Bengali, etc.) do not get adequate attention. To this end, in this paper, we tackle the hostile post detection in Hindi for four dimensions – *fake*, *hate*, *offensive*, and *defamation*. We propose HostileNet, a novel deep learning framework that leverages the HindiBERT-based contextual representations and hand-crafted lexicon features for the hostile post classification. Moreover, we also propose a novel mechanism to further fine-tune the attention vectors w.r.t. each hostile dimension. We evaluate HostileNet on the CONSTRAINT-2021's multi-label Hindi shared task dataset in both coarse-grained (hostile vs. non-hostile) and fine-grained (*fake* vs. *hate* vs. *offensive* vs. *defamation*) setups. Our evaluation shows that HostileNet outperforms various existing systems including the best performing system as reported in the CONSTRAINT-2021 shared task for both the setups. Furthermore, we provide a thorough analyses of the obtained results in forms of ablation study, error analysis, attention heapmap analysis, lexicon feature analysis, etc. We make the code and the curated multi-label hostile lexicon available for research use at https://github.com/LCS2-IIITD/HostileNet.git.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Technological progressions are going above and beyond, gradually sneaking into our lives. The growing fame of the Internet has radically expanded the usage of the OSM platforms. It has transformed into an extraordinary stage for individuals to impart their thoughts and opinions. A massive volume of text information disseminates over the internet, which contains a large portion of textual information consisting of posts shared by individuals on various topics like political issues, religious groups, the economy, and many more. The current outbreak of the COVID-19 pandemic has boosted the consumption of these OSM platforms to a great extent. According to a recent study, there has been an increment of 55% in total time spent by Indians on the Internet (BusinessToday 2020). This remarkable rise, in turn, has led to an increase in the amount of hostile content such as fake news, hate speech, and offensive posts over these platforms[1]. A recent survey inspects millions of websites, including most popular social media websites, and reports an escalation in hate speech by 900% against China and its people over Twitter[2]. A similar trend is observed against Asians, with over 200% increase in traffic on sites and posts that spread hate against Asians. Reichelmann et al. 2020 scrutinized the propagation of online hate speech in six nations and showed that hate speech has no boundaries.

1.1 Impact of Hostile Media

Most of the misinformation transmits through the OSM platforms (D. Joshi 2021). The principal intention behind the hostile content (e.g., fake news, hate speech, offensive posts, etc.) is to spread misinformation, embed fear into the minds of the public, defame someone, or spread hatred (Devakumar 2020). There are several instances where the spread of hostile content had impacted the entire society. During the 45^{th} US Presidential elections, around 25% of Americans visited a fake news website that tried to influence the thought process of the general public and affected the eventual outcome of the election (Grave et al. 2018).

In another example, a fake and defaming post in Bangladesh causes the destruction of several religious places of minority communities by a violent mob (Ahmed and Manik 2012). In the most recent example from India, a renowned celebrity's Twitter account was suspended permanently due to infraction of Twitter's hateful conduct and abusive behavior policies

[1] https://bit.ly/3rQQLTW
[2] https://l1ght.com/Toxicity_during_coronavirus_Report-L1ght.pdf

(Kanyal 2021). Similarly, many posts that spread hate or other hostile nature often lead to catastrophic impact as worst as the loss of human life. Such phenomena are not limited to a cultural, regional, or linguistic group; instead, these are global phenomena. Everyone is a victim of misuse of the social media platforms with such hostile posts. Considering the impact of such posts, timely detection and remedy are of utmost necessity to ensure a civilized environment.

1.2 Motivation

With the second highest number of Internet users in the world[3], India contributes to a dominant chunk of online content. Hindi is the most spoken language in India and 3^{rd} in the world with over 615 million speakers (Ghosh 2020).

Despite having a huge footfall, most of the languages in India still fall under the umbrella of low-resource languages and a significant effort is desirable across the NLP spectrum, and hostile post detection is not an exception. For the hostile post detection, a decent number of research has been carried out for English and other languages (Waseem and Hovy 2016; Davidson et al. 2017; Badjatiya et al. 2017; Mitrović, Birkeneder, and Granitzer 2019; Tran et al. 2020; Kaliyar, Goswami, and Narang 2021); however, the research involving Indian languages (e.g., Hindi) is scarce (Mandl et al. 2019; Kar et al. 2020; Saroj and Pal 2020). A prime reason is the unavailability of the qualitative dataset. Recently, a benchmark dataset (Bhardwaj et al. 2020) in Hindi covering four hostile dimensions, was developed as part of a shared task in the CONSTRAINT-2021 workshop (Patwa, Bhardwaj, et al. 2021). A good number of research teams participated in the competition and submitted their systems. Among the participating systems, we observe that their proposed systems fail to handle all hostile dimensions in equal proportion – there is no single system that reports best results for *all* four cases and the overall case. This prompts us to explore the hostile post detection for a unified solution instead of dimension-specific solutions. To this end, in this work, we propose a novel joint architecture, named HostileNet, to detect four hostile dimensions (i.e., *hate*, *fake*, *defamation*, and *offensive*) in Hindi. HostileNet utilizes label-wise gold attention scores to fine-tune HindiBERT's (Doiron 2020) attention heads to cater to each label specifically, alongside with other lexicon level features.

We evaluate HostileNet on the CONSTRAINT-2021 dataset (Bhardwaj et al. 2020) and compare the performance with the wining systems of the shared task. We study two setups; the first setup, *aka.* coarse-grained setup, identifies hostility in a social media post. The second one, i.e., the fine-grained setup, aims to reveal the presence/absence of four hostile dimensions. HostileNet yields better scores in both coarse-grained and fine-grained setups compared to the wining systems. We also report our analyses of the obtained results in further details.

[3] https://www.internetworldstats.com/top20.htm

1.3 Challenges

The rapid generation of malicious content and misinformation over OSM platforms, news feeds, and blogs have made the automatic identification of hostility an extremely perplexing task. It demands to increase the need for high computational models to tackle the problem of such posts. A hostile post is often decisively composed to spread misinformation, hatred and to mislead common mass. As an effect, it needs deep insight to interpret the hostility even for the human being. Another crucial challenge is the regional and cultural difference that affects how a group interprets a post. For example, the term '*meetha*' has a generic meaning of '*sweet*', but it is treated as '*fa##ot*' (hateful towards the LGBT community) in different culture. Similarly, in India, China, and many other countries, the word '*dog*' is treated as a derogatory and offensive term; however, it symbolizes friendly and loving pets in the majority of the western world.

Furthermore, fine-grained hostile post detection adds significant complexity in the identification process due to the eminent, diverse, yet overlapping characteristics of these subcategories (Bhardwaj et al. 2020). Hate speech and offensive speech are closely related and, in general, are used interchangeably by an ordinary person. The distinction lies in the motivation behind the hostility – hate posts are racial slur towards a group; whereas, offensive comments can be personal and may not target the group as a whole. Similarly, fake news is always false and malicious, whereas the allegation in defamation might be true but lacks proof and does not hold legal liability. At times, many claims are fake with the malicious intent of inciting hatred towards a particular community. Given these discussions, it is not difficult to understand that the discrimination among these hostile dimensions is a highly challenging task and needs careful investigation in an efficient identification model.

1.4 Our Contributions

We summarize our contributions as follows:

- We present a novel hostility detection dataset in Hindi language " CONSTRAINT-2021 " (Bhardwaj et al. 2020). We collect and manually annotate $\sim$ 8200 online posts. The annotated dataset covers four hostility dimensions: fake news, hate speech, offensive, and defamation posts, along with a non-hostile label.
- We address the problem of hostile post detection in social media posts in Hindi. This work explores two setups, coarse-grained hostile post detection as a binary classification and fine-grained hostile post detection as a multi-label multi-class classification problem.
- We propose a unified framework, HostileNet, to handle the identification of four hostile dimensions – *fake*, *hate*, *defamation*, and *offensive*.
- HostileNet incorporates a novel module to optimize the computed attention scores against the label-wise gold attention scores.

- Our evaluation on the CONSTRAINT-2021's dataset shows state-of-the-art performance for both fine-grained and course-grained setups and across all four dimensions.
- We report extensive analyses of HostileNet, including ablation analysis, feature analysis, heatmap analysis, error analysis, etc.

The rest of the work is organized as follows. In Chapter 2, we discuss the prominent work done in the field of hostile post detection. Chapter 3 encloses a brief description of the dataset. In Chapter 4, we shed light upon our proposed methodology. Chapter 5 consists of results and analyses of our proposed model, which is briefly followed by conclusion in Chapter 6.

Chapter 2

Related Work

The proliferation of hostile content on OSM platforms through daily feeds, news blogs, and online newspapers has made it an extremely challenging task to identify real and faithful content. Being one of the powerful and widely used OSM platforms, Twitter provides a perfect playground for individuals with malicious intentions to disseminate toxicity, misinformation, fake news, and hatred over the Internet. Upon perceiving the harmful effects of hostile content over OSM platforms, several studies have been made in the past three decades, which we briefly present below.

2.1 Hostile Text Detection in English

There are plenty of existing hostile post detection methods; most of them are devised for high-resource languages like English. The pioneering work in the hostile text detection was put forward by Spertus 1997. This study leveraged the traditional machine learning technique, namely Decision Tree, to detect hostile messages. Despite their approach to hostile post detection was straightforward, their work clutched a lot of attention and furnished a foundation for this challenging task. Later, a supervised learning approach with uni-grams to detect racism in tweets was proposed by Kwok and Y. Wang 2013. They employed a Naive Bayes (NB) classifier, leveraging acquired labeled data from different Twitter accounts to learn a binary classifier for the labels – "racist" and "non-racist". Most of the early attempts in hostile post detection were based on traditional machine learning methods focusing on predictive features; while in recent times, the study shifted towards the utilization of linguistic and syntactic features. Waseem and Hovy 2016 utilized character n-grams coupled with linguistic features for hate speech detection. Davidson et al. 2017 used SVM for multi-class classification of a tweet into "hate", "offensive" or "neither", employing tf-idf weighted n-grams and POS tag-grams. Additionally, the authors prolonged their feature set with multiple Twitter-specific handcrafted features such as the number of hashtags, URLs, user mentions, characters and words. Some studies likewise coupled linguistic, semantic, and syntactic features for detecting abusive words in text (Basile et al. 2019; Nobata et al. 2016). Surface-level features such as character n-grams, word n-grams, and word skip-grams are broadly used features for hostile post detection. To scrutinize the role of surface-level features, Malmasi and Zampieri 2018 carried out a study and argued that the surface-level features are insufficient to distinguish hate speech from profanity.

In recent times, numerous research trended towards the utilization of deep learning methods. Badjatiya et al. 2017 were the pioneer to employ deep learning for classifying a tweet as "racist", "sexist" or "neither". They used LSTM to learn tweet embeddings with gradient boosting. Upon the same task, Sajjad et al. 2019 practiced CNN trained over GloVe embeddings, alongside other ML-based handcrafted features with Logistic Regression classifier. Zampieri et al. 2019a spawned Offensive Language Identification Dataset (OLID), which comprised $14K$ English tweets for OffenEval 2019 Shared Task to detect, categorize and identify the target of offensive language. This dataset turned out to be another milestone in the hostile post detection task. They used CNN architecture as a baseline. The OffensEval 2019 Shared Task (Zampieri et al. 2019b) witnessed the shift towards deep learning-based systems for offensive language detection. The majority of the top-ranked teams used ensembles, CNN/RNN, and transformer-based models (Pelicon, Martinc, and Novak 2019; Doostmohammadi, Sameti, and Saffar 2019; Mitrović, Birkeneder, and Granitzer 2019). Though all the systems engineered around deep learning techniques perform exceptionally well, but these systems significantly lack interpretability.

The development of OSM has transformed drastically over the past few years. The escalating trend of OSM has led to an increasing amount of hostile content and fake news over the Internet. To tackle the challenging problem of automatic fake news detection, Karimi et al. 2018 proposed a Multi-source Multi-class Fake news Detection (MMFD) framework by amalgamating information from multiple sources with automatically elicited features.

Rasool et al. 2019 used a supervised multi-layered multi-label fake news detection where they continually learned to relabel the dataset correctly and performed the final evaluation on a hold-out test set. Shu, S. Wang, and H. Liu 2019 claimed that fake news is often intentionally fabricated to mislead users consequently; identifying fake news based solely on news content is not very reliable. To confront this, they put forward a tri-relationship embedding framework (TriFN) that modeled an association among the news publisher, news content, and users.

