Identification of hate material in code mixed social media data using demographic and contextual information.

A Major Project Report
Submitted in partial fulfillment of the requirements for the degree of
Bachelor of Technology
in
Computer Science and Engineering

by

Souvik Das	**18BCS099**
Amitabh Paliwal	**18BCS004**
Sahana N H	**18BCS086**
Derik Lytten	**17BCS008**

Supervisor:
Dr. Sunil Saumya

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY

Department of Computer Science and Engineering
Indian Institute of Information Technology
Dharwad (India)

28 May 2022

Certificate

Department of Computer Science and Engineering
Indian Institute of Information Technology, Dharwad

It is certified that the work contained in the project report entitled "Identification of hate material in code mixed social media data using demographic and contextual information." by the following students has been carried out under my/our supervision and that this work has not been submitted elsewhere for a degree.

Souvik Das	18BCS099
Amitabh Paliwal	18BCS004
Sahana N H	18BCS086
Derik Lytten	17BCS008

Date: Dr. Sunil Saumya

This project report entitled "Identification of hate material in code mixed social media data using demographic and contextual information." submitted by the group is approved for the degree of Bachelor of Technology.
The final presentation has been held on _____.

Supervisor(s) Examiner(s)

Declaration

IIIT Dharwad
28 May 2022

I/We declare that this written submission represents my/our ideas in my/our own words and where others' ideas or words have been included, I/We have adequately cited and referenced the original sources. I/We also declare that I/We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I/We understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.[1]

Souvik Das	18BCS099	_____
Amitabh Paliwal	18BCS004	_____
Sahana N H	18BCS086	_____
Derik Lytten	17BCS008	_____

[1]This report contains visualizing of explicit words or slang, reader's discretion is advised. The aforementioned words or slang are not meant for any hurting feelings or sentiments. The data is collected from Twitter, a public domain.

Abstract

In this report we examine methods to detect targeted audience and severity of code-mixed/Hinglish(Hindi + English) hate speech in social media, while distinguishing this from general profanity. We aim to establish lexical baselines for this task by applying NLP algorithms using our own data-set and embeddings for this purpose.
Keywords: **hate speech; social media; english-hindi; hinglish; machine learning; nlp; bi-lstm**

Table of Contents

List of Figures

Chapter 1

Introduction

Social media is changing the face of communication and accessibility in the world. In 2020, over 3.6 billion people were using social media worldwide, a number projected to increase to almost 4.41 billion in 2025. With such a huge user base, it is not uncommon to see a substantial amount of hate directed towards influential people, particular religions/cultures/races, political parties, and other individuals. While positivity does prevail in some parts of social media, the hate spread over these platforms can sometimes lead to huge riots and even an information war between two countries.

In this report, we discuss how we can predict the targeted audience and the severity of ***Hinglish(Hindi + English)*** hate speech posted over social media. For the training/testing dataset we scraped Twitter to collect over 9000 hate tweets posted in Hinglish, and manually labelled them in as classes of targets of hate speech ***Target (Individual, organizational, Religion)*** and classes of severity of the hate speech ***Severity (Negligible, Moderate, Severe)***. In the upcoming chapters, a detailed analysis of the dataset is provided, which helps us visualize the data points. A list of data points is given below for reference:

- `time_created`
- `tweet`
- `retweets, likes, quotes, replies`
- `username, user_name, user_desc, num_tweets, loc`
- `followers, following`
- `target, severity`

We created word vectors to encode the meaning of the words present in our corpus of text. Using these word vectors we were able to use deep learning solutions (Bi-LSTMs) to the problem. The tweets are classified according to targets and severity using tweet text. Another classification using the tweet text and target information to predict severity is also done.

Chapter 2

Motivation

With social media hate spreading like wildfire, and hate tweets receiving a substantial amount of attention, this project aims at identifying severe hate tweets that might have the potential to start riots and information wars between two countries.
With no proper Hinglish dictionary for swear words, and embedding vectors for the same available we built this project from scratch using our own Hinglish data which consisted of tweets, word embeddings, Hinglish stopwords, and Hinglish hate words.
Hate tweets receive a significant amount of re-tweets and likes on the platform, and as such there is a general trend towards increasing animosity on social media platforms such as twitter.
As an effort to stop this malevolence we built this project. We aim to identify and curb the spread of such tweets, and make social media a better place for all.
EDA(s) have been attached below to support the above-mentioned claim.

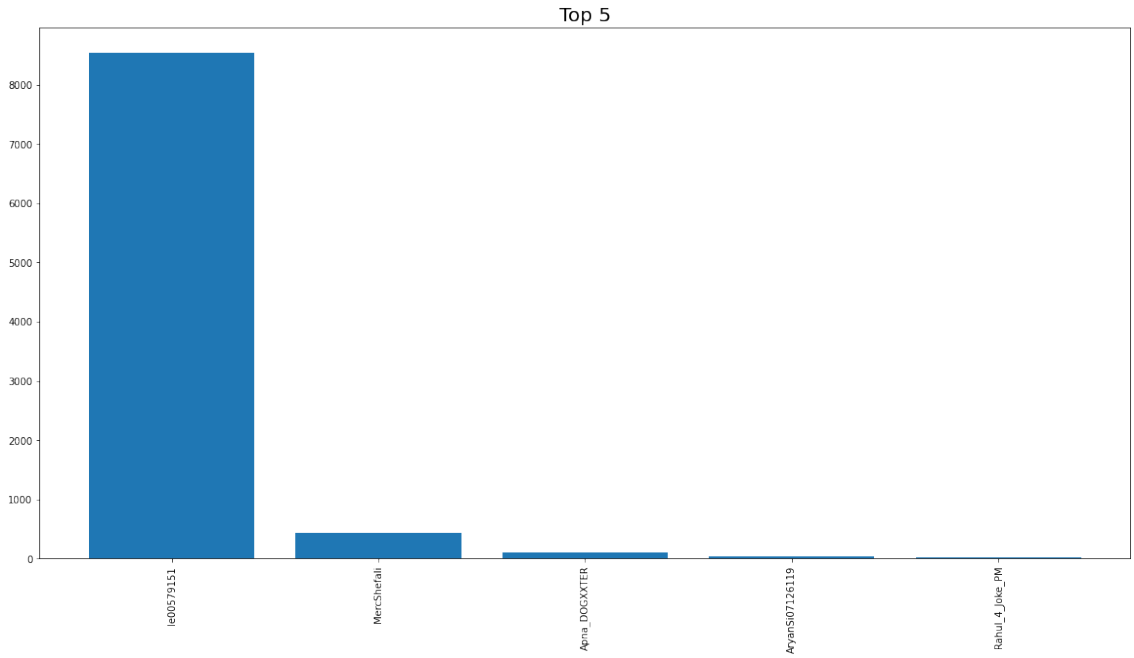

Figure 2.1: Top 5 user whose hate tweets have received most number of re-tweets. *x - username, y-number of re-tweets*

The number of retweets on a hate tweet are more than 6000, other tweets show significant retweet counts in the order of hundreds.

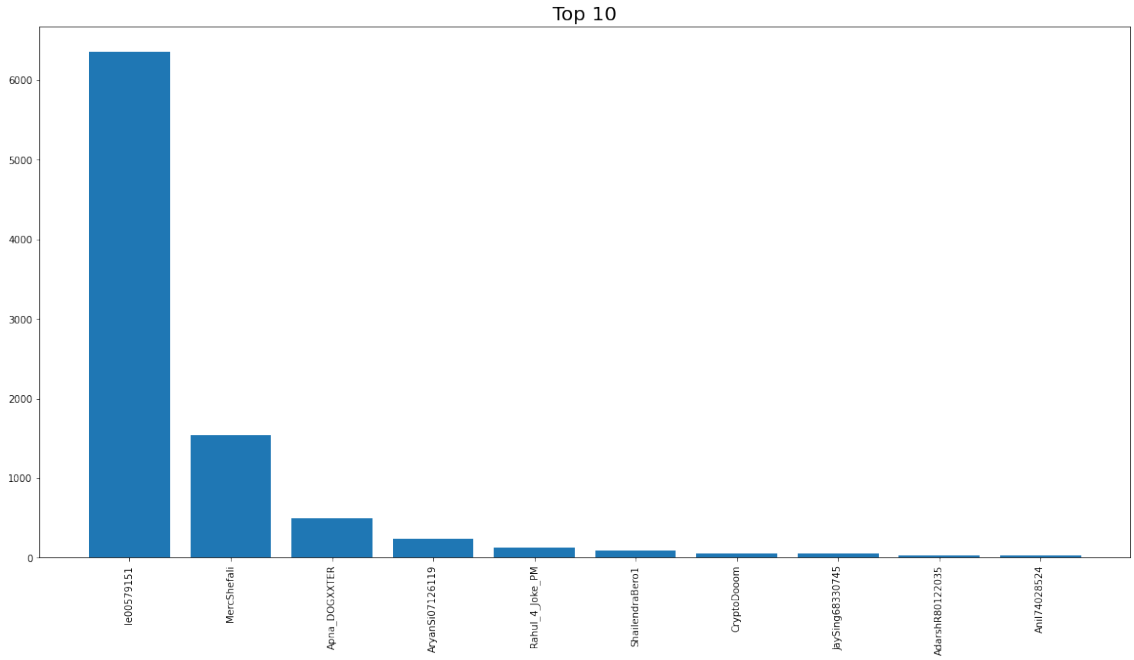

Figure 2.2: Top 10 user whose hate tweets have received most number of likes. *x - username, y-number of likes*

The number of likes on a hate tweet are more than 8000, other tweets show significant like counts in the order of hundreds. Thus, the general trend is that hate speech is receiving more and more attention online.

