Bachelor of Engineering

Project Presentation

DEPARTMENT OF COMPUTER ENGINEERING
SHAH AND ANCHOR KUTCHHI ENGINEERING COLLEGE
CHEMBUR, MUMBAI – 400088.
2020 – 2021

BE PROJECT

ON

CLEANSE
An NLP Based Hate Speech Classifier and Euphemistic Text Generator

By

Rutvi Bheda - BE3-21

Darsh Bhimani - BE3-22

Femin Dharamshi - BE3-27

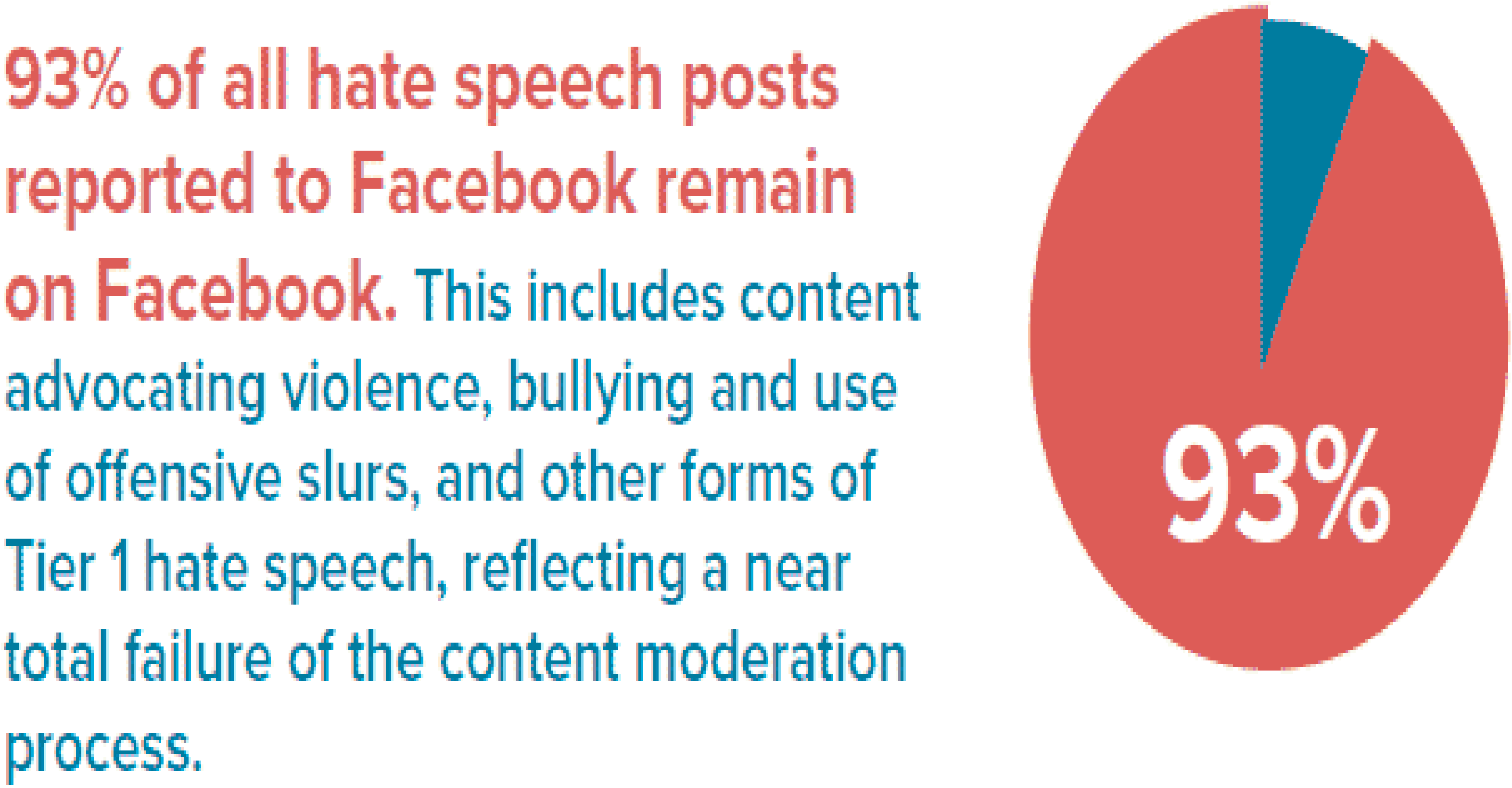
Under the guidance of

Deepti Nikumbh
(Guide)