There have been a series of studies on COVID-19 misinformation detection. (Patwa, Sharma, et al. 2021) developed a new dataset. (Paka et al. 2021) proposed a cross-stitch based model for covid-19 fake news detection. They further extended their models using co-attention mechanism Bansal et al. 2021. (Deepak, Chakraborty, Long, et al. 2021) comprehensively covered various aspects of fake news including detection and diffusion. A series of studies are conducted on collusive fraud attacks (H. S. Dutta, Chetan, et al. 2018a; H. S. Dutta, Chetan, et al. 2018b; Chetan et al. 2019; H. S. Dutta and Chakraborty 2020b; H. S. Dutta, Jobanputra, et al. 2020; Arora et al. 2020; H. S. Dutta, V. R. Dutta, et al. 2020; H. S. Dutta and Chakraborty 2020a; H. S. Dutta, Diwan, and Chakraborty 2021; H. S. Dutta, Aggarwal, and Chakraborty 2021; H. S. Dutta, Arora, and Chakraborty 2021) and harmful memes (Pramanick, Sharma, et al. 2021; Pramanick, Dimitrov, et al. 2021)

Recently, the BERT model (Devlin et al. 2019) has gained tremendous attention. Tran et al. 2020 proposed HaBERTor model for detecting hate speech where they pre-trained BERT purely using 1.4M annotated hate speech comments. Parikh et al. 2019 was the first to work on multi-label detection of accounts of sexism. They used models like BERT, Universal Sentence Encoder for sentence representation and proposed a hierarchical combination of

BiLSTMs and CNNs over word embeddings.

2.2 Hostile Text Detection in Low-Resource Languages

Most literature in hostile post detection concentrates on high-resource languages; consequently, low-resource hostile post detection systems are sparse. Barely a handful of approaches have been investigated for low-resource languages like Hindi, Bangali, Urdu, etc. Ibrohim and Budi 2019 used SVM, Naive Bayes , and Random Forest for multi-label abusive and hate speech detection in Indonesian tweets. They proposed a multi-label classification approach leveraging transformation techniques like Binary Relevance (BR), Classifier Chains (CC), or Label Power Set (Kafrawy, Mausad, and Esmail 2016). Hossain et al. 2020 introduced a fake news detection system for Bengali that employed SVM encapsulated with other linguistic features. By the time development initiated for hostile post detection in low-resource languages, the deep learning era was already in the complete drive; hence, most researchers explored deep learning methods to tackle this issue. Mathur et al. 2018 utilized a multi-channel CNN-LSTM based architecture to classify offensive tweets in Hinglish (Hindi+English) language, pre-trained on English tweets. Likewise, Rizwan, Shakeel, and Karim 2020 proposed a hate speech and offensive language detection system in Romanized Urdu. They leveraged CNN-gram, which used different kernel sizes to learn patterns that are analogous to n-grams. The most recent inclination towards enhancing hostile post detection in regional languages is to conduct shared tasks – SemEval'19 Task 5 (Basile et al. 2019) and GermEval 2018 (Wiegand and Siegel 2018) are the offensive and hate speech detection tasks for Spanish and German, respectively.

2.2.1 Hostile Text Detection in Hindi

Despite being the 3^{rd} most spoken language in the world (Ghosh 2020), research in Hindi hostile post detection commenced quite late. Currently, CONSTRAINT-2021 (Patwa, Bhardwaj, et al. 2021) and HASOC (Mandl et al. 2019) marked as the most prominent shared tasks for hostile text detection in Hindi and grabbed a lot of attention from numerous researchers across the globe. However, majority of the approaches in both the shared tasks used engineered approaches, ensembles, or one vs all strategy for multi-label classification with little interpretability (Zhou, Li, and Ding 2021; Kamal, Kumar, and Vaidhya 2021; Raha et al. 2021; Bhatnagar et al. 2021; Gupta et al. 2021; Sarthak et al. 2021). Kar et al. 2020 utilized mBERT embeddings with Twitter user-level features for COVID-related fake news detection in Hindi and Bangla alongside English. They showed high efficacy in zero-shot learning among Hindi and Bengali due to their linguistic similarity as both are derived from the Indo-Aryan family of Indian languages.

2.3 Shortcomings of the Existing Systems

Identification of hostile content on the OSM platforms has gained a tremendous amount of interest in recent years. Numerous attempts have been made to tackle this challenging problem. They, however, had several pitfalls. The main shortcomings of existing systems are as follows:

- **Low-Resource Languages:** The most significant deficiency in the domain of hostile post detection is the lack of research in low-resource languages. India has the 2^{nd} highest number of internet users in the world[1], which shows how Indian languages will soon dominate a massive chunk of OSM platforms. Additionally, it is evident from the fact that now 50% of the tweets posted over Twitter by Indians are in non-English languages (Mandavia and Krishnan 2019). Yet, the majority of the Indian languages fall under the low-resource category. Ramchandra Joshi, Goel, and Raviraj Joshi 2020 showed how deep learning approaches like CNN, LSTM perform comparably to Bag-of-words in the case of Hindi language.

- **Hostile Post Detection in Single Dimension Only:** Even in high-resource languages, most existing systems work in only one dimension of hostility. Hence, we do not have systems that can efficiently learn correlations amongst various sub-categories of hostile content such as fake news, hate speech, offensive, and defaming posts (Basile et al. 2019; Nobata et al. 2016; Waseem and Hovy 2016; Davidson et al. 2017). Many of them use binary relevance, majority voting as an ensemble technique, or some engineering approach rather than a research-based approach for multi-label hostile post detection (Zhou, Li, and Ding 2021; Kamal, Kumar, and Vaidhya 2021; Raha et al. 2021; Bhatnagar et al. 2021; Gupta et al. 2021; Sarthak et al. 2021)

- **Less Interpretability:** Even if some multi-label systems tackle a subset of hostile categories, these models have significantly less reliability and interpretability as to why the model predicted a post as, say, hateful and defaming (Rasool et al. 2019; Parikh et al. 2019; Chalkidis et al. 2019; Tran et al. 2020).

[1] https://www.internetworldstats.com/top20.htm

Chapter 3

Dataset

Despite Hindi being the third most spoken language in the world, and a significant presence of Hindi content on social media platforms, to our surprise, we were not able to find any significant dataset on fake news or hate speech detection in Hindi. A survey of the literature suggest a few works related to hostile post detection in Hindi, such as (Kar et al. 2020; Jha et al. 2018; Safi Samghabadi et al. 2020); however, there are two basic issues with these works - either the number of samples in the dataset are not adequate or they cater to a specific dimension of the hostility only. In this report, we present our manually annotated dataset for hostile posts detection in Hindi (Bhardwaj et al. 2020). We collect more than $\sim$ 8200 online social media posts and annotate them as hostile and non-hostile posts. Furthermore, we identify four hostility dimensions for each hostile post as *fake*, *defamation*, *hate*, and *offensive*. Though some of these hostile dimensions sound similar at the abstract-level (e.g., *hate* and *offensive*), their definitions are different, and we define them below following (Mathur et al. 2018) and (Davidson et al. 2017).

- **Fake News**: A piece of information or an alleged claim that is verifiable to be false. The authors label posts that disseminates click-bait, satire, and parody content as fake news as well.
- **Hate Speech**: Any post that targets a specific individual/group of people based on their ethnicity, religious beliefs, geographical belonging, race, etc., with malicious intentions of disseminating hate or emboldening violence.
- **Offensive**: Any post which encompasses profane, impolite, rude, or vulgar language to insult an individual or group.
- **Defamation**: Any pernicious posts which disperse misinformation regarding an individual, or a group/organization, with the aim of destroying their reputation publicly.
- **Non-Hostile**: Posts that do not contain any hostile content.

The dataset development is part of the CONSTRAINT-2021 shared task (Patwa, Bhardwaj, et al. 2021). The CONSTRAINT-2021workshop emphasizes the hostility detection on three major points, i.e., low-resource regional languages, detection in emergency situations, and early detection task[1].

[1] https://constraint-shared-task-2021.github.io/

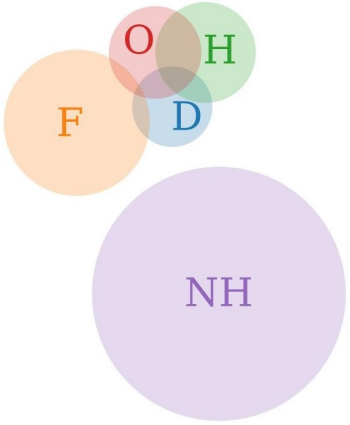


Figure 3.1: Venn Diagram of Multi-Label Hindi hostility dataset. Notations: [$F-$ *Fake*], [$O-$ *Offensive*], [$H-$ *Hate*], [$D-$ *Defamation*], [$NH-$ *Non-hostile*].

3.1 Data Development

During the development of the dataset, we observe that some of the posts overlap across some hostile dimensions; therefore, we adopt the idea of multi-label tags for each post. Figure 3.1 shows the class-wise overlaps among hostile dimensions in form of a Venn diagram[2]. Although it reveals the relationship among for the majority cases, it is inadequate to show the intersection between fake and hate class in 2D. Some of the examples from the dataset are presented in Table 3.1.

	Posts	**Label**
1.	#NonsenseModi चंद्रायन मिशन सक्सेस होने के साथ ही मोदी जी ने वहां जाकर डिस्टर्ब कर दिया था जिसके कारण वैज्ञानिकों के महीनों की मेहनत बेकार *#NonsenseModi Chandrayaan mission was about to succeed but Modi ji went there and disturbed them due to which months of hard work of scientists was wasted.*	Defamation, Hate, Offensive
2.	@Username यह स#ले जिहादी मानसिकता वालों की जांच तो बढ़िया ढंग से होनी चाहिए यह भी स#ला जावेद अख्तर और सलीम का भाई है यह स#ला नेपोटिज्म के खिलाफ नहीं बोलता है *Bl##dy Islamic militant groups @Username should be properly checked for mental illness He is also a bl##dy supporter of Javed Akhtar and Saleem He also doesn't bl##dy speak up against nepotism*	Hate, Offensive
3.	कांग्रेस मूल की कंगना रनौत बिहार चुनाव में भाजपा का प्रचार करेंगी! #NATIONALNEWS *Kangana Ranaut is alleged to be a Congress supporter but is also alleged to promote BJP's propaganda in Bihar elections! #NATIONALNEWS*	Fake
4.	कोराना दौर में प्लेसमेंट ऑफ़र वापस ले रही हैं कंपनियाँ, छात्र परेशान URL *Durning the Covid-19 pandemic, companies are revoking placement offers, which stresses out students URL*	Non-Hostile

Table 3.1: A few samples from the CONSTRAINT-2021 Hindi hostile dataset (Bhardwaj et al. 2020). We present the samples in the original Devanagari, and its English translation for readability.

3.2 Data Collection

We collect $\sim$ 8200 hostile and non-hostile texts from various social media platforms like Twitter, Facebook, WhatsApp, etc. We follow different strategies to collect data for each

[2] https://www.meta-chart.com/venn

category.

- For *fake news* collection, we refer to some of India's top most fact checking websites like BoomLive [3], Dainik Bhaskar [4], etc, and read numerous articles in Hindi. This process helps us identify the topics of the fake news. Subsequently, we compile a topic-wise keyword list for each fake news. Next, we curate online social media platforms such as Twitter, Instagram, etc., for the collection of posts.

- For *hate speech* collection, at first, we target the tweets encouraging violence against minorities based on their race, religious beliefs, etc. Following this process, we analyse the timelines of users with significant hate-related posts. Additionally, we also analyse users who liked or commented in support of the hate speech and scan their timelines for additional hate-related posts as well.

- For *offensive posts*, we employ the list of top swear words used in Hindi language as determined by Jha et al. 2018. For each swear word, we query Twitter API[5] to extract (offensive) tweets. In the next step, we manually verify each collected tweet as offensive. One critical observation that we make during the collection process is that offensive posts against women are more toxic and hate-oriented than the male counterpart.

- For the posts related to the *defamation* category, we read viral news articles where people or a group are publicly shamed due to misinformation, and perform topic-wise search to collect defamation tweets.

- To collect *non-hostile* data, we extract posts from some of the trusted sources (e.g., BBCHindi). We manually iterate over the collected samples to ensure that they are non-hostile in every way. Furthermore, we also annotate around 600 non-hostile texts from many non verified users with small followers count to maintain diversity in our dataset.