Chapter 3

Literature Review

There exist datasets made by tagging tweets and facebook comments as hate or not-hate in code-mixed data[1] [2]. Work was also done on annotating tweets according to verbal aggression and its types, but no classification results or methods are discussed in the paper and the classification task is mentioned as a future objective[3]. Most of the work on code-mixed data is based upon classifying text as hate or not-hate[4][5]. Our method tries to create fine-grained classifications for code-mixed hate speech. The method of creating twitter API search queries using modifiers and targets was an idea we came across during our literature survey. Our survey indicated that the general methods used for classification include SVM, KNN, Random-forest classifiers, 1D CNNs, LSTMs, and Bi-LSTMs. The papers indicate that SVMs, 1D CNNs and LSTMs give the highest accuracies on their respective datasets, but the recall for LSTMs is generally better[1][6]. We have used Bi-LSTMs for classification.

Chapter 4

Datasets

This project is based on Hinglish tweets collected from Twitter using Twitter API v2.[1]

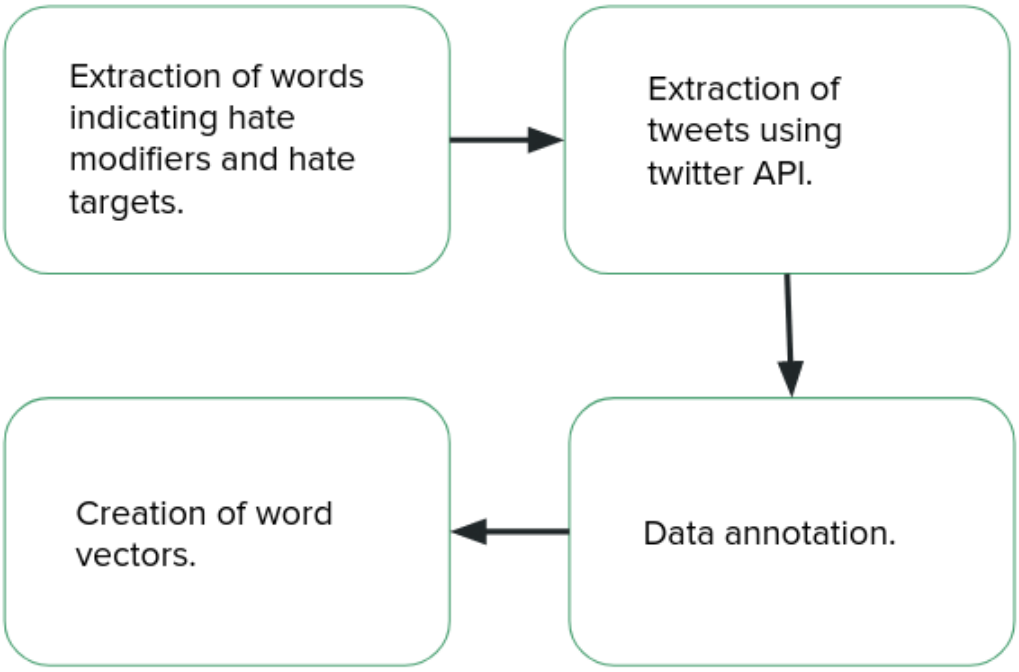

Figure 4.1: An overview of the dataset creation process.

[1]This chapter contains visualizing of explicit words or slang, reader's discretion is advised. The aforementioned words or slang are not meant for hurting any feelings or sentiments. The data is collected from Twitter, a public domain.

4.1 Extraction of Hinglish hate tweets

Extraction of words indicating hate modifiers and hate targets

We used a dataset of Hinglish tweets which contained the tweet text and a column indicating whether the text is hate speech or not. Using the latter column's values, we extracted the text values with 'yes' as the hate indicator. We then collected all the words in this corpus of text that appeared with some significant frequency (more than eight appearances). This list of words was reviewed using our custom *Hinglish stopwords corpus* to remove the Hinglish stopwords. After manual review this list was further split into two separate CSV files consisting of *hate modifiers* (explicit words) and *hate targets* (targeted audience)[7].

Extraction of tweets using twitter API

The hate modifiers and hate targets were combined to form Twitter search queries to be used with the API. Due to a character limit of 512 characters on the search query, the modifiers and targets were used in sets of fifteen. Each set of modifiers was combined with every set of targets present. Each one of these search queries created returned a maximum of 100 results. Using these search queries we collected over 15000 Hinglish hate tweets.
We used expansions provided by the Twitter API to pull the user data, and the location data with the tweet data. Finally the data-points gathered were as follows:

- `time_created`: The timestamp for when the tweet came into existence.
- `tweet`: The text content of the tweet.
- `retweets`: The number of retweets received by the tweet.
- `likes`: The number of likes received by the tweet.
- `quotes`: The number times the tweet was used as a quote tweet.
- `replies`: The number of replies received by the tweet.
- `username`: The username of the user who created the tweet.
- `user_name`: The name provided with the ID by the user who created the tweet.
- `user_desc`: The user description of the creator of the tweet.
- `num_tweets`: The total number of tweets created by the user who created the tweet.

- `loc:` The geo-location of the user who created the tweet.
- `followers:` The number of followers of the user who created the tweet.
- `following:` The number of users the creator of the tweet follows on the platform.

After the dataset was created we came across duplicates where the content of the tweet was the same as another tweet but the mentions and the hashtags used were different. In order to resolve this problem we used a thorough cleaning procedure which included removing mentions, hashtags, emoticons, and URLs to create a new column which was used to drop duplicates from the dataset. Since this column was redundant in the final dataset it was removed from the final version.

4.2 Data annotation

We manually labeled the tweets on two factors: *target and severity*.

- target: It was used to label the tweet based on its target audience.

  'target' has three possible values:

  - 1 : hate speech directed towards an individual.
  - 2 : hate speech directed towards an organization such as political parties, governments, countries, corporations etc.
  - 3 : hate speech directed towards a religion.

- severity: was used to label the tweet based on its outcome potential among the masses.

  'severity' has three possible values:

  - 1 : casual use of abuse words in use as slang.
  - 2 : heavy use of abuse words in use as slang or use of abuse words not used as slang that may incite violence.
  - 3 : heavy use of abuse words that are not used as slang or threats of riots and calls to violence.

time_created	tweet	retweets	likes	quotes	replies	username	user_name	user_desc	followers	following	num_tweets	loc	target	severity
2022-04-16 13:3	har kashmiri hinc	0	1	0	0	MahaKaal10000	Utsav Koul	this is utsav koul	361	956	605		3	2
2022-04-16 12:5	@muntajeemkha	1	1	0	2	TheBaahubali	बाहु #HTL	एक सद्विप्रा बहुधा वद	25990	294	109971		3	2
2022-04-16 11:5	@shivanirajput_	0	0	0	0	arvinderarora12	arvinder arora	iam kind hearted	99	894	471		3	3
2022-04-15 17:0	@fierynature202	0	0	0	0	MohanVerna7	Mohan Verma 7	mohan verma7 h	1364	1390	932		3	3
2022-04-15 08:5	@nagmariyasat	0	0	0	0	Shailes91502385	Shailesh Soni (s	s s hindu	491	955	2501		3	2
2022-04-15 07:3	@sudhirchaudha	0	1	0	0	Sandeep810731	Sandeep Singh	unemployment	1	57	145		1	1
2022-04-15 05:5	@surajraghu1sh	0	0	0	0	LoveEmotions02	Love Emotions	it's a status junct	24	36	939		3	2

Figure 4.2: Hinglish dataset(*df*) as csv

It is evident from the figure above that most of the tweets do not have any *loc* given.

4.3 Hinglish hate modifiers and targets

The custom Hinglish hate modifiers and targets corpus containing Hinglish abuse modifiers and target audiences extracted from the dataset of hate tweets referred to earlier used to build search queries for the Twitter API.

	Unnamed: 0	adjectives
5	5	murderer
94	94	haramiyio
154	154	fucker
23	23	gandi
128	128	barbaad

	Unnamed: 0	groups
22	22	cricket
83	83	bacche
65	65	ammi
145	145	terrorism
7	7	sanghi

Figure 4.3: Hinglish hate modifiers and targets respectively

4.4 Hinglish stopwords corpus

A custom Hinglish stopwords corpus containing over 1000 common Hinglish adjectives, pronouns, prepositions, and articles was created to aid the data cleaning process and exploratory data analysis.

custom_st

stopwords
a
aadi
aaj
aap
aapne
aata
aati
aaya
aaye
ab
abbe

Figure 4.4: Custom Hinglish stopwords corpus

4.5 Creation of Hinglish word vectors

The word vectors were created using the skipgrams variant of the word2vec algorithm which uses distributional semantics to create vectors representing the words. The meaning of a word is based on the context in which it appears. It is a dense representation that uses real-valued vectors and can better capture similarity between words.