Priyanka Abhyankar
(Co-Guide)

Introduction

- Hate Speech in written text and/or toxic comments are rising as the reach of social media continues to grow across the world.
- A new study conducted by Norton by Symantec found that eight out of ten people have experienced some form of online harassment in India. The most common forms of online harassment were abuse and insults (63 per cent) and malicious gossip and rumours (59 per cent).
- Hate speech, has "*exacerbated societal and racial tensions, inciting attacks with deadly consequences around the world*", according to an open letter released by nearly two dozen independent United Nations experts.

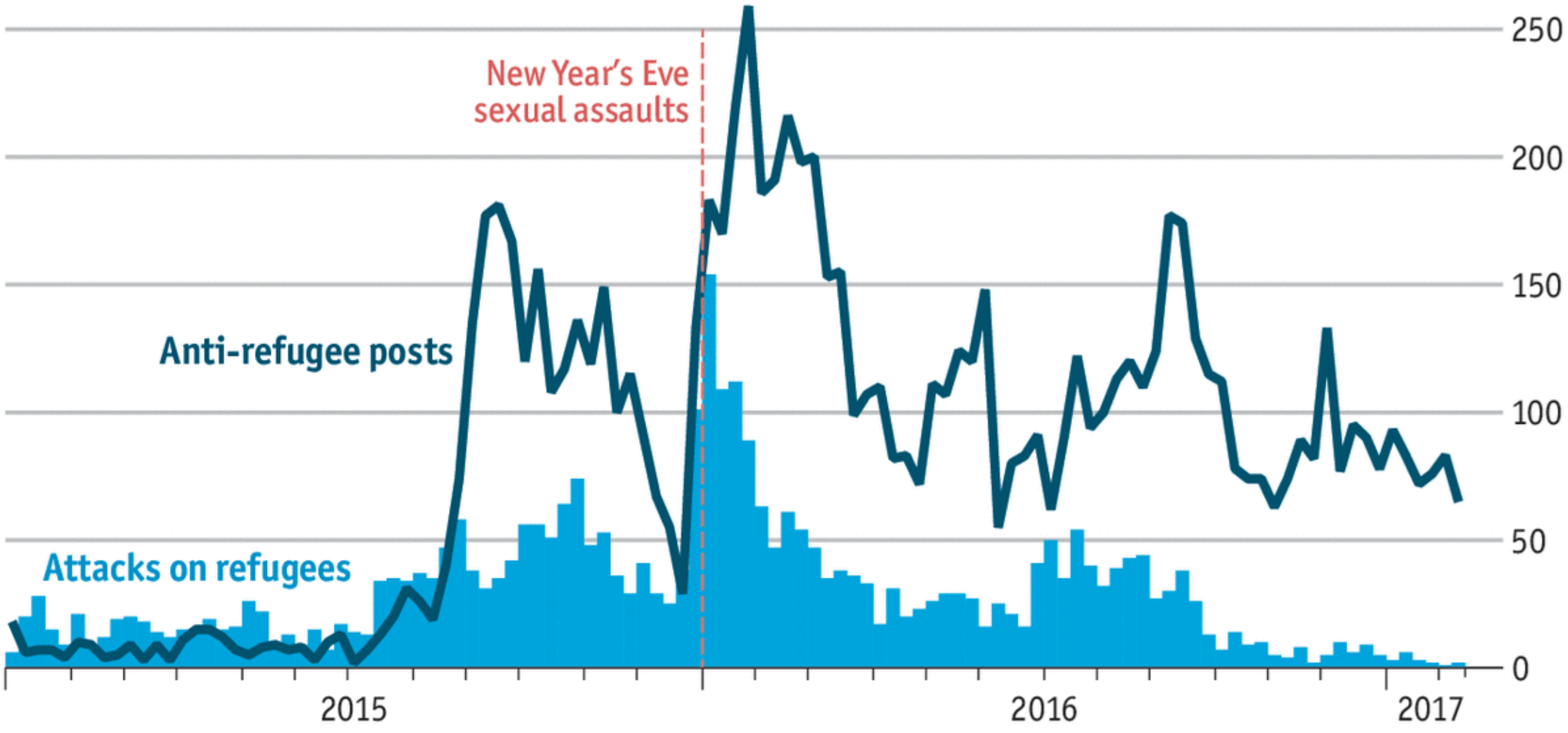


93% of all hate speech posts reported to Facebook remain on Facebook. This includes content advocating violence, bullying and use of offensive slurs, and other forms of Tier 1 hate speech, reflecting a near total failure of the content moderation process.