3.3 Dataset Stats

We present a brief statistics of the dataset in Table 3.2. Due to the overlapping nature of the hostile dimensions, a post can be labelled with multiple hostile dimensions. From Table 3.2, we observe that the hostile posts are not perfectly balanced across four dimensions – *defamation* has $\sim$ 50% samples in comparison to the *fake* class. Moreover, the skewness escalates with posts having single-dimension only. We observe that 37% of the samples are fake, while only 11%, 15%, and 17% of the samples are defaming, offensive, and hateful, respectively. We also list some samples from different classes in Table 3.1.

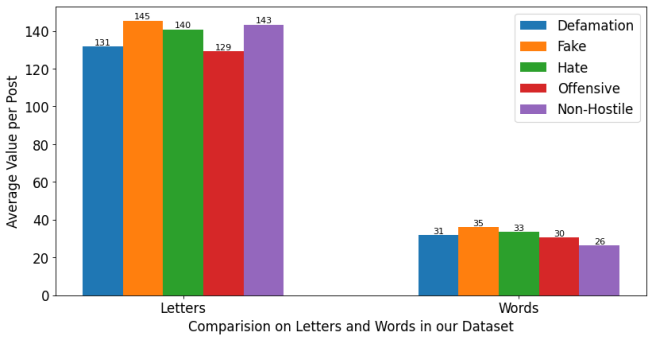
[3] https://hindi.boomlive.in/fake-news
[4] https://www.bhaskar.com/no-fake-news/
[5] https://developer.twitter.com/en/docs/twitter-api

Dataset	Hostile Posts					Non-Hostile
	Defamation	Fake	Hate	Offensive	Total*	
Train	564	1144	792	742	2678	3050
Validation	77	160	110	103	376	435
Test	169	334	234	219	780	873
Overall	810	1638	1136	1064	3834	4358

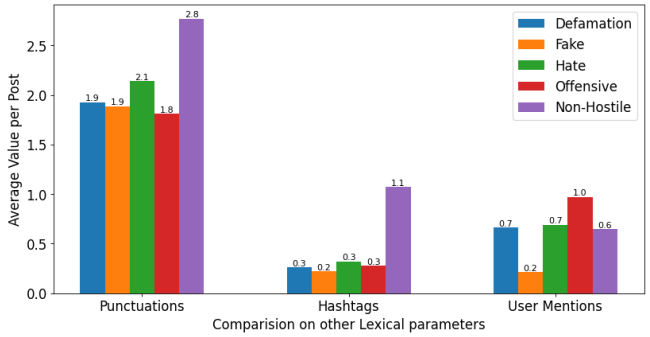
Table 3.2: CONSTRAINT-2021 Hindi Shared Task Dataset (Bhardwaj et al. 2020) description. Defamation, Fake, Hate and Offensive denotes the number of posts associated with these respective fine-grained hostile dimensions. Since this is a multi-label hostile dataset, * indicates the total count for hostile posts.

3.4 Dataset Analysis

On analyzing our dataset, we find multiple interesting patterns. Figure 3.2a shows the average number of letters and words per post across each hostile and non-hostile dimension. It is interesting to note that unlike other languages, in Hindi even though the non-hostile posts have a higher average number of letters per post, the average number of words in hostile posts is higher than the non-hostile posts. This might suggest that hostile posts contain more short words in place of long common words. Similarly from Figure 3.2b, we can observe that on average a non-hostile post has roughly 32% more punctuation marks than a hostile post, which suggests that people who spread hostile content bother less about the syntactic correctness of their content and more on the harmful aspect.



(a) Average number of characters and words per post.



(b) Average Punctuations (| , : ? _ ” ; !), Hashtags, and User Mentions per post.

Figure 3.2: Class-wise distribution.

We also show the word clouds[6] in hostile and non-hostile posts in Figures 3.3a and 3.3b, respectively. There are some common popular words which belong to both hostile and non-hostile categories. This is because words like Corona, Modi, Nation, and many more were over social media throughout our entire annotation process in all sorts of conversations. Still, the amount of negation and offensiveness against the ruling party, against religions, or countries like China, Pakistan is clearly visible in Figure 3.3a.

(a) Hostile Word Cloud (b) Non-Hostile Word Cloud

Figure 3.3: CONSTRAINT-2021 dataset Word clouds

Our analysis reveals that hostile content can be direct or indirect, i.e., whether the post is harmful to an individual/group or is in general (Waseem, Davidson, et al. 2017). On the similar lines, a hostile post could be either implicit or explicit as well. We observe that offensive posts have one user mention on average, reflecting that the dataset mainly consists of directed offensive content. For experiments, we follow the train, validation, and test split ratio of 70:10:20, respectively.

[6] https://www.wordclouds.com/

Chapter 4

Proposed Methodology

A high-level architecture of our proposed model, HostileNet is shown in Figure 4.1. It has three main components – a backbone network consisting of the HindiBERT (Doiron 2020) framework, a module to optimize the multi-label attention vectors, and a module to incorporate the lexicon and other handcrafted features. Given a Hindi tweet to HostileNet, we fine-tune HindiBERT. Moreover, during training, we compute attention vectors, one for each hostile dimension, and optimize the Kullback–Leibler (KL) divergence score between the computed attention vectors and the gold attention vectors. The objective is to learn the relevant and important tokens as close as to the training distribution. In parallel, we encode lexicon-specific features and fuse them into the network through concatenation. Finally, we employ a small multi-layer perceptron (MLP) network for the classification. Since we address the multi-label classification – one tweet can belong to more than one hostile dimension, we utilize four sigmoid neurons at the output layer and optimize the classification loss through binary cross-entropy. In the following section, we furnish details of each component in HostileNet.

4.1 HostileNet Architecture

Formally, let $p = \{w_1, w_2, ..., w_n\}$ be a post in the dataset consisting of n words. At first, we normalize the text. For this, we replace each emoticon in the post with its corresponding textual definition, e.g., we convert 🙏 to "*folded_hands*"[1]. Moreover, we tokenize the post using sentence piece tokenizer (Kudo and Richardson 2018) and subsequently pad the sequence up to T length for consistency among all posts.

4.1.1 Preprocessing

At first, we describe the compilation of multi-label lexicon for hostile texts. We utilize the lexicon for leveraging the hand-crafted features in HostileNet and to obtain the gold attention vectors for each hostile dimension.

- *Multi-label Lexicon Creation:* We summarize the lexicon creation process in Algorithm 1. Given a set of posts as input, the algorithm returns a multi-label lexicon dictionary, where a key is a valid token of the dataset and its value consists of a list of five normalized

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

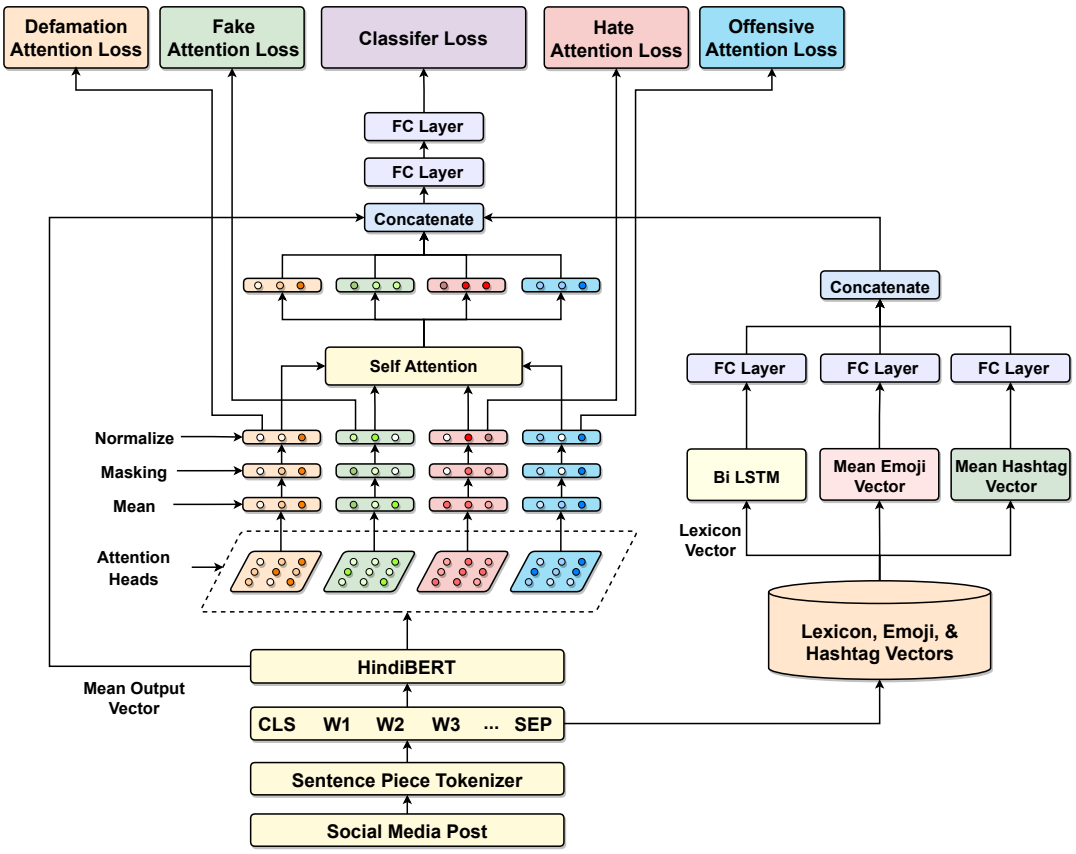


Figure 4.1: The architecture of HostileNet for the multi-label hostile post detection in Hindi.

scores that sum up to unity. These scores show the token's association with the four hostile dimensions – *defamation*, *fake*, *hate*, and *offensive*, and one *non-hostile* dimension, respectively. The l^{th} lexicon score for a token is a value in the interval $[0, 1]$, where a value close to 1 denotes strong association of the token towards the l^{th} label and vice-versa.

In the next step, we calculate the token frequency f_w^l for each token w in vocabulary V, against each label l, and normalize the counts by the total number of label counts for that token as follows:

$$f_w^l = \frac{f_w^l}{\sum_j f_w^j}$$

This allows us to minimize the effect of frequent and common words in computing a token's association with the hostile label. Further, we normalize the scores once again to handle the skewness in the dataset, e.g., a token may have higher frequency for a label than others due to imbalance dataset. Moreover, to ensure a good segregation amongst labels, we subtract the normalized score by the cumulative scores of all other labels. Finally, we compute the softmax function to obtain the absolute scores in the interval [0,1].

$$Lex_w^l = Softmax(f_w^l - \sum_{k \in L, k \neq l} f_w^k) \quad \forall l \in L$$

ALGORITHM 1: Multi-Label Lexicon Creation

Input: The set of all posts P, in the training set

Output: A multi-label lexicon score dictionary Lex, where a key can be any sentence piece tokenenized token from the dataset, and its value consists of a list of 5 scores, which shows the key's association within defamation, fake, hate, offensive, and non-hostile dimension

$L \leftarrow$ Total number of labels in the training set

$Lex \leftarrow$ Empty dictionary

$C \leftarrow \{C^1, C^2, ..., C^L\}$ ▷ *Number of training samples for each label*

for *each tokenized post* $p \in P$ **do**

  for *each token* $w \in p$ **do**

    if w *not in dictionary Lex* **then**

      $f_w^1 = f_w^2 = f_w^3 = f_w^4 = f_w^5 = 0$ ▷ f_w^l *is frequency of token* w *in label* l

      $Lex_w = [f_w^1, f_w^2, f_w^3, f_w^4, f_w^5]$ ▷ *Initialize list with zero frequencies*

    end

    ▷ *Increment frequencies associated with labels of post* p *by 1. For e.g. if post is defaming and fake, then increment* f_w^1 *and* f_w^2 *of* Lex_w *by 1*

  end

end

for *each token* $w \in Lex$ **do**

  for *each label* $l \in L$ **do**

    $f_w^l = \frac{f_w^l}{\sum_j f_w^j}$

  end

  for *each label* $l \in L$ **do**

    $f_w^l = \frac{f_w^l}{C^l}$

  end

  for *each label* $l \in L$ **do**

    $Lex_w^l = Softmax(f_w^l - \sum_{k \in L, k \neq l} f_w^k)$

  end

end

We compute the lexicon score for each token in the vocabulary except for the stopwords[2] and a few punctuation marks (',', '.', etc.). However, we keep tokens like '!', '?', etc. as they are highly correlated with all four hostile dimensions.

- *Gold Attention Vectors:* For each hostile dimension l, we create one gold attention vector g^l using the multi-label lexicon created in the previous step. Given an input post, we extract the l^{th} lexicon score for each token in the post and form a vector $v \in \mathbb{R}^T$ – the padded length. Next, we compute the gold attention vector by applying a softmax function over the vector v. These gold attention vectors ($\forall_{l \in L}\ g^l$) represent the segregation of a post in four hostile dimensions as they truly highlight the portion of the text contributing towards respective hostile labels.