The context of a word is defined as the set of 'm' surrounding words.

Eg. for m = 2 context of the word 'fox' in "The quick brown fox jumped over the lazy dog." is [quick, brown, jumped, over].

Our word vectors used a window size of 8 for context.

The skipgrams word2vec algorithm predicts the distribution of the context words using the center word. The representations of the words are spread over 300 dimensions which capture different things about the word. This would include semantic information such as singular/plural, male/female, etc. along the different dimensions.

Before using custom word vectors, GloVe 6B 50D - stanfordnltk and inltk, have also been used to create a bi-LSTM model but unfortunately, neither of them have vectors for the hate words used in the tweets which led to reduced accuracy and increased false positives.

Chapter 5

Data Analysis

In this chapter, a detailed analysis of the dataset is provided using pertinent and self-explanatory graphs and a word cloud. A custom Hinglish stopword corpus has been used to remove the commonly used Hinglish pronouns, prepositions, and articles to aid the data analysis.[1]

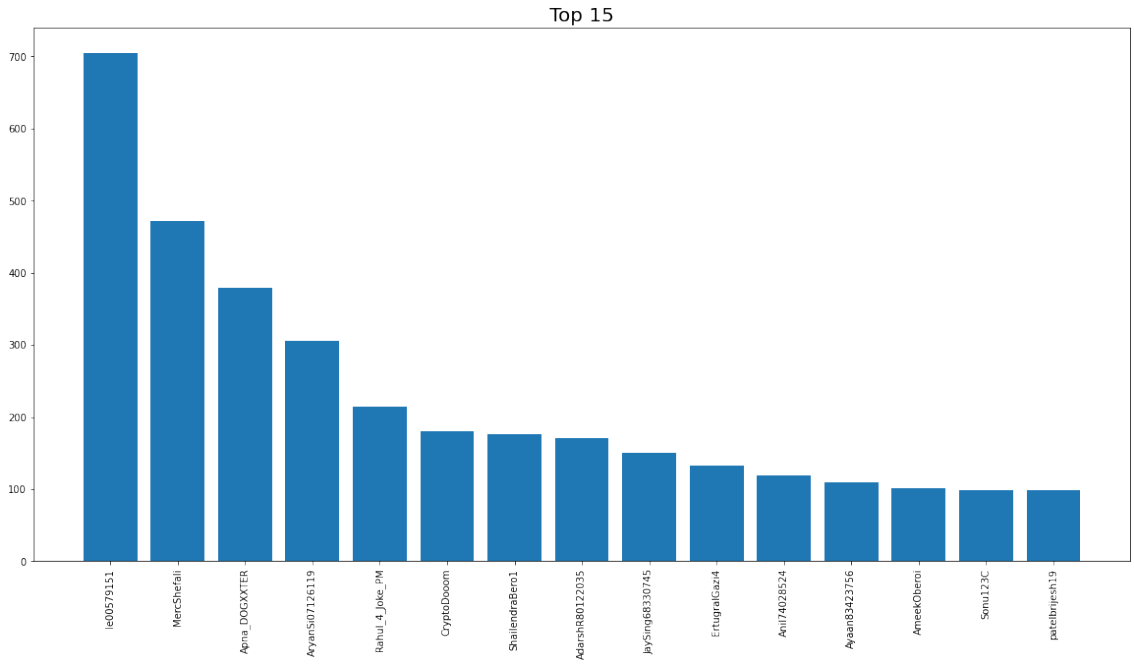

Figure 5.1: Top 15 users who have posted most hate on twitter.
x - username, y-number of hate tweets

[1]This chapter contains visualizing of explicit words or slang, reader's discretion is advised. The aforementioned words or slang are not meant for hurting any feelings or sentiments. The data is collected from Twitter, a public domain.

This graph shows the users who have been engaged in hate speech on twitter. It shows their usernames against their tweet count.

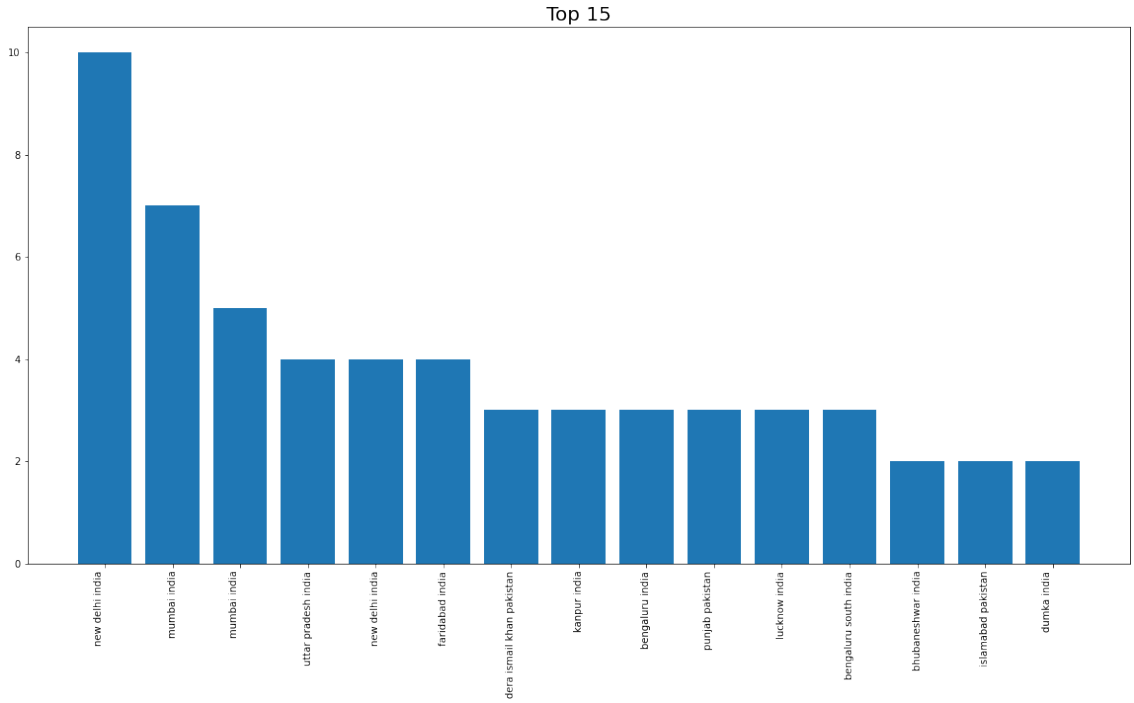

Figure 5.2: Top 15 locations from where most hate is generated.

This graph is not representative of the complete dataset. A lot of the tweets did not have location values provided. These are the locations that appeared the most times among the tweets that did have location data.

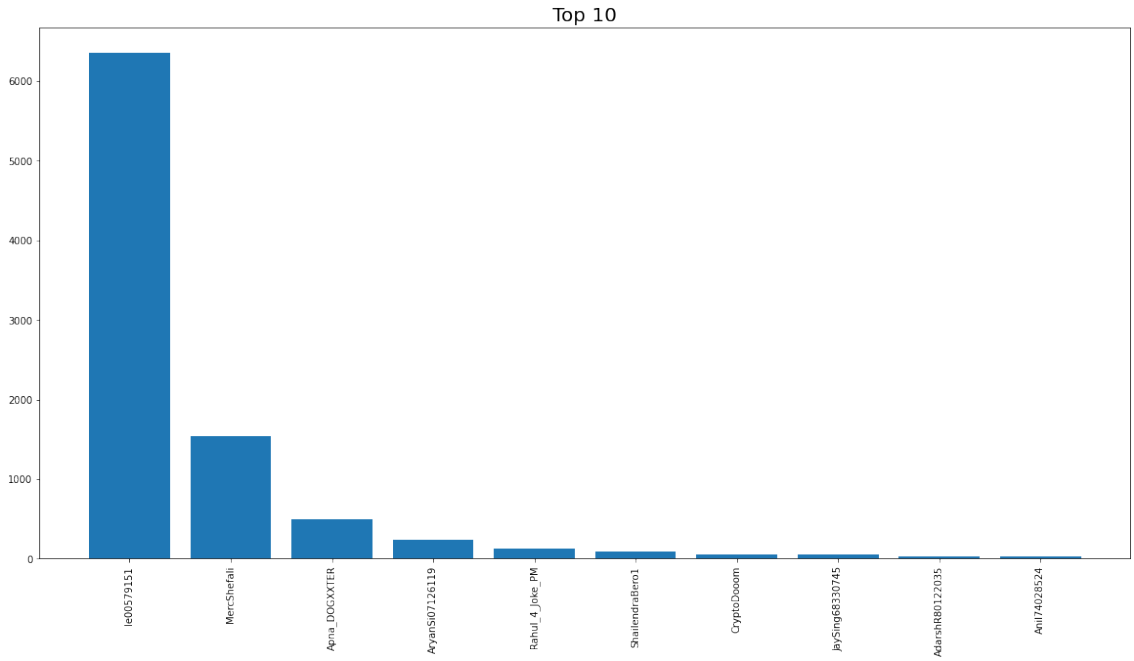

Figure 5.3: Top 10 users whose hate tweets have received most number of likes.
x - username, y-number of likes

The users who have received the most number of likes on a single tweet. This data was used to establish the relevance of the problem and the motivation for our project.