Fig 1. Statistics of hate speech content on Facebook

Anti-social media

Germany, number of anti-refugee posts on the AfD’s Facebook page and attacks on refugees



Source: “Fanning the Flames of Hate: Social Media and Hate Crime”, by K. Müller and C. Schwarz, Dec 6th 2017

Economist.com

Fig 2. Influence of hate speech over the past few years

Problem Statement

The level of hateful and abusive comments on the Internet is growing to the point where the existing structures are not in a position to fight effectively. The quick elimination of such comments is an ineffective disciplinary step and also removes useful comments. In addition, it is also delayed.

Literature Survey

Sr No.	Title	Publishing Year	Author	Description
1.	Hate Speech On Twitter.A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection	IEEE-2018	Hajime Watanabe, Mondher Bouazizi, Tomoaki Ohtsuki	This paper initially explains what exactly is Hate Speech and how has it impacted people by the growth on internet.It further classifies a particular text into 3 classes: Clean, Offensive and Hateful followed by the dat apre-processing like removing the urls and tags present in it.The author explains the 4 main features that can used : Sentiment based features, Semantic features, Unigram features and Pattern features.Sentiment features would allow to extract the polarity of the text, Semantic Analysis looks after the expression if it contains punctuations at proper places, Unigram features allows us to detect any explicit form of hate speech and Pattern allows to detect any longer forms of hate speech.The author says that the combination of these 4 features would help in detection of the hate speech .

Literature Survey

Sr No.	Title	Publishing Year	Author	Description
2.	A Bag-of-Phonetic-Codes Model for CyberBullying Detection in Twitter	IEEE-2018	Ankita Shekhar, M Venkatesan	In this paper, the author talks about pronunciation features to detect cyber-bullying. The pronunciation features can help in the identification of misspelled and censored words. It also talks about the approaches for Sentiment Analysis .
3.	Hate speech classification in social media using emotional analysis	IEEE-2018 7th Brazilian Conference on Intelligent Systems	Ricardo Martins, Marco Gomes, Jose Jo ´ ao Almeida, Paulo Novais, Pedro Henriques	This paper presents the combination of lexicon based and machine learning approaches to predict hate speech contained in a text, using an emotional approach through sentiment analysis.

Literature Survey

Sr No.	Title	Publishing Year	Author	Description
4.	Detection of Hate and Offensive Speech in Text	Springer-2019	Abid Hussain Wani(B) , Nahida Shafi Molvi, and Sheikh Ishrah Ashraf	The dataset in initially preprocessed to remove all the duplicated and redundant text including URLs ,punctuations, space patterns, etc. Conventional ML methods and Deep Learning methods were applied on it. They performed Conventional ML techniques in combination to Bag of Words and TFIDF where they found that Decision Tree Classifier using BiGrams with Bag of Words and TFIDF combined to have the heighest F1 Score. Doing the same with DeepLearning Techniques(CNN and LSTM) combined with Glove / Word2Vec / Random Embeddings, the authors concluded that LSTM with WOrd2Vec provides highest classification performance.

Literature Survey

Sr No.	Title	Publishing Year	Author	Description
5.	Extracting Word Synonyms from Text using Neural Approaches	The International Arab Journal of Information Technology, Vol. 17, No. 1, January 2020	Nora Mohammed	Word2vec was used to construct word embeddings and performed a qualitative evaluation of the most similar words to a few target words.Similarity by semantic analysis doesn’t necessarily indicate synonymy.Concludes that embedding created using word2vec should be input to a neural network for synonymy detection.