4.1.2 Context Rich Representation

Efficient representation learning is one of the crucial aspects of any deep learning architecture for an NLP task. To obtain the hidden representation for each token $w_i \in p$, we employ a

[2] https://data.mendeley.com/datasets/bsr3frvvjc/1

pre-trained HindiBERT (Doiron 2020) model. It incorporates the Electra (Clark et al. 2020) architecture and has been pre-trained on 8GB of OSCAR common crawl dataset and 1GB of Wikipedia dataset in Hindi. We extract 256-dimensional embedding vectors to represent the post as the mean of T tokens.

4.1.3 Fine-tuning Attention Vectors

To learn the relevant and important tokens as close as to the training distribution, we fine-tune four attention vectors, one for each dimension, in HostileNet. We hypothesize that the association of one attention head per label will help the model cater specifically to each label. We utilize L attention heads of HindiBERT[3] corresponding to L hostile dimensions, with one attention head corresponds to one hostile dimension.

Let $A = \{A^1, A^2, ..., A^L\}$ be the set of query-key attention matrices of HindiBERT for post p. Let $A^l \in A$ represent the l^{th} query-key attention matrix for post p. We compute the mean over query attention scores and obtain an attention vector $a^l \in \mathbb{R}^T$. Subsequently, we mask the attention scores of Hindi stopwords, various non-relevant tokens, such as [CLS], [SEP], [PAD], etc. to obtain the masked attention vector $m^l \in \mathbb{R}^T$. Further, we normalize m^l to obtain $n^l \in \mathbb{R}^T$, by applying a masked softmax function. It allows us to redistribute the probability mass to the remaining tokens such that $\sum_t n_t^l = 1$ and still maintains 0 as the attention scores for all masked tokens.

Finally, we optimize the label-wise KL divergence loss between the gold attention vector g^l and normalized BERT attention scores n^l for every label l in a post s.

$$\mathbb{L}_{KD}^l(g^l||n^l) = \{k_1^l, k_2^l, ..., k_T^l\} \qquad \text{where} \qquad k_t^l = n_t^l(log(n_t^l) - g_t^l) \tag{4.1}$$

Moreover, for L labels, we have total KL divergence loss as:

$$\mathbb{L}_{KD}(g||n) = \sum_{l=1}^{L} \lambda_{KD}^l * \mathbb{L}_{KD}^l(g^l||n^l) \tag{4.2}$$

where λ_{KD}^l is a hyper-parameter to control label imbalance issue in KL divergence. This allows the model to tune each head pertinent to the respective hostile dimension.

4.1.4 Lexicon Features

To supplement neural network-based contextual representation, we incorporate multi-label hostile lexicon vectors computed through our Algorithm 1 and encode them through a BiLSTM layer. In addition, we encode hashtags and emoticons present in the input post to leverage their semantics in HostileNet. We combine these three vectors with the self-attended vectors of HostileNet for the final classification.

[3]The pre-trained HindiBERT has four attention heads by default; therefore, we associate each attention head to one hostile dimension in our case. Please note that, in case of more hostile dimensions, we can train HindiBERT with more attention heads.

- **Lexicon Embedding:** To create a context-aware lexicon embedding using our multi-label lexicon, we pass the tokenized post through a Bi-LSTM (Graves and Schmidhuber 2005) to get a set of hidden state vectors $h = \{h_1, h_2, ..., h_T\}$. We take the sum over all the hidden states h_t, $t \in [1, T]$. Then we pass it through a fully-connected layer to obtain the lexicon embedding for the post.

- **Emoticon Embedding:** We take the mean of the vector representations of all the emojis present in the input post using emoji2vec (Eisner et al. 2016) and pass it through a fully-connected layer.

- **Hashtag Embedding:** To incorporate hashtag information, we use Twitter's hashtag segmenter (Baziotis, Pelekis, and Doulkeridis 2017) to segment hashtags in the input post. For example, we segment the hashtag "#संजय_सिंह_गुंडा_है" (#Sanjay_Singh_Gunda_Hai | #Sanjay_Singh_is_Gangster) into a set of four words 'संजय' (Sanjay | Sanjay), 'सिंह' (Singh | Singh), 'गुंडा' (Gunda | Gangster), 'है' (Hai | Is). Then we use the multilingual IndicFT[4] (Kakwani et al. 2020) word embedding model to obtain 300 dimensional static representation for each segment of the hashtag. Finally, we take the mean of all segments obtained from all the hashtags in the post and pass them through a fully-connected layer to get the overall hashtag embedding for a post.

Final Prediction: Subsequent to the optimization of the attention vectors for each label, we fuse them through a self-attention mechanism followed by a concatenation operation. The concatenated vector along with the hand-crafted lexicon-based feature vector are fed to a multi-layered perceptron for the final classification. As mentioned earlier, the samples in the CONSTRAINTS dataset are of multi-label nature; therefore, we employ four sigmoid neurons with the binary cross-entropy (BCE) loss for the predictions. For optimizing HostileNet, we sum up BCE and KL divergence attention losses.

$$\mathbb{L}(s) = BCE(s) + \sum_{l=1}^{L} \lambda^l * \mathbb{L}^l_{KD}(g^l || n^l) \tag{4.3}$$

[4]FastText based word embedding model trained over English and 11 Indian languages.

Chapter 5

Experimental Results

In this section, we discuss our experimental results and report comparative analysis against various baselines. We also illustrate a thorough analyses of HostileNet's performance using ablation, and different choices of supervised attention losses for our model. We then demonstrate explainability and error analysis of our best model.

5.1 Baseline Models

Here, we define various existing systems that we employ for the comparative study. All these systems were part of the CONSTRAINT-2021 shared task challenge and ranked amongst the best systems.

- **CONSTRAINT Baseline** (Bhardwaj et al. 2020): The authors use multilingual BERT (Devlin et al. 2019) to extract contextual representation of a post, and train SVM (Hearst et al. 1998) for the classification. For each hostile dimension, one binary SVM classifier is trained in one-vs-all setup.

- **Albatross** (Bhatnagar et al. 2021): The authors use a two-step approach for the hostile posts classification. At first, a coarse-grained classifier is trained to segregate the hostile posts form non-hostile posts. Subsequently, another classifier is trained for each dimension. Finally, the predictions are accumulated through an ensemble technique. The authors primarily fine-tune BERT for the classification, except for the defamation class, where they use an SVM classifier.

- **Bestfit AI** (Sarthak et al. 2021): The authors use Relational Graph convolutional Networks (RGCN) (Schlichtkrull, Kipf, and Bloem 2018) and multilingual BERT's pooler output to capture the semantic and contextual knowledge, respectively. They take the concatenation of these embeddings and pass it through a series of fully-connected layers for the classification. To extract the semantic knowledge, the authors translate Hindi post to English and employ Spacy[1] to obtain the dependency tree. Subsequently, the dependency parse tree is used to create labelled directed graphs where each node in graphs represents a token in the input post.

[1] https://spacy.io/api/dependencyparser

- **Monolith** (Kamal, Kumar, and Vaidhya 2021): The authors utilize IndicBERT (Kakwani et al. 2020) to train a binary coarse-grained classifier for hostile post detection and four separate classifiers for the fine-grained classification. They combine the output of coarse-grained classification with each fine-grained model in order to instil general hostile information learned by their coarse-grained model.

- **IREL IIIT** (Raha et al. 2021): The authors utilize pre-trained IndicBERT (Kakwani et al. 2020) and further fine-tune IndicBERT using AllenAI's pre-training implementation[2] to generate contextual embeddings. They convert all emojis to their equivalent vector representations using Emoji2vec (Eisner et al. 2016). For hashtag embeddings, they use pre-trained IndicBERT directly.

- **Zeus** (Zhou, Li, and Ding 2021): The authors fine-tune five BERT classifiers and apply majority voting-based ensemble for the final predictions.

In addition to these systems, we also compare our results with three other systems (Quark, Fantastic Four, and Cean) presented at the workshop; however, they did not report their findings publicly. We take their numerical results from the CONSTRAINT-2021's shared task description paper (Patwa, Bhardwaj, et al. 2021).

5.2 Experimental Setup

We pad each tokenized post to a maximum of 128 tokens. We apply a dropout of 0.25 and run all experiments with batch size of 16. We use a learning rate of 0.0001 and train the model for maximum 50 epochs with early stopping criteria. We use Adam as the optimizer with a decay of 0.001 and linear scheduler with warm up. For the coarse-grained classification, we optimize binary cross entropy with the *hostile* class_weight as 1.13. Similarly, in fine grained classification, the class_weights are taken as 4.74, 2.34, 3.38, and 3.64 for the *defamation*, *fake*, *hate*, and *offensive* classes, respectively. In both cases, class_weight is calculated using $k/|l|$, where $|l|$ is the number of samples for the label l and k is the total number of samples in our training set.

5.3 Performance of HostileNet

We present our comprehensive result in Table 5.1 for both setups – coarse-grained and fine-grained tasks. In coarse-grained task, IREL IIIT (Raha et al. 2021) reports the best weighted F1-score of 97.16 in the CONSTRAINT-2021 shared task closely followed by the Albatross (Bhatnagar et al. 2021) model with weighted F1 of 97.10. In comparison, HostileNet yields slightly better score (97.52) than the wining system.

In the fine-grained setup, we report F1-scores for each hostile dimension along with the weighted-average F1-score. The top performing systems at the shared task are Zeus (45.52) (Zhou, Li, and Ding 2021), Bestfit AI (82.44) (Sarthak et al. 2021), IREL IIIT (59.78) (Raha

[2] https://github.com/allenai/dont-stop-pretraining

Model	Fine-Grained					Coarse-Grained
	Def F1	**Fake F1**	**Hate F1**	**Off F1**	**w-F1**	**w-F1**
Constraint Baseline	39.92	68.69	49.26	41.98	54.20	84.22
Quark	30.61	79.15	42.83	56.99	56.60	96.91
Albatross	42.80	81.40	49.69	56.49	61.11	97.10
Fantastic Four	43.29	78.64	56.64	57.04	62.06	96.67
Bestfit AI	31.54	82.44	58.56	58.95	62.21	96.61
Monolith	42.00	77.41	57.25	61.20	62.50	95.83
IREL IIIT	44.65	77.18	59.78	58.80	62.96	97.16
Cean	44.50	78.33	57.06	62.08	63.22	96.67
Zeus	45.52	81.22	59.10	58.97	64.40	96.07
HostileNet	**48.96**	**82.93**	60.14	61.02	**66.32**	**97.52**
(−) Hashtags	48.81	81.80	58.51	**62.16**	65.76	97.21
(−) Emoticons	45.74	82.35	**60.88**	59.17	65.31	96.24
(−) Lexicons	47.32	79.74	57.19	60.90	64.17	96.85
(−) Pretraining	44.44	80.86	58.98	58.82	64.02	96.55

Table 5.1: Results of our HostileNet architecture compared with top baselines on (Bhardwaj et al. 2020) dataset along with ablation results of HostileNet (last four rows of of the table). Hashtags, Emoticons, and Lexicons denote the lexicon features as described in Chapter 4.

et al. 2021), and Cean (62.08) (Patwa, Bhardwaj, et al. 2021) for the defamation, fake, hate, and offensive dimensions, respectively. In comparison, HostileNet obtains improved performances in defamation (48.96), fake (82.93), and hate (60.14) classes. Moreover, on average, HostileNet outperforms the best system by $\sim$ 2% – it reports 66.32 weighted F1-score compared to 64.40 of Zeus (Zhou, Li, and Ding 2021). Note that none of the top performing systems are consistent – they report best result for one dimension only even though they trained separate systems for each dimension. On the other hand, our proposed HostileNet is a unified system and achieves the state-of-the-art performances in three out of four dimensions – it reports comparative sores in the offensive dimension. Furthermore, it obtains the state-of-the-art performance in both the fine-grained and coarse-grained setups on average. Thus the obtained results signify the robustness of HostileNet in detecting four hostile dimensions.

5.4 Ablation Analysis

After establishing the efficacy of HostileNet, we perform a series of ablation study to understand the effect of various sub-modules in the architecture. We begin by removing the lexicon-based embeddings (hashtag, emoticon, and lexicon embeddings) from HostileNet in sequence. We report the ablation results at the lower part of Table 5.1. In fine-grained setup, we observe a decrease of 0.56 in weighted F1-score with the removal of hashtag embeddings from HostileNet. For the same setup, a drop of 0.4 is observed in case of coarse-grained. The drop in performance reflects the role of hashtags in influencing the information virality and social movement, as shown initially by (R. Wang, W. Liu, and Gao 2016).