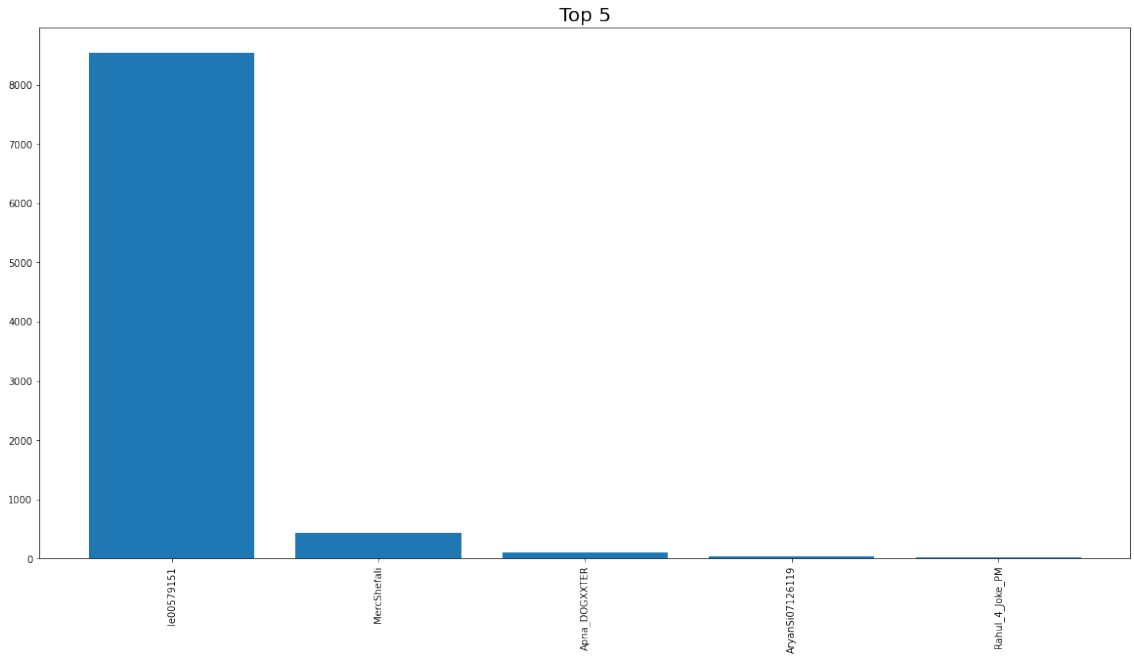

Figure 5.4: Top 5 user whose hate tweets have received most number of re-tweets.
x - username, y-number of re-tweets

The users who have received the most number of re-tweets on a single tweet. This data was used to establish the relevance of the problem and the motivation for our project.

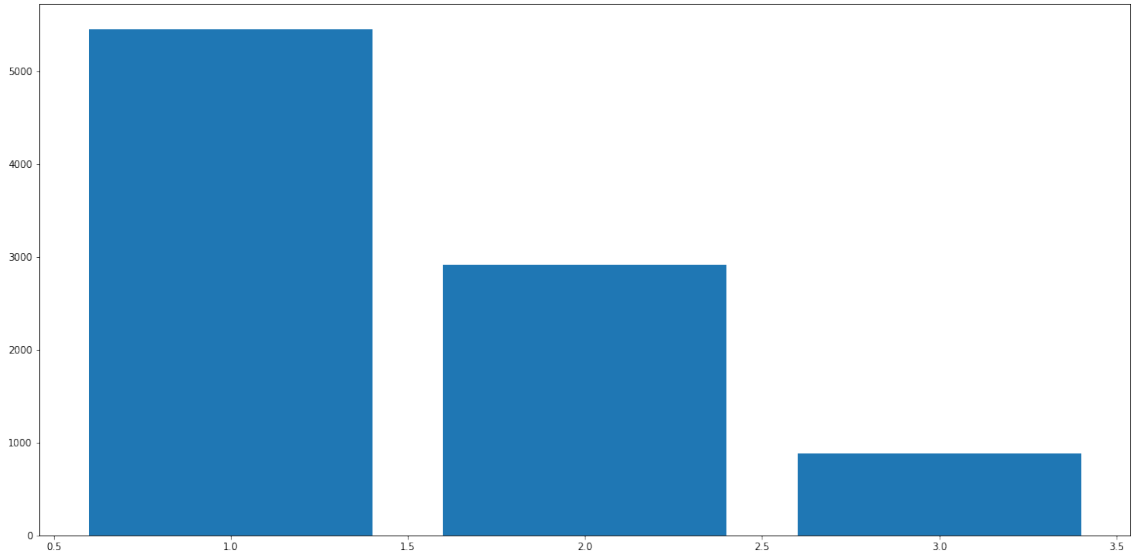

Figure 5.5: Number of tweets in each severity segment.
Severity(1-Negligible, 2-Moderate, 3-Severe)
x - severity, y-number of tweets

This graph shows the number of tweets in each category of severity. Here, 1 stands for negligible severity, 2 stands for moderate severity, and 3 stands for severe hate.

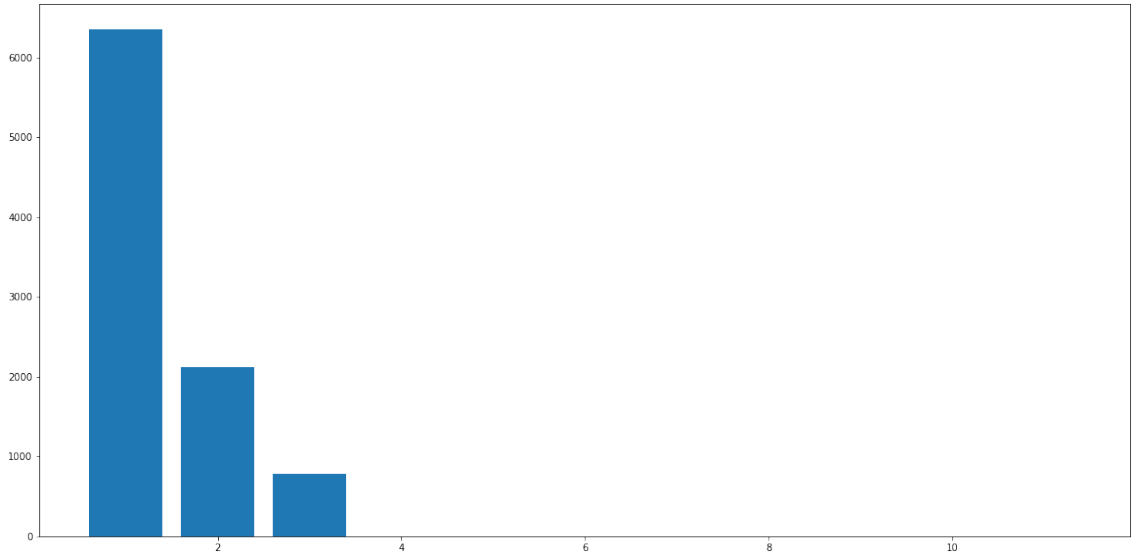

Figure 5.6: Number of tweets in each target segment.
Target(1-Individual, 2-organizational, 3-Religion)
x - target, y-number of tweets

This graph shows the number of tweets in each target category. Here, 1 stands for individual as a target, 2 stands for organization as a target, and 3 stands for religion as a target.

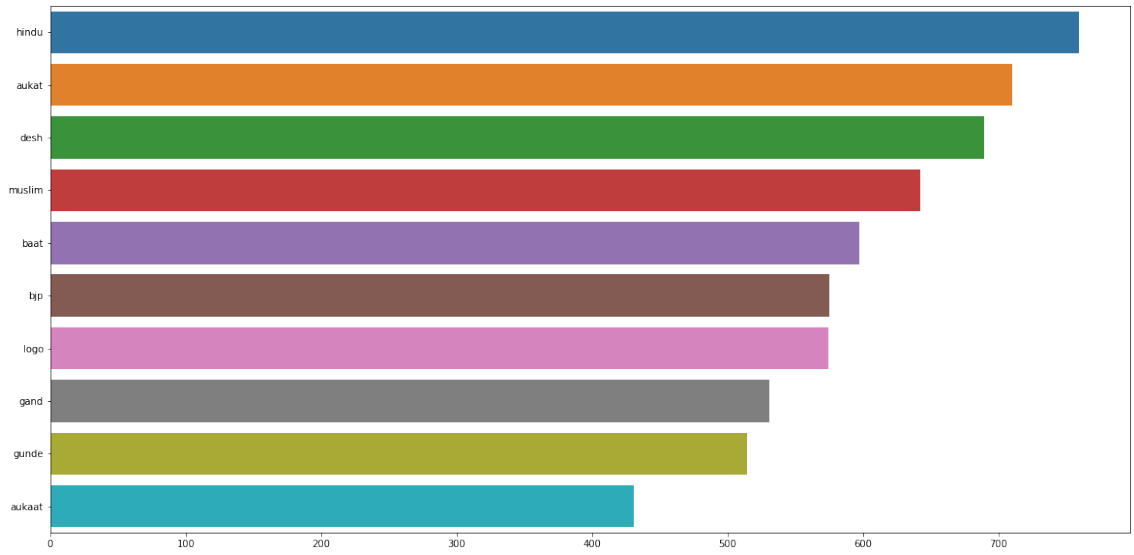

Figure 5.7: Top 10 Hinglish hate words.
x - frequency, y-Hinglish words

This figure shows the words that appeared the most number of times in the hate tweets in the database, excluding the stopwords.