Research Gaps

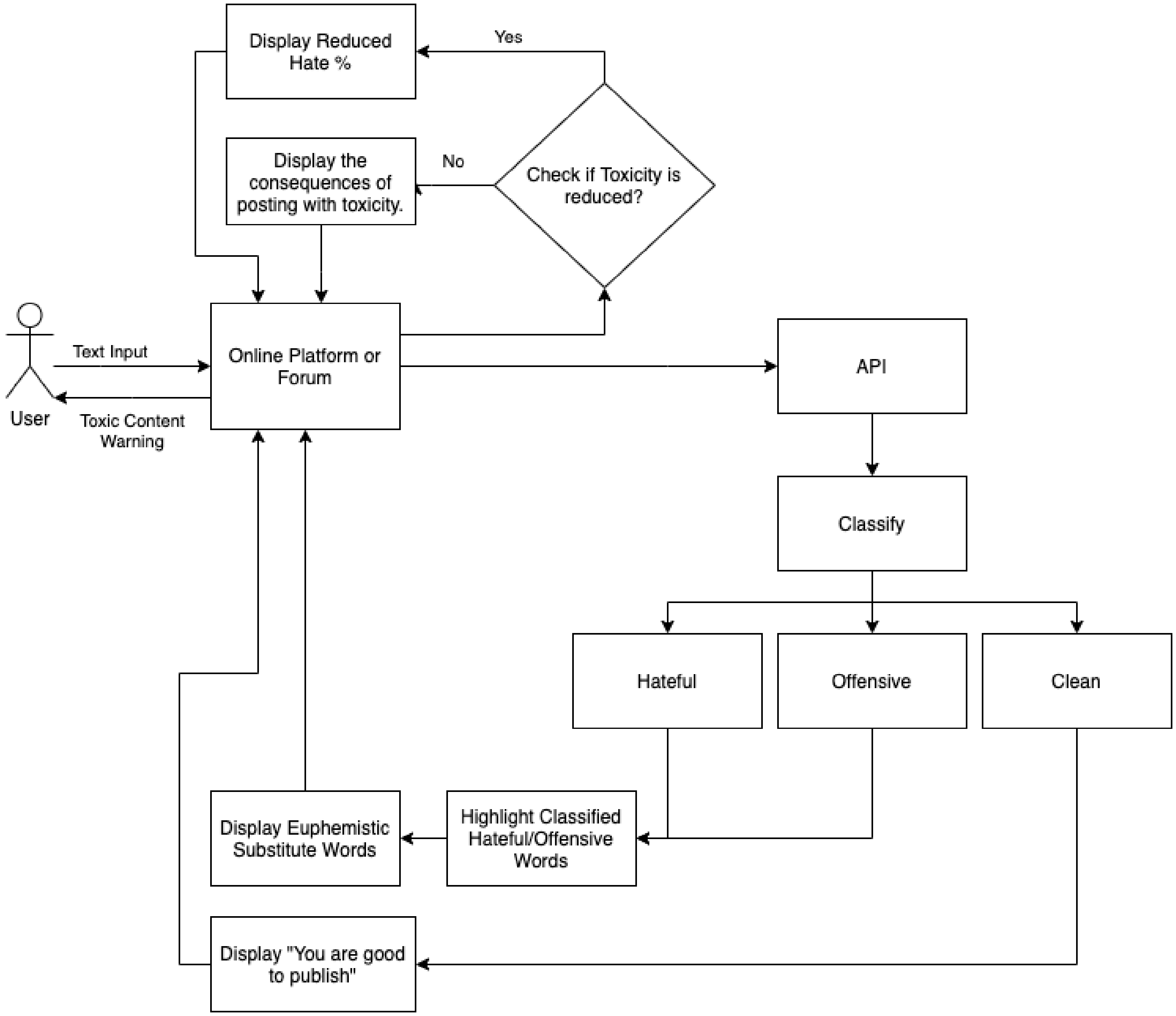
- The literature review did reveal certain gaps and issues in the systems that have been studied. One of the gap found was that how the biases of the labeler may affect the classifier so modeled. This issue must be checked in the creation of the classifier model in order to ensure impartiality.
- While the research on finding words based on similarity is quite advanced, methods studied do not find an inoffensive substitution of the word, i.e. euphemism.
- Another gap is that although the classifier may work good and give good results but it needs to be updated regularly and continuously. We know that English Language keeps on updating as well as evolving each day.
- A final gap observed was the inadequacy of semantic analysis being used alone to generate synonyms. It ought to be assisted by other approaches to be improved.