	Attention Loss	**Fine-Grained**					**Coarse-Grained**
		Def F1	**Fake F1**	**Hate F1**	**Off F1**	**w-F1**	**w-F1**
HostileNet	$None$	46.81	81.39	58.31	58.15	64.31	96.36
	$\mathbb{L}_{MSE}(g^l, n^l)$	42.81	81.02	58.84	**61.06**	64.26	96.79
	$\mathbb{L}_{ASL}(g^l, n^l)$	46.59	80.96	59.79	60.08	64.92	96.55
	$\mathbb{L}_{KLD}(g^l \| \| n^l)$	**48.96**	**82.93**	**60.14**	61.02	**66.32**	**97.52**

Table 5.2: Results of our model with different choice of loss functions in order to tune HindiBERT's attention heads. MSE, ASL, and KD stands for Mean Squared Error, Asymmetric, and KL Divergence Loss functions. Here g^l and n^l are gold and normalized HindiBERT's attention vectors for label l respectively.

Subsequently, we ignore the emoticon embeddings and observe performance drops of 0.45 and 1 F1 points in the fine-grained and coarse-grained setups, respectively. In the next step, once again the performance drop is observed when we skip the lexicon embedding in HostileNet as well. Additionally, we also observe the effect of utilizing the pre-trained HindiBERT model on HostileNet's training in the last row of Table 5.1. Overall, the removal of lexicon-based features and pre-training have adverse effects on both fine-grained and coarse-grained setups with a significant drop of 2.30 and 1.28 weighted F1 points, respectively. The above ablation results cement our intuition of leveraging the lexicon-based features for improved learning of HostileNet.

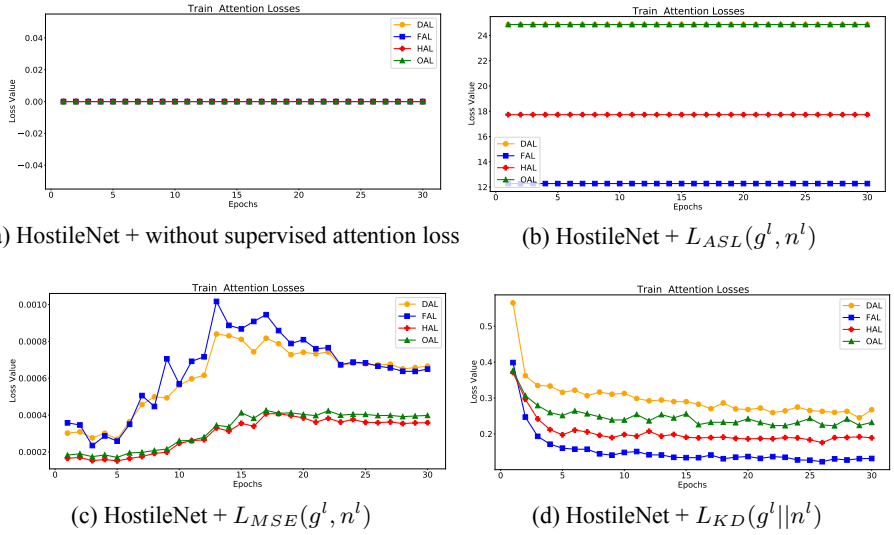
5.5 Supervised Attention Loss Analysis

In Table 5.2, we report our analysis of various loss functions that we employed to fine-tune the label-specific attention vectors. We experiment with optimizing mean-squared-error (MSE), asymmetric loss (ASL), and KL divergence (KLD) between the normalized HindiBERT's attention vectors (n^l) and the gold attention vectors (g^l). Moreover, we also experiment without optimizing the attention vector to provide support to the incorporation of fine-tuning the attention vectors.

It is evident from Table 5.2 and Figure 5.1 that KL divergence loss has the best effect on the learning of HostileNet in fine-grained setup followed by asymmetric loss – 66.32 F1-score with $\mathbb{L}_{KLD}$ in comparison with 64.92 F1-score with $\mathbb{L}_{ASL}$. Furthermore, in the absence of the optimization of attention vectors, HostileNet reports a performance degradation of 2 points in F1-score, thus supporting our claim that fine tuning attention vectors for each hostile dimension indeed has a positive effect on the overall performance. Moreover, we observe similar trend in the coarse-grained setup as well.

5.6 Pre-processing Analysis

To analyze the effect of pre-processing text on HindiBERT, we perform several experiments with slightly different pre-processing approaches on the input post before tokenization. All other hyperparameters and choice of loss functions remain fixed as mentioned in Section 5.2.



(a) HostileNet + without supervised attention loss

(b) HostileNet + $L_{ASL}(g^l, n^l)$

(c) HostileNet + $L_{MSE}(g^l, n^l)$

(d) HostileNet + $L_{KD}(g^l || n^l)$

Figure 5.1: HostileNet's supervised attention loss comparison. All the above experiments use P_{OP+EED} as the pre-processing approach. MSE, ASL, and KD stands for Mean Squared Error, Asymmetric, and KL Divergence Loss functions. Here g^l and n^l are gold and normalized BERT attention vectors for a label l. The legends DAL, FAL, HAL, and OAL stands for attention loss of defamation, fake, hate, and offensive classes respectively.

Since our algorithm creates lexicon vectors on the basis of tokenized corpus, for each pre-processing configuration, we have a slightly different set of tokens, lexicon scores, and gold attention vectors. Following is the explanation of all the pre-processing approaches:

- P_{OP}**:** In this approach we pass the original post (OP) as it is to HindiBERT.
- P_{OP+EED}**:** In this approach we embed the Emoticon's English description (EED) present within the original post (OP) instead of the emoticon itself. For example, if we have the emoji "🙏" present within the original post, then we will substitute it with "folded_hands" which is the description of the emoticon in English.
- $P_{OP-Emoticons}$**:** Here we remove all emoticons present in the original post and use this text as input to HindiBERT.
- P_{PP}**:** In this approach we remove all URLs, user mentions, and punctuation marks except '|' which is used to denote the end of sentence in Hindi.
- P_{PP+EED}**:** In this approach, we follow the same pre-processing steps mentioned in P_{PP}, and in addition we substitute an emoticon with their English description.
- $P_{PP-Emoticons}$**:** In this approach also we follow the same pre-processing steps mentioned in P_{PP} with the only difference being that we remove all emoticons completely from the input post.

As we can observe from Table 5.3 incorporation of textual description of emoticons in English gives us better results in both with and without pre-processing scenarios. Although

there is not much difference in weighted-F1 scores of with or without preprocessing approaches, HindiBERT model fails to produce interpretable outputs in case of all pre-processed approaches. We can infer that BERT has a hard time trying to fine-tune attention heads in case of pre-processed inputs as the attention loss barely reduces. This might be because BERT needs punctuation marks, user mentions etc. in order to capture the true context of the post.

Preprocessing Style	**Fine-Grained**					**Coarse-Grained**
	Def F1	**Fake F1**	**Hate F1**	**Off F1**	**w-F1**	**w-F1**
P_{OP}	47.60	80.23	59.61	60.11	64.80	97.33
P_{OP+EED}	48.96	**82.93**	60.14	61.02	**66.32**	**97.52**
$P_{OP-Emoticons}$	**49.23**	79.25	58.43	59.83	64.39	97.33
P_{PP}	49.21	81.04	57.00	60.39	64.79	97.09
P_{PP+EED}	46.96	82.05	**60.51**	**61.64**	65.89	96.79
$P_{PP-Emoticons}$	48.40	79.72	58.58	60.78	64.67	96.97

Table 5.3: Results of our model with different preprocessing approaches on input post. Here OP, PP, and EED stands for original post, preprocessed post, and embedded Emoticons in English description respectively. Substraction of Emoticons shows removal of emoticons from the entire post

5.7 Multi-label Lexicon Analysis

In this section, we present our analysis of the multi-label lexicon for each hostile dimension. For each token, we select the label with the maximum score; thus creating a list of tokens for each label. The number of tokens associated with the *defamation*, *fake*, *hate*, and *offensive* dimensions are 2652, 3646, 2146, and 2417, respectively. The remaining 4923 tokens correlate with the non-hostile dimension. Table 5.4 lists a few sampled tokens for each dimension. We present our observations for each label as follows:

- For the defamation dimension, we observe that a significant number of tokens revolve around politics – especially in terms of the two major and rival political parties of India (BJP and Congress). It correlates with the fact that supporters of these parties try to malign or defame each other.
- In case of fake news, we observe the presence of various country names, such as *India*, *China*, *Japan*, and COVID19-related terms. It could be because the dataset curation period Bhardwaj et al. 2020 overlaps with the early stage of the pandemic and comprises many unverified and fake news.
- Hate posts in India majorly revolve around the religious and casteism slurs. Our curation of hate lexicon rightfully captures such tokens (*Hindu*, *Muslim*, *Caste*, *Religious*, etc.) as listed in Table 5.4. Moreover, various political parties use terms such as '*foreigner*', and '*patriot*' in order to breed hate against some individual or a community.

Label	Lexicons in Hindi (English)
Defamation	भक्त (Devotee), आतंकी (Terrorist), फूल (Fool), बराबर (Equal), डूब (Drown), मरो (Die), जवाब (Answer), मोदीजी (Modiji), हिटलर (Hitler), कंगना (Kangana), बीजेपी (BJP), कांग्रेस (Congress), प्रवक्ता (Spokesman), बेटा (Son), बेटी (Daughter), चुनाव (Elections), भ्रष्ट (Corrupt), क्रांति (Revolution), ड्रामा (Drama)
Fake	पुलिस (Police), कोरोना (Corona), हमेशा (Always), और (more), भारत (India), जापान (Japan), चीन (China), छूट (Discount), रिकॉर्ड (Record), जल्द (Immediate), गिरफ्तार (Arrest), गाँधी (Gandhi), शक्तियों (Powers), विघटन (Dissolution), वक्तव्य (Statement), लहरा (Hoist)
Hate	हिन्दु (Hindu), मुसलमान (Muslim), धर्मप (Religious), जातियों (Castes), सरकार (Government), अधिकार (Rights), विरोध (Protest), संविधान (Constitution), नरक (Hell), हत्या (Killing), फांसी (Hanging), अभिमान (Pride), बर्बादी (Waste), अनौपचारिक (Informal), बख्तरबंद (Armored), सामूहिक (Collective), सता (haunting), देश (Country), देशभक्त (Patriot), विदेशी (Foreigner)
Offensive	स#ला (Rascal), कु#ता (Dog), कु#या (Bi##h), क###ने (Ra##al), बह###द (S###erF###er), मा####द (M###erF###er), तर्क (Argument), आपदा (Disaster), चर्चा (Discussion), काले (Black), मर्द (Male), भिख (Beg), टूट (Break), दलितों (Dalits), ताना (Taunt), बेचना (Sell), गोमांस (Beef), पागल (Mad), विफलता (Failure), गंदगी (Mess), लुटेरों (Robbers), ब्लाक (Block), साम्प्रदायिक (Communal)
Non-Hostile	बिजनेस (Business), अर्न (Earn), सितंबर (September), मेट्रो (Metro), रैल (Rail), फंक्शन (Function), मुस्कान (Smile), पाठक (Reader), प्रोफेसर (Professor), सैन्य (Military), उड़ान (Flight), सावधानी (Cautious), दोपहर (Afternoon), युवक (Youth), राष्ट्रीय (National), अंतर्राष्ट्रीय (International), बढ़ती (Increase), प्रकाश (Light)

Table 5.4: Mapping of tokens to the dimension having maximum lexicon score calculated using algorithm 1. We present the tokens in original Devanagari, followed by its English translation for readability.

- We observe that our algorithm correctly maps majority of the swear and slang words in the dataset, such as '*dog*'[3], '*bi##h*', '*ra##al*', '*s##erF##er*', '*m##erF##er*', etc., to offensive dimension. We also observe a few hateful words (e.g., '*black*') belong to the offensive dimension instead of hate speech. This could be because, skin-color racism is not very apparent in Indian context, and a common mass treats them as offensive rather than hateful.

- In case of the non-hostile category, most of the tokens are neutral in nature and simple day-to-day innocuous words such as *earn*, *reader*, *professor*, *afternoon*, *metro*, etc.

Overall, our analysis shows that the multi-label lexicon was able to capture the semantics of the tokens with reasonable precision and assist the model in improved performance.