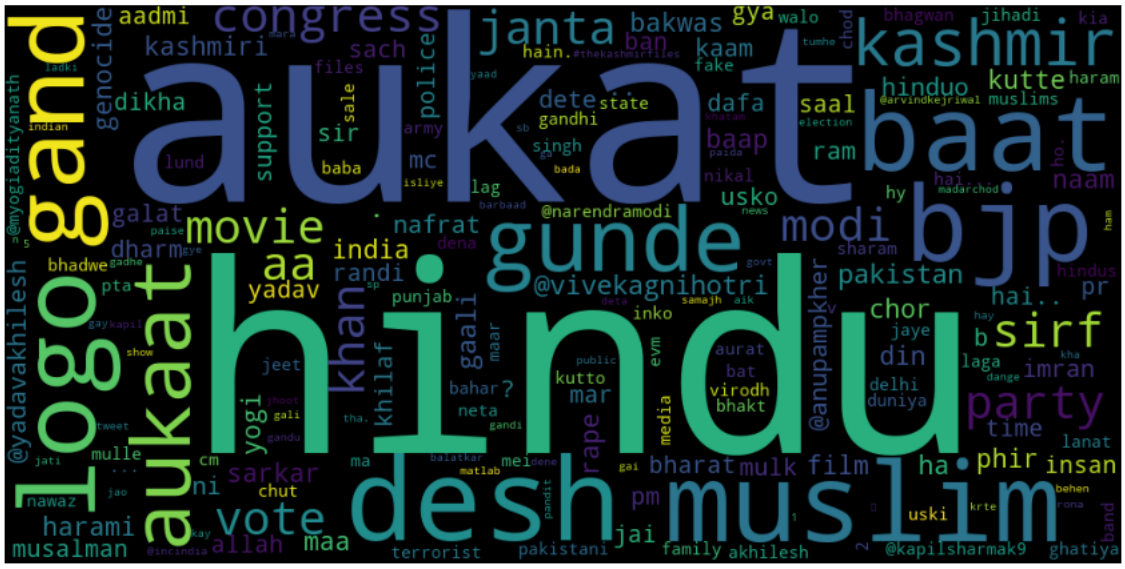

Figure 5.8: Wordcloud

Chapter 6

Proposed solution and Results

6.1 Bi-LSTM

Our project uses ***Bi-LSTM*** to *predict severity using hate tweet and target*. The procedure discussed in this chapter can also be followed for *predicting target using hate tweet and severity*.

Bidirectional Long Short-Term Memory (*Bi-LSTM*) is a type of recurrent neural network. It processes data in two directions since it works with two hidden layers. Bi-LSTMs effectively increase the amount of information available to the network, improving the context available to the algorithm (e.g. knowing what words immediately follow and precede a word in a sentence).

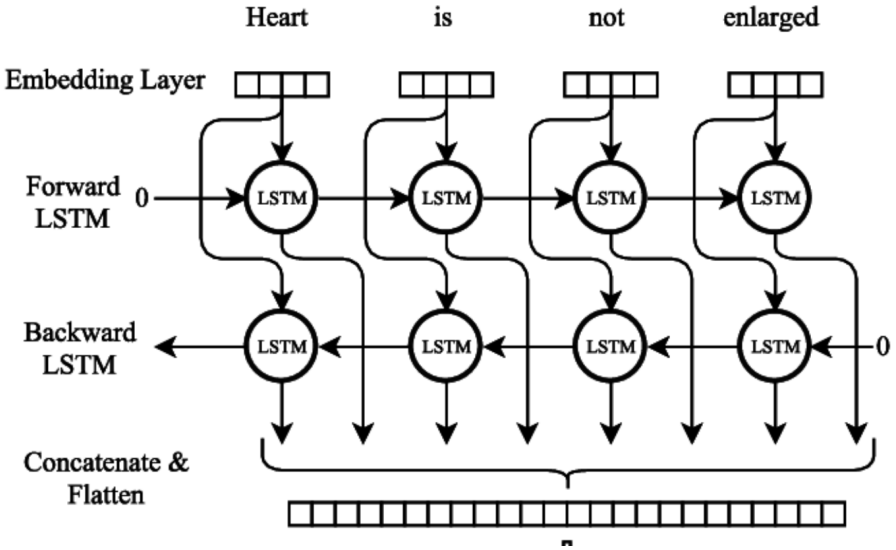

Figure 6.1: Image Source: Modelling Radiological Language with Bidirectional Long Short-Term Memory Networks, Cornegruta et al

6.2 Preparing the data

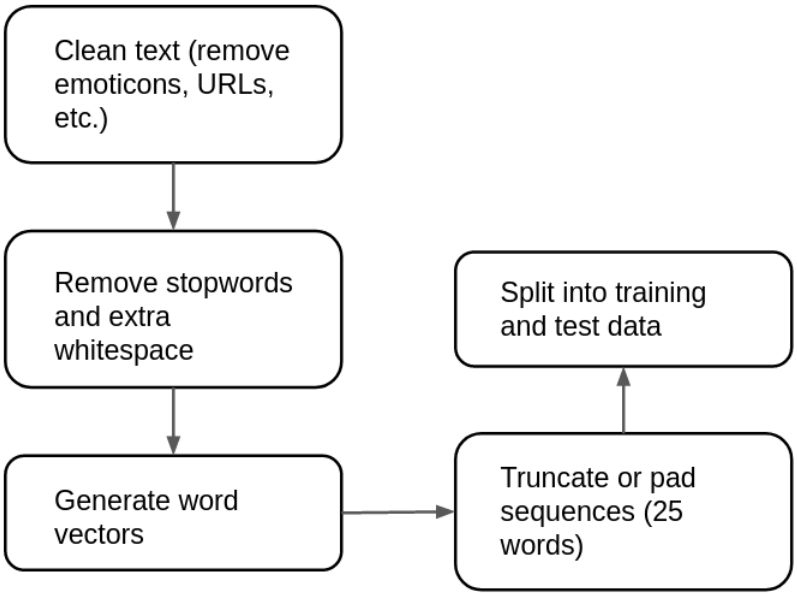

Figure 6.2: Summary of the pre-processing done to prepare the data.

The annotated dataset containing 9200+ tweets with target and severity labelled are loaded as the dataset (*balanced-ds.csv*). First we remove the emoticons, mentions, hashtags, and URLs from the tweet data to create a new column containing the cleaned text. Then we use this cleaned text and remove extra whitespace and stopwords from the text. The word vectors for the dataset were generated using word2vec algorithm using the skipgrams variant where the context window size was 8 and the number of dimensions was 300.

We split the dataset into training and test data using a 70/30 train/test split. The sequence length was limited to 25 words after analysing the sentence lengths in the data.

Three classifications were made. For the first classification we used the tweet text to predict targets of the hate tweets. For the second classification we used the tweet text to predict severity of the hate tweets. For the third classification we used target data as additional data along with the tweet text. Data preparation for the third classification required an additional creation of another input variable for the model.

6.3 The Bi-LSTM model

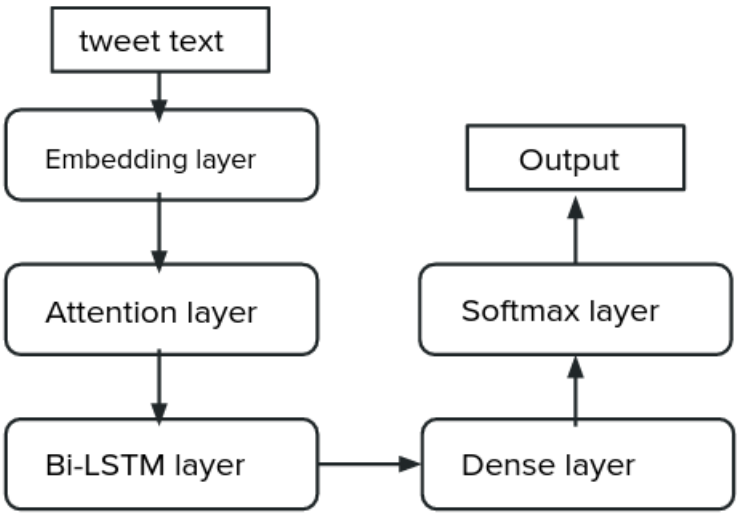

Figure 6.3: The structure of the model used for classifying target and severity using tweet text as input data.

Hyper-parameters of the model used for classifying target and severity using tweet text as input data.