Objectives

1. Real time classify text as toxic or non toxic to identify hatefulness and suggestion of euphemistic substitution.
2. Design an algorithm to find less toxic substitution for a hateful phrase .
3. Measure hate objectively, reducing subjectivity.
4. To have automatic moderator which acts before publishing.

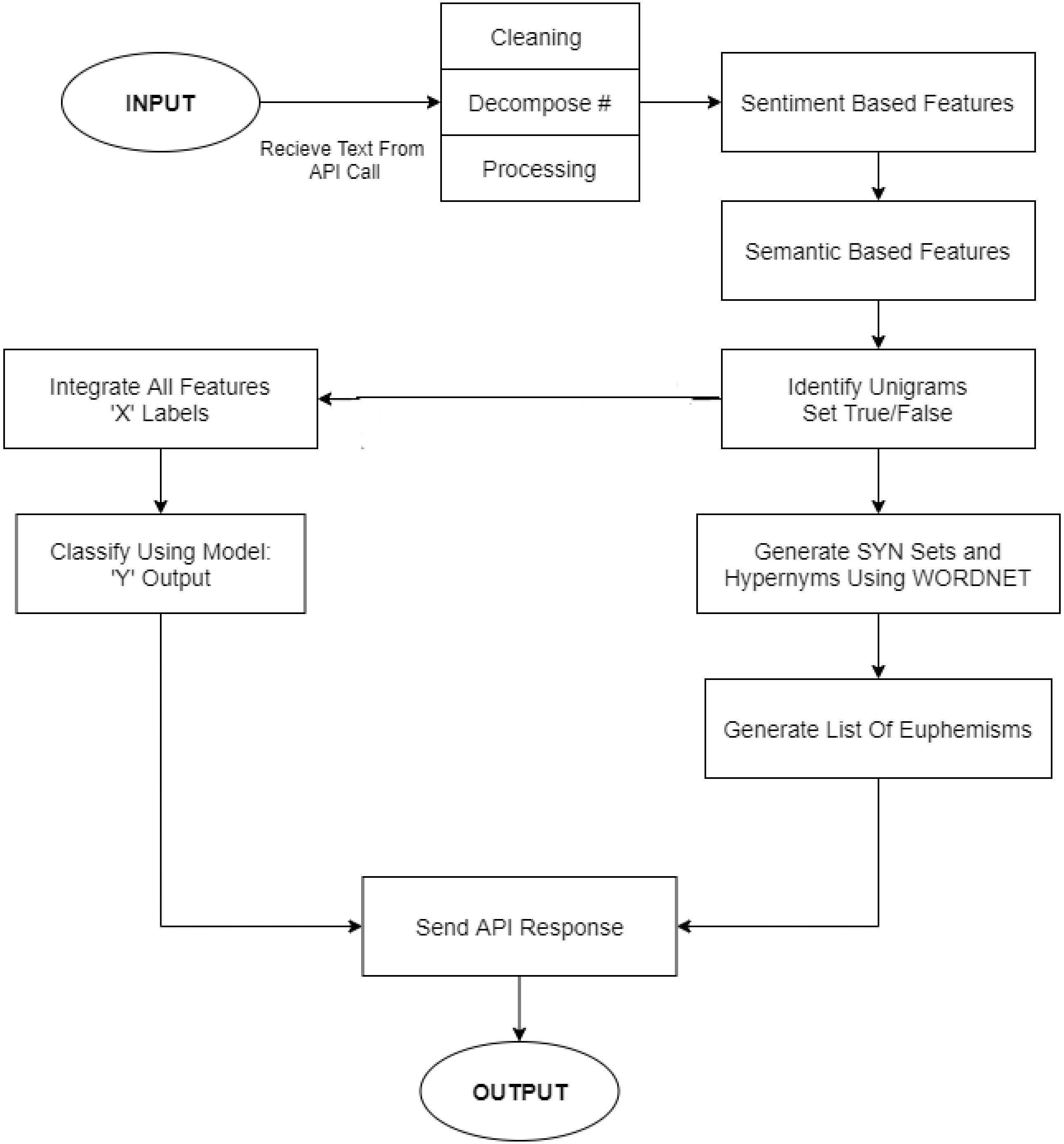
Methodology

System Flow


Display Reduced Hate %
Yes
Display the consequences of posting with toxicity.
No
Check if Toxicity is reduced?
User
Text Input
Toxic Content Warning
Online Platform or Forum
API
Classify
Hateful
Offensive
Clean
Display Euphemistic Substitute Words
Highlight Classified Hateful/Offensive Words
Display "You are good to publish"