5.8 Explainability using Tuned Attention Vectors

We also analyze the attention vectors as computed by HostileNet. Table 5.5 demonstrates the heatmaps for two test samples. In addition, we also report the gold attention scores for each hostile dimension for comparison. The ground-truth labels for the samples are [*defamation* and *fake*] and [*hate* and *offensive*], which HostileNet correctly predicts with attention tuning. In sample 1, we observe that HostileNet put greater attention on the words like 'आरोप' (Aarop | Blame), 'सोची समझी' (Sochi Samji | Though out), etc. for the defamation class and words like 'पुलवामा' (Pulwama | Pulwama) (related to Pulwama Attack 2019[4]) and 'अभिनंदन' (Abhinandan | Abhinandan), an Indian Air Force pilot who was held captive in Pakistan in counter strike, are very well highlighted for the fake class.

Similarly, in sample 2, more attention is given to the words 'दंगों' (Dango | Riots) and 'युद्ध' (Yudh | War) which goes on to show the provocative nature of this hateful post. On

[3] A derogatory term in hindi

[4] https://en.wikipedia.org/wiki/2019_Pulwama_attack

		Attention vector	Attention Heat Map
Sample 1	Defamation	Gold	अभिनंदन ने बीजेपी पर आरोप लगाते हुए बयान दिया है कि पुलवामा हमला बीजेपी की सोची समझी साजिश थी #FactCheck
		HostileNet w/ tuning	अभिनंदन ने बीजेपी पर आरोप लगाते हुए बयान दिया है कि पुलवामा हमला बीजेपी की सोची समझी साजिश थी #FactCheck
		HostileNet w/o tuning	अभिनंदन ने बीजेपी पर आरोप लगाते हुए बयान दिया है कि पुलवामा हमला बीजेपी की सोची समझी साजिश थी #FactCheck
	Fake	Gold	अभिनंदन ने बीजेपी पर आरोप लगाते हुए बयान दिया है कि पुलवामा हमला बीजेपी की सोची समझी साजिश थी #FactCheck
		HostileNet w/ tuning	अभिनंदन ने बीजेपी पर आरोप लगाते हुए बयान दिया है कि पुलवामा हमला बीजेपी की सोची समझी साजिश थी #FactCheck
		HostileNet w/o tuning	अभिनंदन ने बीजेपी पर आरोप लगाते हुए बयान दिया है कि पुलवामा हमला बीजेपी की सोची समझी साजिश थी #FactCheck
Sample 2	Hate	Gold	@Username ट्रम्प हमारे लोकतंत्र के लिए खतरा हैं ! इसके अलावा , अगर मेरा पक्ष किसी एक मुद्दे पर हार जाता है , तो मैं इस कु#या को बंद करने की योजना बनाता हूं ! दंगों ! गृह युद्ध !
		HostileNet w/ tuning	@Username ट्रम्प हमारे लोकतंत्र के लिए खतरा हैं ! इसके अलावा , अगर मेरा पक्ष किसी एक मुद्दे पर हार जाता है , तो मैं इस कु#या को बंद करने की योजना बनाता हूं ! दंगों ! गृह युद्ध !
		HostileNet w/o tuning	@Username ट्रम्प हमारे लोकतंत्र के लिए खतरा हैं ! इसके अलावा , अगर मेरा पक्ष किसी एक मुद्दे पर हार जाता है , तो मैं इस कु#या को बंद करने की योजना बनाता हूं ! दंगों ! गृह युद्ध !
	Offensive	Gold	@Username ट्रम्प हमारे लोकतंत्र के लिए खतरा हैं ! इसके अलावा , अगर मेरा पक्ष किसी एक मुद्दे पर हार जाता है , तो मैं इस कु#या को बंद करने की योजना बनाता हूं ! दंगों ! गृह युद्ध !
		HostileNet w/ tuning	@Username ट्रम्प हमारे लोकतंत्र के लिए खतरा हैं ! इसके अलावा , अगर मेरा पक्ष किसी एक मुद्दे पर हार जाता है , तो मैं इस कु#या को बंद करने की योजना बनाता हूं ! दंगों ! गृह युद्ध !
		HostileNet w/o tuning	@Username ट्रम्प हमारे लोकतंत्र के लिए खतरा हैं ! इसके अलावा , अगर मेरा पक्ष किसी एक मुद्दे पर हार जाता है , तो मैं इस कु#या को बंद करने की योजना बनाता हूं ! दंगों ! गृह युद्ध !

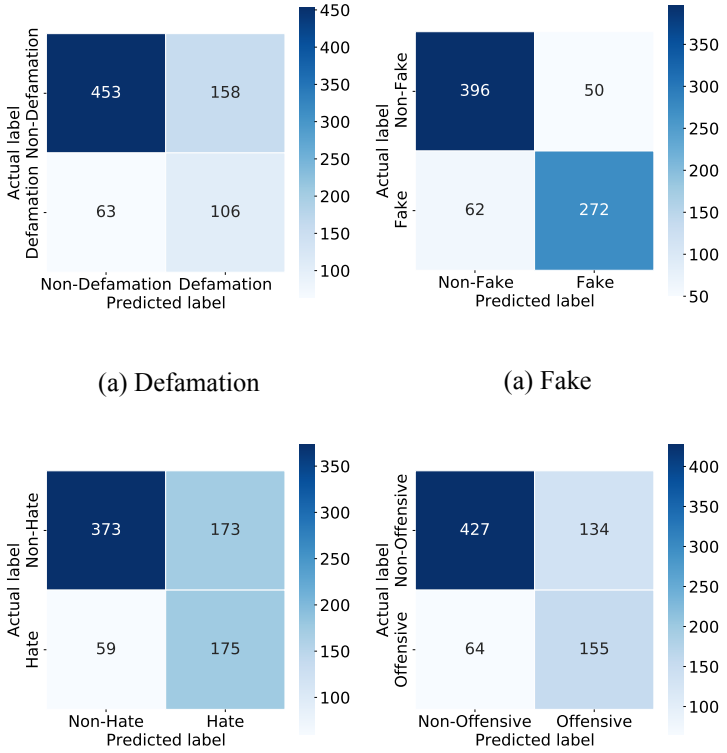
Sample 1 Gold label: [Defamation, Fake]; HostileNet w/ tuning: [Defamation, Fake]; HostileNet w/o tuning: [Defamation, Hate, Offensive];

Sample 2 Gold label: [Hate, Offensive]; HostileNet w/ tuning: [Hate, Offensive]; HostileNet w/o tuning: [Defamation, Hate, Offensive];

Table 5.5: Attention heatmaps for two hostile samples from the test set. For each dimension, we present the respective attention scores (darker shade represents higher weight) as computed by HostileNet with (w/) and without (w/o) tuning the attention vectors. We also report the gold attention vectors for each dimension for comparison. For the given samples, the ground-truth labels are [Defamation and Fake] and [Hate and Offensive] respectively. HostileNet, when subjected to tuning the attention vectors, predicts both the sample accurately. On the other hand, without tuning HostileNet misclassifies both the samples as Defamation, Hate, and Offensive. The learning of HostileNet is also evident from the heatmap as with (w/) tuning the model computes attention weights closer to the gold vectors in most of the cases. In comparison, without tuning, the model fails to assign appropriate attention scores in comparison with the gold attention scores.

the other hand, for offensive class, we notice that the word 'कु#या' (Ku##ia | B##ch) is the second most attended word after the username of the victim. In both cases, it can be further observed that the HostileNet's attention scores are very close to the gold attention scores. It suggests that the optimization of the KL divergence between the model's attention vectors and gold attention vectors facilitates the model to learn the relevant and important tokens as close as to the training distribution.

To further establish the efficiency of the attention vector tuning, we also present the heatmaps of the HostileNet's attention vectors without any fine-tuning (i.e., no optimization w.r.t. the gold attention vector). It is evident that the model without (w/o) tuning finds its difficult to attend to the relevant words in the post; hence, it fails to predicts the hostile dimensions correctly – it predicts [*defamation*, *hate*, and *offensive*] as the hostile labels for both the samples.



(a) Defamation (a) Fake

(b) Hate (a) Offensive

Figure 5.2: Confusion matrix plot across four hostile dimensions.

5.9 Error Analysis

In this section, we quantitatively and qualitatively analyze the errors committed by HostileNet. In Figure 5.2, we plot the confusion matrices for each hostile dimension. We observe that in all cases except for the *fake* label, the *false-positives* are significant. Consequently, it pulls down the F1-scores despite having high recall. Moreover, the precision for the *defamation* label is particularly low as the system reports higher *false-positives* than the *true-positives*. On the other hand, both *false-positives* and *false-negatives* are comparatively on the lower side in *fake news* detection; hence, HostileNet yields good F1-scores of 82.93. We relate the above phenomena with the available number of samples for each dimension in the dataset (c.f. Table 3.2) – *defamation* has the least number of samples (810), whereas, the *fake* samples are the highest (1638). We hypothesize that with more number of samples and relatively balanced data, HostileNet would perform even better.

Next, we move on to investigate some error cases of HostileNet and the best baseline system – IREL IIIT-H Raha et al. 2021 for the coarse-grained analysis and Zeus Zhou, Li, and Ding 2021 for the fine-grained analysis.

Table 5.6 presents the predictions and gold label for a few samples in the coarse-grained analysis. The first example is *non-hostile*; however, both HostileNet and the best baseline misclassify it as *hostile*. The possible reason might be the presence of the term 'बलिदान' (Balidaan | Oblation) in the post. Similarly, we observe misclassifications by both systems in

	Example	Gold	Prediction	
			HostileNet	**IREL IIIT-H**
1	आज हज़रत इमाम हुसैन साहब की बहादुरी और बलिदान को याद करते हुए हम सच्चाई और इंसाफ़ की राह पर चलने का संकल्प लेते हैं। Today, in the remembrance of bravery and sacrifice of Late Hazrat Imam Hussain, we take resolution to walk on the path of truth and justice.	Non-Hostile	Hostile	Hostile
2	मुहर्रम के लिए छूट है पर गणेश उत्सव में कुछ लोग मूर्ति विसर्जन को भी जाय तो पाबंदी । तेलंगाना सरकार को शर्म आनी चाहिए। @Username @Username @Username @Username URL There is leniency for Muharram but not for immersion of statues in river during Ganesh Chaturthi. Telegana government should be ashamed. @Username @Username @Username @Username URL	Hostile (Offensive)	Non-Hostile	Non-Hostile
3	18 जून से दिल्ली में राष्ट्रपति शासन लागू होगा और अगले चार सप्ताह के लिए पूरे दिल्ली एनसीआर में कम्पलीट लॉकडाउन रहेगा। है। President's rule will be implemented in Delhi from June 18 and there will be complete lockdown in entire Delhi NCR for the next four weeks.	Hostile (Fake)	Hostile	Non-Hostile
4	#AAP के @Username की बात तो कुछ हद तक विपक्ष ने मानी, राज्यसभा में किसान विरोधी बिलों का ज़बरदस्त विरोध भी किया, #AAP के @Username का विरोध तो बहुत ज़ोरदार था,पर मोदी सरकार ही क#नी है, वोटिंग के बग़ैर ही बिल पास करवा लिया , अब किसान ही भाजपा को फ़ेल करेंगे URL #AAP's @Username was accepted by the opposition to some extent, there was a strong opposition to the anti-farmer bills in Rajya Sabha. #AAP's opposition to @Username was very strong, but Modi government is a bi##h, got the bill passed without voting, now farmers fail BJP Got it passed, now it is the farmer who will fail the BJP URL	Hostile (Defamation, Fake, Offensive)	Hostile	Non-Hostile
5	पूर्व राष्ट्रपति PranabMukherjee का लंबी बीमारी के बाद निधन हो गया है. वो दिल्ली के RR अस्पताल में भर्ती थे, उनकी ब्रेन सर्जरी हुई थी और उनकी कोविड 19 रिपोर्ट भी पॉज़िटिव आई थी. Former President PranabMukherjee has passed away after prolonged illness. He was admitted to RR Hospital in Delhi, had undergone brain surgery and his COVID 19 report also came positive.	Non-Hostile	Non-Hostile	Hostile
6	बुद्धि व छमता अवसर मिलने पर दिखती है। पिछले वर्ष उप्र में होमिओपैथी में मेडिकल अफसरों की नियुक्ति में पिछडों की मेरिट 99% थी जबकि सवर्ण की 86%. Wisdom and versatility are seen when given the opportunity. In the previous year UP, the merit of the backwards in the appointment of medical officers in Homeopathy was 99% while that of the Savarnas was 86%.	Non-Hostile	Hostile	Non-Hostile

Table 5.6: Error analysis for coarse-grained hostile post classification using miss-classified examples by HostileNet and IREL IIIT-H Raha et al. 2021 (best coarse-grained hostile post detection baseline in CONSTRAINT-2021 Hindi shared task).

the second example as well – both systems tag the post as *non-hostile*. These two examples show that both systems fail to understand the *hostility* in the posts. In the next two cases, HostileNet correctly identifies the *hostile* lables in the posts; however, the baseline misclassifies both instances as *non-hostile*. It shows the inability of the baseline to comprehend the presence of the *offensive* word 'क#नी' (Kam##i | Bi##h). For the last example in Table 5.6, HostileNet wrongly tags the post as *hostile*; however, the baseline identifies the *non-hostile* nature of the post correctly.