Model	Bi-LSTM with attention
No. of neurons in attention layer	25
Attention layer activation function	softmax
No. of Bi-LSTM layers	2
No. of units in each Bi-LSTM layer	15
Bi-LSTM activation function	softmax
Bi-LSTM loss function	sparse_categorical_crossentropy
Bi-LSTM optimizer	adam

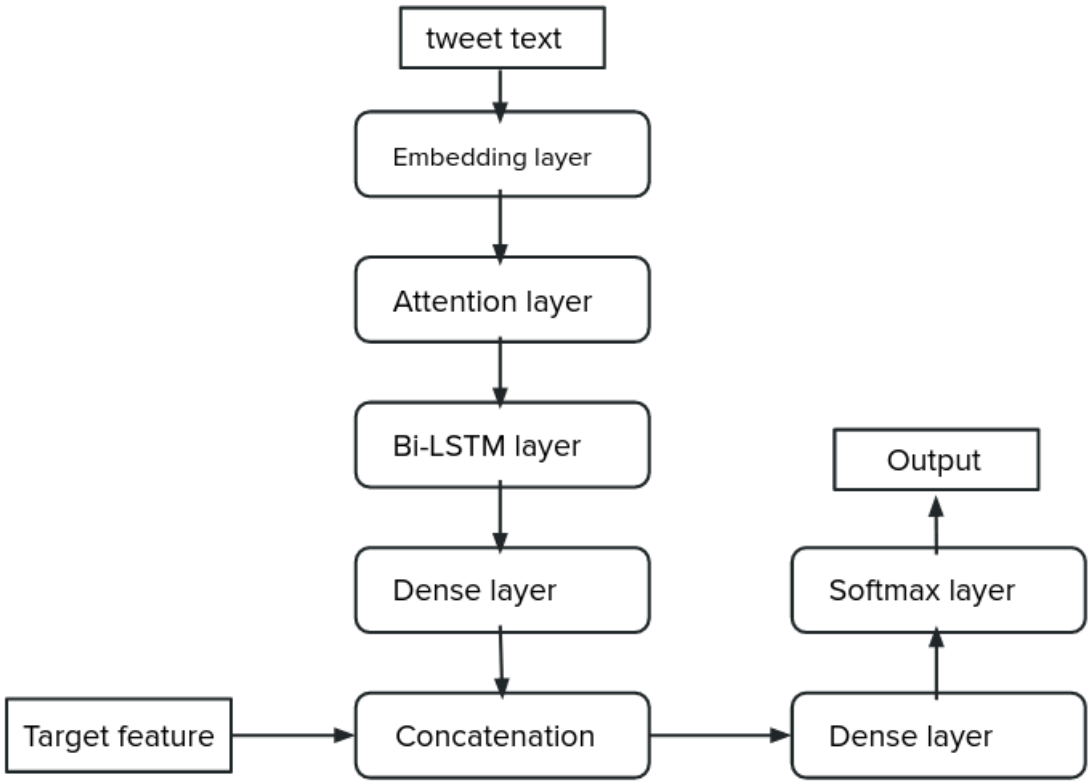

Figure 6.4: The structure of the model used for classifying severity using tweet text and target information as input data.

Hyper-parameters of the model used for classifying severity using tweet text and target information as input data.

Model	Bi-LSTM with attention
No. of neurons in attention layer	25
Attention layer activation function	softmax
No. of Bi-LSTM layers	2
No. of units in each Bi-LSTM layer	15
Bi-LSTM activation function	softmax
Bi-LSTM loss function	sparse_categorical_crossentropy
Bi-LSTM optimizer	adam

Our model uses the word vectors generated using the tweet corpus. The model consists of an attention layer. This is followed by two bidirectional layers containing LSTMs, which use a dropout value of 0.2 to prevent overtraining, and working together implement a Bi-LSTM layer. This is followed by a dense layer of neurons using the relu activation function. For the first two classifications which use the tweet text to make the classifications we then use a softmax layer to get the probabilities. For the third classification we use another layer to concatenate the target value feature with the features obtained by the attention layer, the Bi-LSTM layer, and the dense layer. Another dense layer is used. Then the softmax layer is used to get the outputs. The model is fitted using adam optimizer. The loss used was categorical cross entropy loss. There was a further training/validation split of 70/30. The model was trained over 10 epochs using a batch size of 256.

6.4 Process Summary

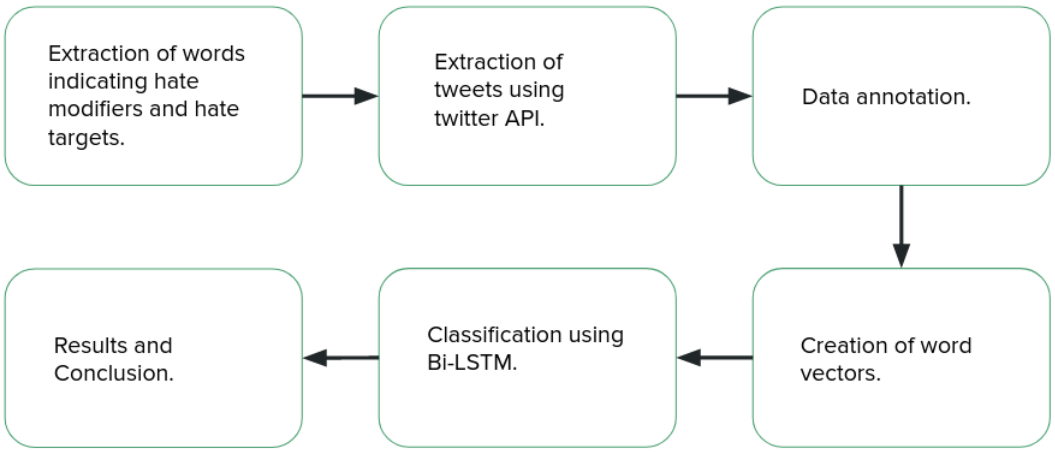

Figure 6.5: A summary of the process followed.

6.5 Results

6.5.1 Predicting target using tweet text

Training and validation

Using the Bi-LSTM model for predicting targets using tweet text we get a validation accuracy of 68.83 %.

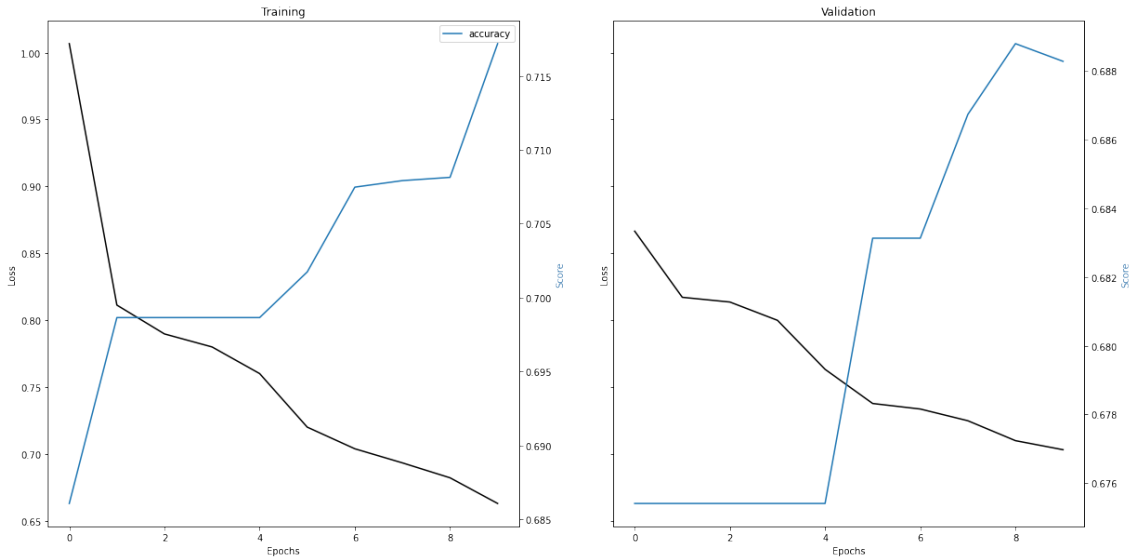

Figure 6.6: A graph representing training accuracy/loss and validation accuracy/loss respectively for prediction of targets using tweets.

The figure shows the training accuracy and loss and the validation accuracy and loss over 10 epochs of training.