Methodology

Data Flow



INPUT

Recieve Text From API Call

Cleaning

Decompose #

Processing

Sentiment Based Features

Semantic Based Features

Identify Unigrams
Set True/False

Integrate All Features
'X' Labels

Classify Using Model:
'Y' Output

Generate SYN Sets and
Hypernyms Using WORDNET

Generate List Of Euphemisms

Send API Response

OUTPUT

Expected Outcomes:

The final product of this project must be able to:

1. Quantify hate speech in a text
2. To perform [Objective 1] in an unbiased and robust manner
3. Suggest replacements for Hateful word
4. Measure change in the amount of hate (before and after substitution)
5. Enable platforms to get Hate Report easily and quickly from the API

Datasets

- **Automated Hate Speech Detection and the Problem of Offensive Language Dataset:** The dataset contains tweets that are annotated into 3 classes ie Hateful, Offensive and Clean.
- **NAACL_SRW Dataset:** A hate speech twitter dataset which contains 16907 tweet ids which are classified as racial, sexism, or none

Data Pre-processing

- Cleaning the tweet (Removal of URLs and user tags).
- Removal of hashtags.
- Tokenization and Stemming (with the help of Porter Stemmer available in the NLTK library).

Semantic Analysis

- Semantic features looks after the punctuation and capitalizations in the text.
- The various features that are considered here are: Question Marks, Exclamation Marks and Capitalized Words.
- This feature can change the meaning of the sentence when applied.

Sentiment Analysis

- This feature is important in understanding the meaning and context of the text.
- The feature considered here are-dual: positive and negative, scale and binary sentiment.
- We use these features because not all negative speech is hate speech.

Unigram Based Feature

- These features represent words or expressions that are explicitly hateful
- They are considered features only when they are marked by a minimum number of occurrences
- Such most frequent words are collected and used as features

Future Scope

- Extend the application to determine user as per their past behaviour in hate speech
- To identify hateful phrases or statements present rather than just words
- Giving a warning to user for excessive usage of hate speech

References

1. H. Watanabe, M. Bouazizi and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," in IEEE Access, vol. 6, pp. 13825-13835, 2018, doi: 10.1109/ACCESS.2018.2806394.
2. A. Shekhar and M. Venkatesan, "A Bag-of-Phonetic-Codes Modelfor Cyber-Bullying Detection in Twitter," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, 2018, pp. 1-7, doi: 10.1109/ICCTCT.2018.8550938.
3. R. Martins, M. Gomes, J. J. Almeida, P. Novais and P. Henriques, "Hate Speech Classification in Social Media Using Emotional Analysis," 2018 7th Brazilian Conference on Intelligent Systems (BRACIS), Sao Paulo, 2018, pp. 61-66, doi: 10.1109/BRACIS.2018.00019.
4. Wani A.H., Molvi N.S., Ashraf S.I. (2020) Detection of Hate and Offensive Speech in Text. In: Tiwary U., Chaudhury S. (eds) Intelligent Human Computer Interaction. IHCI 2019. Lecture Notes in Computer Science, vol 11886. Springer, Cham. https://doi.org/10.1007/978-3-030-44689-5_8

References

5. Mohammed, Nora. (2019). Extracting Word Synonyms from Text using Neural Approaches. The International Arab Journal of Information Technology. 45-51. 10.34028/iajit/17/1/6.

6. Davidson, Thomas & Warmsley, Dana & Macy, Michael & Weber, Ingmar. (2017). Automated Hate Speech Detection and the Problem of Offensive Language.

7. A. Veloso, W. Meira Jr, T. A. Macambira, D. O. Guedes, and H. M. P.de Almeida, "Automatic moderation of comments in a large on-line journalistic environment.".

8. Nobata, Chikashi & Tetreault, Joel & Thomas, Achint & Mehdad, Yashar & Chang, Yi. (2016). Abusive Language Detection in Online User Content. 145-153. 10.1145/2872427.2883062.

Any Questions?

Thank You