We also report the fine-grained results obtained by HostileNet and the best performing baseline model, Zeus Zhou, Li, and Ding 2021, in Table 5.7. As expected, the predictions in the fine-grained setup in much more complex than the coarse-grained setup due to the multi-label classification. In majority of the cases, we observe that HostileNet makes atleast one correct prediction. In the first example, our model predicts *hate* as the correct label; however, it fails to recognize the *offensiveness* in the post. Moreover, it wrongly assigns the *defamation* tag to the post possibly due to the presence of the named-entity 'अम्बानी' (Ambani).

	Example	Gold	Prediction	
			HostileNet	**Zeus**
1	चलो अम्बानी के यहां पड़ा है.... कहीं बम बन के फटेगा तब भी नहीं....तुम लोग तो ऐसी जगह गिरवी पडयते हो जा के कि फिर उठोगे कि फटोगे कुछ पता नहीं रहता.... Come on, it (*for context, it refers to a car full of explosives*) lies at Ambani's place... even if a bomb will explode, not even you ... You guys are mortgaged to such a place that you will get up again that you will not know anything will explode….	Hate, Offensive	Hate, Defamation	Hate, Defamation
2	देखो देखो यह है मोदी दिखाई कुछ नहीं दे रहा है क्यों विकास गायब है Look look this is Modi nothing is visible because development is missing	Defamation	Defamation, Hate	Hate, Offensive
3	facebook.com/pram####46 ये प्रमोद सिंह जी की फेसबुक आईडी है जिस पर एक विडिओ डाला गया है जिसमे अबु आजमी और मुंबई पुलिस व साजिद भाई के समर्थन में नारे लगाए जा रहे है जिसे इन्होंने #पाकिस्तान जिंदाबाद बताया है जिससे गलत संदेश जाता है @Username इस पर कार्यवाही करे...@Username facebook.com/pram####46 This is the Facebook ID of Pramod Singh ji, on which a video has been put in which slogans are being raised in support of Abu Azmi and Mumbai Police and Sajid Bhai, which he described as #PakistanZindabad Wrong message goes @Username take action on this ... @ Username	Hate, Offensive	Hate	Non-Hostile
4	महाराष्ट्र में क्या हो रहा है ये तो मुझे नही पता पर हां ये जरूर पता है वहां अब हिंदुत्व खतरे में है शिवसेना और उद्धव ठाकरे ने हिंदुत्व से समझौता कर लिया है। @Username भगवान उद्धव को सद्बुद्धि प्रदान करे I do not know what is happening in Maharashtra, but yes I do know that now Hindutva is in danger, Shiv Sena and Uddhav Thackeray have compromised with Hindutva. @Username may god bless uddhav	Defamation, Hate	Defamation, Hate	Hate, Offensive
5	"पुलवामा हमला बीजेपी की सोची समझी सजीश थी" वाली राजनितीक सजीश 'कॉंग्रेस' को नये अस्त्र की तरकीब आजमानी चाहीए । The political conspiracy of 'Congress' with "Pulwama attack was a well thought out conspiracy of BJP". They should try the idea of a new weapon.	Defamation, Hate	Defamation, Hate	Fake
6	अक्षय कुमार की पत्नी उतरी रिया चक्रवर्ती के समर्थन मे अंध भक्तों अब तुमरा धर्म संकट में आ गए हो Akshay kumar's wife supports Rhea Chakraborty, now the religion of blind devotee's is in problem	Defamation	Defamation	Fake

Table 5.7: Error analysis for multi-label fine-grained hostile post classification using miss-classified examples by HostileNet and Zeus Zhou, Li, and Ding 2021 (best fine-grained hostile post detection baseline in CONSTRAINT-2021 Hindi shared task Patwa, Bhardwaj, et al. 2021).

We observe the similar behaviour for Zeus as well. Similarly, HostileNet makes one correct (*defamation*) and one incorrect (*hate*) prediction in the second example; whereas, Zeus predicts two incorrect (*hate* and *offensive*) labels and fails to identify the *defamation* class. In both cases, HostileNet predicts one extra class – *defamation* in the first example (precision and recall@50% each) and *hate* in the second example (precision@100% and recall@50%). On the other hand, HostileNet takes a conservative approach and predicts one correct class only (precision@100% and recall@50%) in the third example. It is interesting to note that Zeus marks the post as non-hostile. We also report a few other cases where HostileNet performs better than the baseline system.

Chapter 6

Conclusion

In this work, we verged upon hostile post detection on OSM in low-resource language – Hindi. We presented a state-of-the-art deep neural network architecture, HostileNet, for hostile post detection for four dimensions – *fake*, *hate*, *offensive*, and *defamation*. To captivate effective associations of a token within each hostile dimension, we proposed a novel multi-label lexicon scoring algorithm. To the best of our knowledge, this was the first time a study practically endeavored to characterize lexicon-based scores for the multi-label hostile post detection in low-resource language. Experiments illustrated the superiority of our proposed model HostileNet juxtaposed against various existing state-of-the-art systems. We experimentally showed an improvement of 0.36% and 1.92% in the weighted-F1 score for the coarse-grained and fine-grained hostile post detection tasks, respectively, over the best performing baseline systems. Furthermore, we visualized and illustrated the robustness and interpretability of HostileNet through attention heatmap analysis and token's association score for each dimension. We observed that HostileNet with attention fine tuning attends to relevant tokens corresponding to the associated hostile dimension. Our analysis also showed that the attention scores as computed by HostileNet during inference time is closely aligned with the gold attention scores.

Furthermore, to qualitatively appraise the performance of HostileNet, we conducted an exhaustive error analysis and compared the outcome against the existing state-of-the-art coarse-grained and fine-grained hostile post detection systems. In most cases, we observed that HostileNet yielded better prediction and made at least one correct prediction for the majority of the posts. In contrast, the best baseline systems often failed to comprehend the underlying hostility in a social media post.

Our analysis showed that HostileNet performed comparatively well for the majority (fake) class than the minority (defamation) class. Therefore, our future work would involve improving the performance of the minority class as well increasing the size of the dataset. Also, we plan to extend hostile post detection for other low-resource languages such as Bengali and Marathi.

Bibliography

Ahmed, Inam and Julfikar Ali Manik (2012). *Attacks on Buddhist Templates A hazy picture appears*. https://www.thedailystar.net/news-detail-252212.

Arora, Udit et al. (2020). "Analyzing and Detecting Collusive Users Involved in Blackmarket Retweeting Activities". In: *ACM Trans. Intell. Syst. Technol.* 11.3, 35:1–35:24. doi: 10.1145/3380537. url: https://doi.org/10.1145/3380537.

Badjatiya, Pinkesh et al. (2017). "Deep learning for hate speech detection in tweets". In: *Proceedings of the 26th International Conference on World Wide Web Companion*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, pp. 759–760.

Bansal, Rachit et al. (2021). "Combining exogenous and endogenous signals with a semi-supervised co-attention network for early detection of COVID-19 fake tweets". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp. 188–200.

Basile, Valerio et al. (June 2019). "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 54–63. doi: 10.18653/v1/S19-2007. url: https://www.aclweb.org/anthology/S19-2007.

Baziotis, Christos, Nikos Pelekis, and Christos Doulkeridis (Aug. 2017). "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: Association for Computational Linguistics, pp. 747–754. doi: 10.18653/v1/S17-2126. url: https://www.aclweb.org/anthology/S17-2126.

Bhardwaj, Mohit et al. (2020). "Hostility Detection Dataset in Hindi". In: *ArXiv* abs/2011.03588.

Bhatnagar, Varad et al. (2021). "Divide and Conquer: An Ensemble Approach for Hostile Post Detection in Hindi". In: *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Cham: Springer International Publishing, pp. 244–255. isbn: 978-3-030-73696-5.

BusinessToday (2020). *Average time spent on smartphone up 25% to almost 7 hours amid pandemic: Vivo report*. https://www.businesstoday.in/technology/news/average-time-spent-on-smartphone-up-25-percent-to-almost-7-hours-amid-pandemic-vivo-report/story/424790.html.

Chalkidis, Ilias et al. (July 2019). "Large-Scale Multi-Label Text Classification on EU Legislation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 6314–6322. doi: 10.18653/v1/P19-1636. url: https://www.aclweb.org/anthology/P19-1636.

Chetan, Aditya et al. (2019). "CoReRank: Ranking to Detect Users Involved in Blackmarket-Based Collusive Retweeting Activities". In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*. Ed. by J. Shane Culpepper et al. ACM, pp. 330–338. doi: 10.1145/3289600.3291010. url: https://doi.org/10.1145/3289600.3291010.

Clark, Kevin et al. (2020). "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators". In: *ArXiv* abs/2003.10555.

Davidson, Thomas et al. (2017). "Automated hate speech detection and the problem of offensive language". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. 1.

Deepak, P, Tanmoy Chakraborty, Cheng Long, et al. (2021). *Data Science for Fake News: Surveys and Perspectives*. Vol. 42. Springer Nature.

Devakumar, Delan (2020). *Racism and discrimination in COVID-19 responses*. https://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736(20)30792-3.pdf.

Devlin, Jacob et al. (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. doi: 10.18653/v1/N19-1423. url: https://www.aclweb.org/anthology/N19-1423.

Doiron, Nick (2020). *Monsoon NLP - Hindi BERT*. https://huggingface.co/monsoon-nlp/hindi-bert.

Doostmohammadi, Ehsan, Hossein Sameti, and Ali Saffar (June 2019). "Ghmerti at SemEval-2019 Task 6: A Deep Word- and Character-based Approach to Offensive Language Identification". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 617–621. doi: 10.18653/v1/S19-2110. url: https://www.aclweb.org/anthology/S19-2110.

Dutta, Hridoy Sankar, Kartik Aggarwal, and Tanmoy Chakraborty (2021). "DECIFE: Detecting Collusive Users Involved in Blackmarket Following Services on Twitter". In: *CoRR* abs/2107.11697. arXiv: 2107.11697. url: https://arxiv.org/abs/2107.11697.

Dutta, Hridoy Sankar, Udit Arora, and Tanmoy Chakraborty (2021). "ABOME: A Multi-platform Data Repository of Artificially Boosted Online Media Entities". In: *CoRR* abs/2103.15250. arXiv: 2103.15250. url: https://arxiv.org/abs/2103.15250.

Dutta, Hridoy Sankar and Tanmoy Chakraborty (2020a). "Blackmarket-Driven Collusion Among Retweeters-Analysis, Detection, and Characterization". In: *IEEE Trans. Inf.*

Forensics Secur. 15, pp. 1935–1944. doi: 10.1109/TIFS.2019.2953331. url: https://doi.org/10.1109/TIFS.2019.2953331.

Dutta, Hridoy Sankar and Tanmoy Chakraborty (2020b). "Blackmarket-driven Collusion on Online Media: A Survey". In: *CoRR* abs/2008.13102. arXiv: 2008.13102. url: https://arxiv.org/abs/2008.13102.

Dutta, Hridoy Sankar, Aditya Chetan, et al. (2018a). "Retweet Us, We Will Retweet You: Spotting Collusive Retweeters Involved in Blackmarket Services". In: *CoRR* abs/1806.08979. arXiv: 1806.08979. url: http://arxiv.org/abs/1806.08979.

— (2018b). "Retweet Us, We will Retweet You: Spotting Collusive Retweeters Involved in Blackmarket Services". In: *IEEE/ACM 2018 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018, Barcelona, Spain, August 28-31, 2018*. Ed. by Ulrik Brandes, Chandan Reddy, and Andrea Tagarelli. IEEE Computer Society, pp. 242–249. doi: 10.1109/ASONAM.2018.8508801. url: https://doi.org/10.1109/ASONAM.2018.8508801.

Dutta, Hridoy Sankar, Nirav Diwan, and Tanmoy Chakraborty (2021). "Weakening the Inner Strength: Spotting Core Collusive Users in YouTube Blackmarket Network". In: *CoRR* abs/2111.14086. arXiv: 2111.14086. url: https://arxiv.org/abs/2111.14086.