Testing results

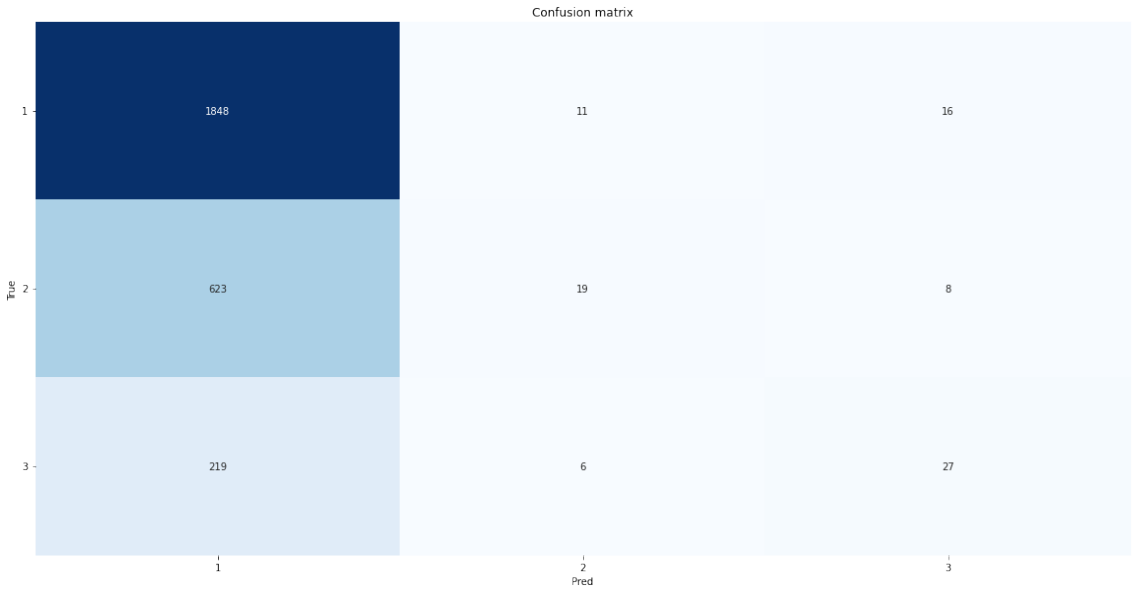

Figure 6.7: Confusion matrix for prediction of targets using tweets.

The confusion matrix shows us that the target prediction for the category 1 (individual target) was very accurate. However, the results for the category 2 (organisational target) and category 3 (religion target) were not good. Many tweets from the target category 2

and 3 were wrongly classified as target category 1 tweets.

```
Accuracy: 0.68
Auc: 0.75
Detail:
              precision    recall  f1-score   support

           1       0.69      0.99      0.81      1875
           2       0.53      0.03      0.06       650
           3       0.53      0.11      0.18       252

    accuracy                           0.68      2777
   macro avg       0.58      0.37      0.35      2777
weighted avg       0.64      0.68      0.58      2777
```

Figure 6.8: Classification report for prediction of targets using tweets.

The classification report for the test data shows an accuracy of 68%. The precision scores do not show extreme variation. This means that a good proportion of positive identifications were actually correct. But the recall scores for target category 2 and 3 are not good compared to the the category 1. Thus, the proportion of actual positives that were identified correctly was great for category 1 but not good for category 2 and 3. This is consistent with the confusion matrix obtained.

6.5.2 Predicting severity using tweet text

Training and validation

Using the Bi-LSTM model for predicting severity using tweet text we get a validation accuracy of 65.17 %.

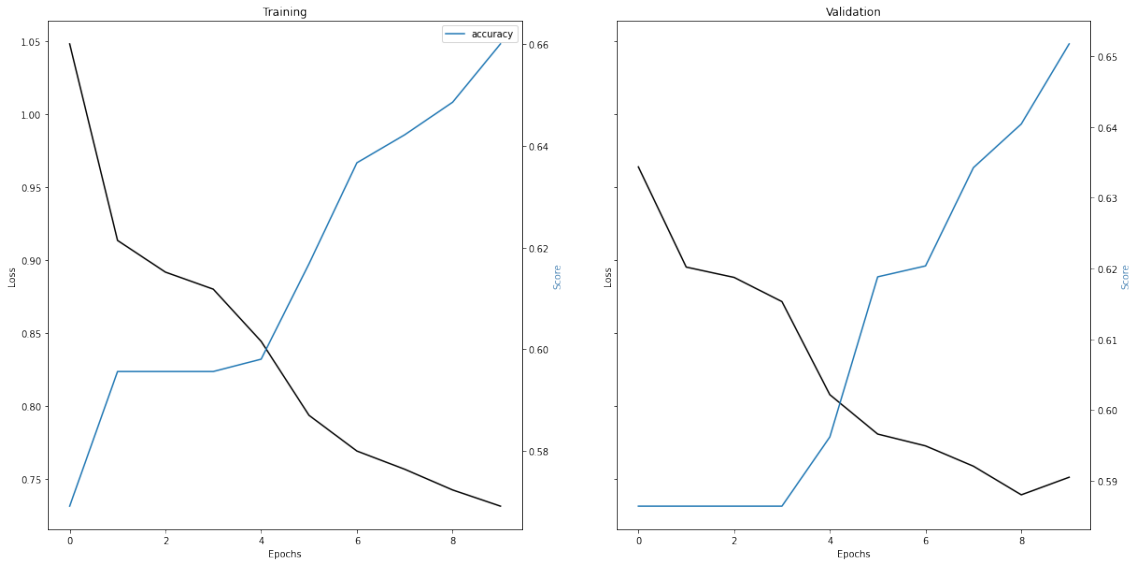

Figure 6.9: A graph representing training accuracy/loss and validation accuracy/loss respectively for prediction of severity using tweets.

The figure shows the training accuracy and loss and the validation accuracy and loss over 10 epochs of training.

Testing results

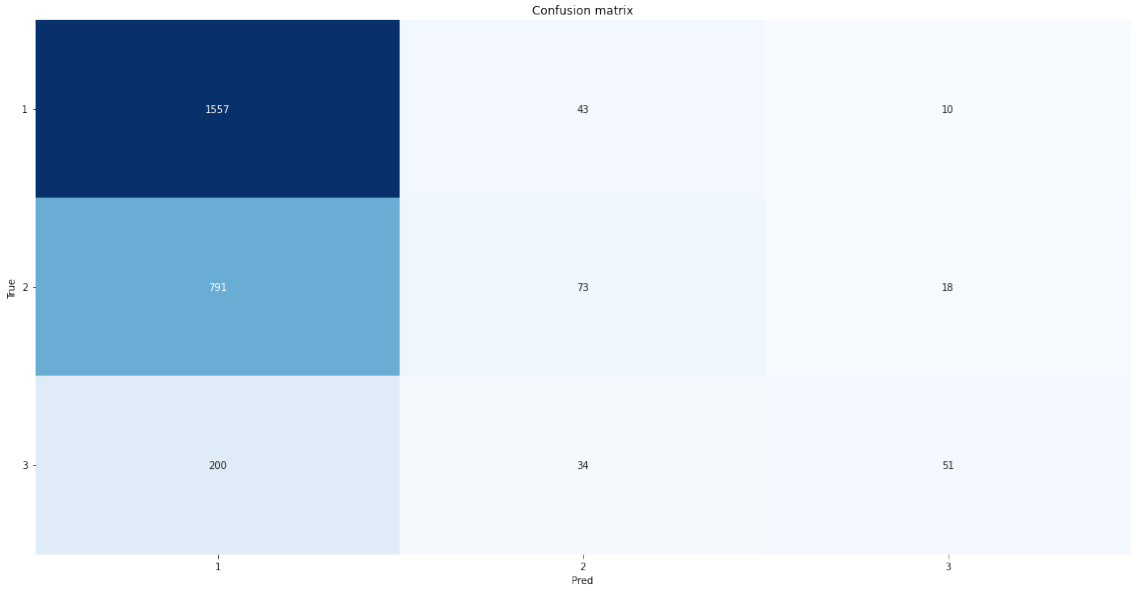

Figure 6.10: Confusion matrix for prediction of severity using tweets.

The confusion matrix shows us that the severity prediction for the severity class 1 (negligible severity) was very accurate. However, the results for the severity class 2

(moderate severity) and severity class 3 (severe severity) were not good with many tweets of severity class 2 and 3 classified as class 1 severity tweets.

```
Accuracy: 0.61
Auc: 0.75
Detail:
              precision    recall  f1-score   support

           1       0.61      0.97      0.75      1610
           2       0.49      0.08      0.14       882
           3       0.65      0.18      0.28       285

    accuracy                           0.61      2777
   macro avg       0.58      0.41      0.39      2777
weighted avg       0.58      0.61      0.51      2777
```

Figure 6.11: Classification report for prediction of severity using tweets.

The classification report for the test data shows an accuracy of 61%. The precision scores for class 1 severity and class 3 severity were the best and class 2 severity received a lower precision score. This means that a good proportion of positive identifications were actually correct for class 1 and class 3 classification of severity and less so for class 2. The recall scores for severity class 2 and 3 are not good compared to the the class 1. Thus, the proportion of actual positives that were identified correctly was great for severity class 1 but not good for severity class 2 and 3. This is consistent with the confusion matrix obtained.

6.5.3 Predicting severity using target information and tweet text

Training and validation

This is a table showing the training and validation accuracy over 10 epochs of training.