Dutta, Hridoy Sankar, Vishal Raj Dutta, et al. (2020). "HawkesEye: Detecting Fake Retweeters Using Hawkes Process and Topic Modeling". In: *IEEE Trans. Inf. Forensics Secur.* 15, pp. 2667–2678. doi: 10.1109/TIFS.2020.2970601. url: https://doi.org/10.1109/TIFS.2020.2970601.

Dutta, Hridoy Sankar, Mayank Jobanputra, et al. (2020). "Detecting and analyzing collusive entities on YouTube". In: *CoRR* abs/2005.06243. arXiv: 2005.06243. url: https://arxiv.org/abs/2005.06243.

Eisner, Ben et al. (Nov. 2016). "emoji2vec: Learning Emoji Representations from their Description". In: *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*. Austin, TX, USA: Association for Computational Linguistics, pp. 48–54. doi: 10.18653/v1/W16-6208. url: https://www.aclweb.org/anthology/W16-6208.

Ghosh, Iman (2020). *Ranked: The 100 Most Spoken Languages Around the World*. https://www.visualcapitalist.com/100-most-spoken-languages/.

Grave, Edouard et al. (2018). "Learning word vectors for 157 languages". In: *arXiv preprint arXiv:1802.06893*.

Graves, A. and J. Schmidhuber (2005). "Framewise phoneme classification with bidirectional LSTM and other neural network architectures". In: *Neural networks : the official journal of the International Neural Network Society* 18 5-6, pp. 602–10.

Gupta, Ayush et al. (2021). "Hostility Detection and Covid-19 Fake News Detection in Social Media". In: *ArXiv* abs/2101.05953.

Hearst, Marti A. et al. (1998). "Support vector machines". In: *IEEE Intelligent Systems and their applications* 13.4, pp. 18–28.

Hossain, Md Zobaer et al. (May 2020). "BanFakeNews: A Dataset for Detecting Fake News in Bangla". In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 2862–2871. url: https://www.aclweb.org/anthology/2020.lrec-1.349.

Ibrohim, Muhammad Okky and Indra Budi (Aug. 2019). "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter". In: *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, pp. 46–57. doi: 10.18653/v1/W19-3506. url: https://www.aclweb.org/anthology/W19-3506.

Jha, V. et al. (2018). "Hindi Language Stop Words List". In.

Joshi, Devashree (Jan. 2021). "COVID-19 Infodemic: Analysis of the Spread and Reach of Misinformation". In: *International Journal of Recent Technology and Engineering* 9, pp. 195–201. doi: 10.35940/ijrte.E5260.019521.

Joshi, Ramchandra, Purvi Goel, and Raviraj Joshi (2020). "Deep Learning for Hindi Text Classification: A Comparison". In: *Proceedings of the Intelligent Human Computer Interaction*. Cham: Springer International Publishing, pp. 94–101. isbn: 978-3-030-44689-5.

Kafrawy, Passent El, Amr Mausad, and Heba Esmail (2016). "Experimental Comparison of Methods for Multi-label Classification in different Application Domains". In: *International Journal of Computer Applications* 114, pp. 1–9.

Kakwani, Divyanshu et al. (Nov. 2020). "IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 4948–4961. doi: 10.18653/v1/2020.findings-emnlp.445. url: https://www.aclweb.org/anthology/2020.findings-emnlp.445.

Kaliyar, Rohit Kumar, Anurag Goswami, and Pratik Narang (2021). "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach". In: *Multimedia Tools and Applications* 80.8, pp. 11765–11788.

Kamal, Ojasv, Adarsh Kumar, and Tejas Vaidhya (2021). "Hostility Detection in Hindi Leveraging Pre-trained Language Models". In: *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Cham: Springer International Publishing, pp. 213–223. isbn: 978-3-030-73696-5.

Kanyal, Jyoti (2021). *Kangana Ranaut's Twitter account permanently suspended for violating rules*. https://www.indiatoday.in/movies/celebrities/story/kangana-ranaut-s-twitter-account-suspended-for-violating-rules-1798655-2021-05-04.

Kar, Debanjana et al. (2020). "No Rumours Please! A Multi-Indic-Lingual Approach for COVID Fake-Tweet Detection". In: *ArXiv* abs/2010.06906.

Karimi, Hamid et al. (2018). "Multi-source multi-class fake news detection". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1546–1557.

Kudo, Taku and John Richardson (Nov. 2018). "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, pp. 66–71. doi: 10.18653/v1/D18-2012. url: https://www.aclweb.org/anthology/D18-2012.

Kwok, Irene and Yuzhou Wang (2013). "Locate the hate: Detecting tweets against blacks". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 27. 1.

Malmasi, S. and Marcos Zampieri (2018). "Challenges in discriminating profanity from hate speech". In: *Journal of Experimental & Theoretical Artificial Intelligence* 30, pp. 187–202.

Mandavia, Megha and Raghu Krishnan (2019). *"Non-English tweets are now 50% of the total: Twitter India MD"*. https://economictimes.indiatimes.com/industry/tech/non-english-tweets-are-now-50-of-the-total-twitter-india-md/articleshow/72000048.cms.

Mandl, Thomas et al. (2019). "Overview of the HASOC Track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages". In: *Proceedings of the 11th Forum for Information Retrieval Evaluation*. New York, NY, USA: Association for Computing Machinery, pp. 14–17. isbn: 9781450377508. doi: 10.1145/3368567.3368584. url: https://doi.org/10.1145/3368567.3368584.

Mathur, Puneet et al. (Oct. 2018). "Did you offend me? Classification of Offensive Tweets in Hinglish Language". In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, pp. 138–148. doi: 10.18653/v1/W18-5118. url: https://www.aclweb.org/anthology/W18-5118.

Mitrović, Jelena, Bastian Birkeneder, and Michael Granitzer (June 2019). "nlpUP at SemEval-2019 Task 6: A Deep Neural Language Model for Offensive Language Detection". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 722–726. doi: 10.18653/v1/S19-2127. url: https://www.aclweb.org/anthology/S19-2127.

Nobata, Chikashi et al. (2016). "Abusive language detection in online user content". In: *Proceedings of the 25th International Conference on World Wide Web*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, pp. 145–153.

Paka, William Scott et al. (2021). "Cross-SEAN: A cross-stitch semi-supervised neural attention model for COVID-19 fake news detection". In: *Applied Soft Computing* 107, p. 107393.

Parikh, Pulkit et al. (Nov. 2019). "Multi-label Categorization of Accounts of Sexism using a Neural Framework". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 1642–1652. doi: 10.18653/v1/D19-1174. url: https://www.aclweb.org/anthology/D19-1174.

Patwa, Parth, Mohit Bhardwaj, et al. (2021). "Overview of CONSTRAINT 2021 Shared Tasks: Detecting English COVID-19 Fake News and Hindi Hostile Posts". In: *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Cham: Springer International Publishing, pp. 42–53.

Patwa, Parth, Shivam Sharma, et al. (2021). "Fighting an infodemic: Covid-19 fake news dataset". In: *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Springer, pp. 21–29.

Pelicon, Andraž, Matej Martinc, and Petra Kralj Novak (2019). "Embeddia at semeval-2019 task 6: Detecting hate with neural network and transfer learning approaches". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 604–610.

Pramanick, Shraman, Dimitar Dimitrov, et al. (2021). "Detecting Harmful Memes and Their Targets". In: *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*. Ed. by Chengqing Zong et al. Vol. ACL/IJCNLP 2021. Findings of ACL. Association for Computational Linguistics, pp. 2783–2796. doi: 10.18653/v1/2021.findings-acl.246. url: https://doi.org/10.18653/v1/2021.findings-acl.246.

Pramanick, Shraman, Shivam Sharma, et al. (2021). "MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets". In: *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*. Ed. by Marie-Francine Moens et al. Association for Computational Linguistics, pp. 4439–4455. doi: 10.18653/v1/2021.findings-emnlp.379. url: https://doi.org/10.18653/v1/2021.findings-emnlp.379.

Raha, Tathagata et al. (2021). "Task Adaptive Pretraining of Transformers for Hostility Detection". In: *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Cham: Springer International Publishing, pp. 236–243. isbn: 978-3-030-73696-5.

Rasool, Tayyaba et al. (2019). "Multi-Label Fake News Detection Using Multi-Layered Supervised Learning". In: *Proceedings of the 2019 11th International Conference on Computer and Automation Engineering*. Perth, WN, Australia: Association for Computing Machinery, pp. 73–77. isbn: 9781450362870. doi: 10.1145/3313991.3314008. url: https://doi.org/10.1145/3313991.3314008.

Reichelmann, Ashley et al. (2020). "Hate Knows No Boundaries: Online Hate in Six Nations". In: *Deviant Behavior*, pp. 1–12.

Rizwan, Hammad, Muhammad Haroon Shakeel, and Asim Karim (Nov. 2020). "Hate-Speech and Offensive Language Detection in Roman Urdu". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 2512–2522. doi: 10.18653/v1/2020.emnlp-main.197. url: https://www.aclweb.org/anthology/2020.emnlp-main.197.

Safi Samghabadi, Niloofar et al. (May 2020). "Aggression and Misogyny Detection using BERT: A Multi-Task Approach". In: *Proceedings of the Second Workshop on Trolling,*

Aggression and Cyberbullying. Marseille, France: European Language Resources Association (ELRA), pp. 126–131. isbn: 979-10-95546-56-6. url: https://www.aclweb.org/anthology/2020.trac-1.20.

Sajjad, M. et al. (2019). "Hate Speech Detection using Fusion Approach". In: *2019 International Conference on Applied and Engineering Mathematics (ICAEM)*, pp. 251–255.

Saroj, Anita and Sukomal Pal (May 2020). "An Indian Language Social Media Collection for Hate and Offensive Speech". In: *Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language*. Marseille, France: European Language Resources Association (ELRA), pp. 2–8. url: https://www.aclweb.org/anthology/2020.restup-1.2.

Sarthak et al. (2021). "Detecting Hostile Posts using Relational Graph Convolutional Network". In: *ArXiv* abs/2101.03485.

Schlichtkrull, Michael, Thomas N. Kipf, and Peter Bloem (2018). "Modeling Relational Data with Graph Convolutional Networks". In: *The Semantic Web*. Cham: Springer International Publishing, pp. 593–607. isbn: 978-3-319-93417-4.

Shu, Kai, Suhang Wang, and Huan Liu (2019). "Beyond news contents: The role of social context for fake news detection". In: *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*. New York, NY, United States: Association for Computing Machinery, pp. 312–320.

Spertus, Ellen (1997). "Smokey: Automatic recognition of hostile messages". In: *Aaai/iaai*, pp. 1058–1065.

Tran, Thanh et al. (Nov. 2020). "HABERTOR: An Efficient and Effective Deep Hatespeech Detector". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 7486–7502. doi: 10.18653/v1/2020.emnlp-main.606. url: https://www.aclweb.org/anthology/2020.emnlp-main.606.

Wang, R., Wenlin Liu, and Shuyang Gao (2016). "Hashtags and information virality in networked social movement: Examining hashtag co-occurrence patterns". In: *Online Inf. Rev.* 40, pp. 850–866.

Waseem, Zeerak, Thomas Davidson, et al. (2017). "Understanding Abuse: A Typology of Abusive Language Detection Subtasks". In: *Proceedings of the First Workshop on Abusive Language Online*. Vancouver, BC, Canada: Association for Computational Linguistics, pp. 78–84. doi: 10.18653/v1/W17-3012. url: https://www.aclweb.org/anthology/W17-3012.

Waseem, Zeerak and Dirk Hovy (2016). "Hateful symbols or hateful people? predictive features for hate speech detection on twitter". In: *Proceedings of the NAACL student research workshop*. San Diego, California: Association for Computational Linguistics, pp. 88–93.

Wiegand, Michael and Melanie Siegel (2018). "Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language". In.

Zampieri, Marcos et al. (2019a). "Predicting the Type and Target of Offensive Posts in Social Media". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume*

1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1415–1420. doi: 10.18653/v1/N19-1144. url: https://www.aclweb.org/anthology/N19-1144.

Zampieri, Marcos et al. (June 2019b). "SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 75–86. doi: 10.18653/v1/S19-2010. url: https://www.aclweb.org/anthology/S19-2010.

Zhou, Siyao, Jie Li, and Haiyan Ding (2021). "Fake News and Hostile Posts Detection Using an Ensemble Learning Model". In: *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Cham: Springer International Publishing, pp. 74–82. isbn: 978-3-030-73696-5.