Epochs	Training accuracy	Validation accuracy
Epoch 1/10	55.75 %	58.59 %
Epoch 2/10	58.99 %	59.52 %
Epoch 3/10	59.54 %	58.69 %
Epoch 4/10	59.43 %	59.88 %
Epoch 5/10	60.31 %	61.93 %
Epoch 6/10	63.84 %	64.87 %
Epoch 7/10	65.67 %	66.46 %
Epoch 8/10	67.55 %	67.85 %
Epoch 9/10	68.21 %	67.75 %
Epoch 10/10	69.38 %	68.72 %

It is evident from the table above that our model has a validation accuracy of 68.72 %.[1]

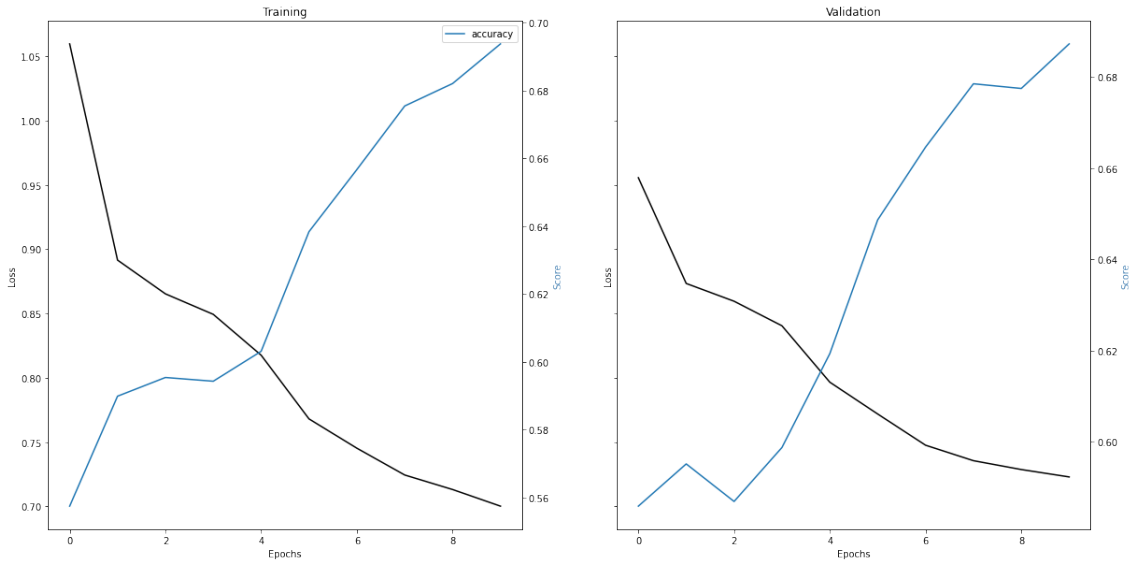

Figure 6.12: A graph representing training accuracy/loss and validation accuracy/loss respectively for prediction of severity using tweets and target as input.

The same data about the training epochs is visualised in the graph with the training and validation loss added.

[1]The model was fit with epochs = 10, batch_size = 256, and a validation split of 70-30/0.3

Testing results

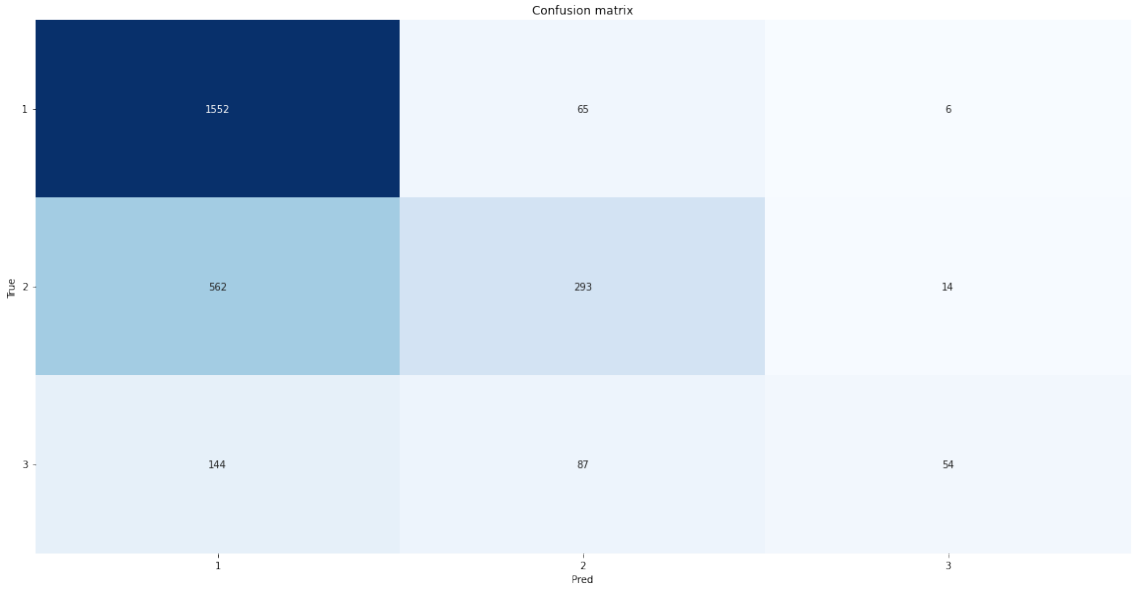

Figure 6.13: Confusion matrix for prediction of severity using tweets and target as input.

The confusion matrix shows us that the severity prediction for the severity class 1 (negligible severity) was very accurate. However, the results for the severity class 2 (moderate severity) and severity class 3 (severe severity) were not good with many tweets of severity class 2 and 3 classified as class 1 severity tweets. The classification was much better for the severity class 2 when compared to the classification done using the tweet text alone.

```
Accuracy: 0.68
Auc: 0.79
Detail:
              precision    recall  f1-score   support

           1       0.69      0.96      0.80      1623
           2       0.66      0.34      0.45       869
           3       0.73      0.19      0.30       285

    accuracy                           0.68      2777
   macro avg       0.69      0.49      0.52      2777
weighted avg       0.68      0.68      0.64      2777
```

Figure 6.14: Classification report for prediction of severity using tweets and target as input.

The classification report for the test data shows an accuracy of 68%. The precision scores for class 1 severity and class 3 severity were the best and class 2 severity received a lower precision score. This means that a good proportion of positive identifications were actually correct for class 1 and class 3 classification of severity and less so for class 2. The classification of class 3 severity was the best in this regard and the difference between class 1 and class 2 was not as drastic as previously observed when classifying tweets according to severity using tweet text alone. The recall scores for severity class 2 and 3 are not good compared to the the class 1. But the scores are significantly better for class 2 severity when compared to the classification according to severity using tweet text alone. Thus, the proportion of actual positives that were identified correctly was great for severity class 1 but not good for severity class 2 and 3. The results for class 2 severity were much improved when target was included as a feature. This is consistent with the confusion matrix obtained.

Chapter 7

Conclusion and Future Scope

7.1 Conclusion

The results indicate that information about the target of hate speech improves the classification of severity of hate as opposed to using just the tweet text. As such, information about the tweet targets is a helpful feature that can be explored further.

This study helped us re-discover Bi-LSTM potential. The custom designing of word vector files and stopword corpus led us to new frameworks like word2Vector, stanford-nltk, inltk, and many more. We are more than proud to say that the learnings we take away from this study will help us in our future NLP-related projects and research.

Although similar projects exist for standard languages like English, Hindi, Bengali, Tamil, etc, Hinglish is an exception. With no proper Hinglish corpus available, we built this project from scratch using our datasets, word vectors, and stopwords. We are more than happy to give our peers the assets related to this project. We look forward to publishing a research paper for the same.

It has been an absolute honor to work with Dr.Sunil Saumya. We look forward to working with Dr. Saumya in the future.

7.2 Future Scope

After analysing the data we observed that individual targets are most common and can be broken down into further categories to improve the dataset. The category of individual target can be further broken down into: individual-personal, individual-organisation, and individual-religion. Increasing the size of the dataset can also improve our word-vectors and improve the classification.

References

[1] P. Rani, S. Suryawanshi, K. Goswami, B. R. Chakravarthi, T. Fransen, and J. P. McCrae, "A comparative study of different state-of-the-art hate speech detection methods in Hindi-English code-mixed data," in *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 42–48. [Online]. Available: https://aclanthology.org/2020.trac-1.7

[2] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava, "A dataset of Hindi-English code-mixed social media text for hate speech detection," in *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*. New Orleans, Louisiana, USA: Association for Computational Linguistics, Jun. 2018, pp. 36–41. [Online]. Available: https://aclanthology.org/W18-1105

[3] R. Kumar, A. N. Reganti, A. Bhatia, and T. Maheshwari, "Aggression-annotated corpus of Hindi-English code-mixed data," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://aclanthology.org/L18-1226

[4] S. Dowlagar and R. Mamidi, "A survey of recent neural network models on code-mixed indian hate speech data," in *FIRE 2021: Forum for Information Retrieval Evaluation, Virtual Event, India, December 13 - 17, 2021*, D. Ganguly, S. Gangopadhyay, M. Mitra, and P. Majumder, Eds. ACM, 2021, pp. 67–74. [Online]. Available: https://doi.org/10.1145/3503162.3503168

[5] K. Sreelakshmi, B. Premjith, and K. P. Soman, "Detection of hate speech text in hindi-english code-mixed data," *Procedia Computer Science*, vol. 171, pp. 737–744, 2020.

[6] S. Kamble and A. Joshi, "Hate speech detection from code-mixed hindi-english tweets using deep learning models," *CoRR*, vol. abs/1811.05145, 2018. [Online]. Available: http://arxiv.org/abs/1811.05145

[7] P. Vijayaraghavan, H. Larochelle, and D. Roy, "Interpretable multi-modal hate speech detection," 2021. [Online]. Available: https://arxiv.org/abs/2103.01616