Hate Speech Detection

Devansh Mody (1130532)
Department of Computer Science
Lakehead University
dmody@lakeheadu.ca

Yidong Huang (1116935)
Department of Computer Science
Lakehead University
yhuang36@lakeheadu.ca

***Abstract*—There is a significant development in the deep learning and natural language processing algorithms in the last few decades. Because of that text summarization, question answering system, chat bots and many more such task have become easier to implement. Presently in the era of internet and electronic devices, hate speech has proven consequential to youth internet users and previous methods relied heavily on the use of manually developed dictionaries. Due to the massive rise of user generated web content on social media, the amount of hate speech is also steadily increasing. Therefore detecting hate speech on the internet nowadays is a state of the art topic in the machine learning and natural language processing field. This paper will focus on building a system to detect whether a text is hate speech or not. Further in this paper we will describe the most effective methods proposed by few authors in their research work on hate speech and different types of approach to solve the problem.**

Keywords: Convolutional Neural Network (CNN), Unigram (UI), Bigram (BI), Trigram (TI), Long Short Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Recurrent Neural Network (RNN), Support Vector Machine (SVM), Bidirectional Encoder Representations from Transformers (BERT), Hate Speech (HS), Parts of speech (POS), Named entity recognition (NER), Term frequency inverse document frequency (TFIDF), Gated Recurrent Unit (GRU).

I. Introduction

In today's digital world also known as the era of the internet, people are living around social media like youtube, twitter, and the use of social media is a very common thing for teenagers, but this may lead to bad and harmful situations, because of the hate speech in our world wide web. Hate speech used to exists before digital media but now in age of internet it spreads faster on social media platforms and other kinds of media and its difficult to track the source of the hate speech or the one who is spreading hate and provoking anger. According to the paper from Michel, the meaning of hate speech is ”speech designed to promote hatred on the basis of race, religion, ethnicity or national origin poses vexing and complex problems for contemporary constitutional rights to freedom of expression” [19]. With the time change, the hate speech can extend to personal attacks like body shame, or other bullying behaviour on the world wide web. This is connected even with concrete public health issues, since recent studies show that victims are more likely to suffer from psycho-social difficulties and affective disorders [20].

In this paper, we are going to present a system for detecting hate speech on the internet. The system aims to support the people who have the right to supervise social media(e.g. the manager of Twitter). This paper would explain a system based on different algorithms which include KNN, CNN, LSTM, Naive Bayes and the pre-train model, BERT. At the same time, this paper will take the text similarity into account, as a part of checking hate speech and different evaluation metrics like ROUGE-L, BLEAU score will also be used to with LIME methods to interpret the performance of the model. Below figure gives the idea of our approach our first aim is to build a one of the proper hate speech model using pretrained model like BERT by changing the model according to our requirements and perform a proper context and subjective analysis to achieve the highest output, then we intend to check the interpretability of the model using LIME methods and we would like to develop a voting mechanism to effectively predict the output of different models and make the final decision based on the output.

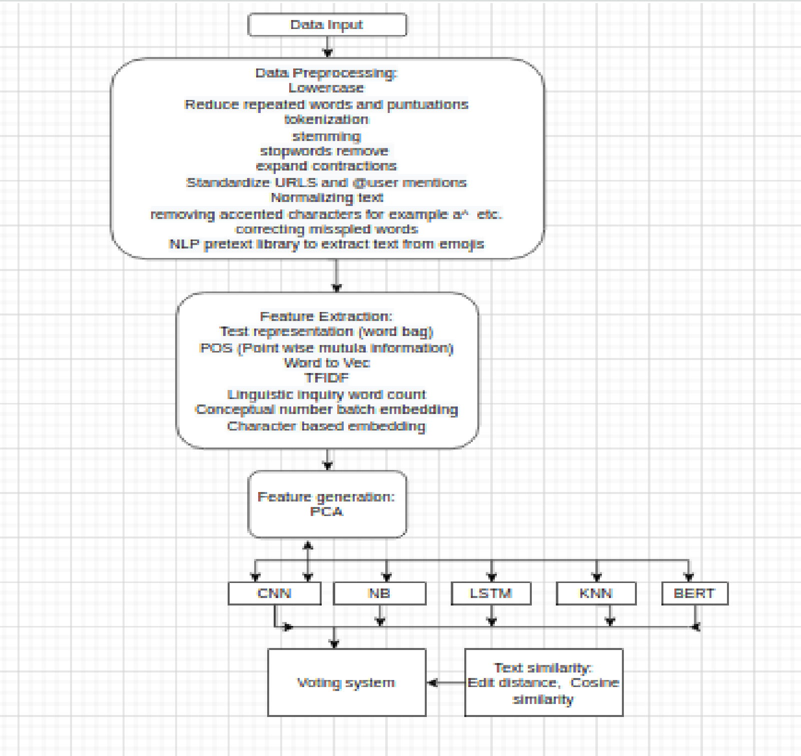
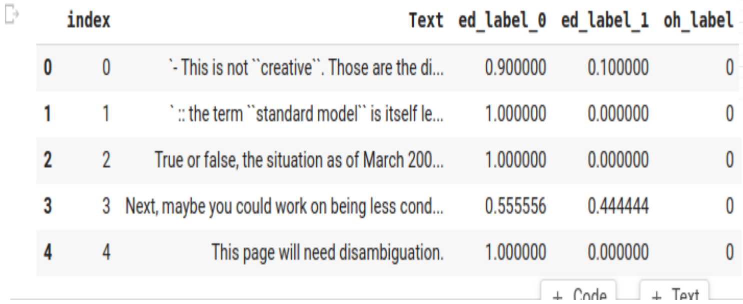

Data Input
Data Preprocessing:
Lowercase
Reduce repeated words and puntuations
tokenization
stemming
stopwords remove
expand contractions
Standardize URLS and @user mentions
Normalizing text
removing accented characters for example a^ etc.
correcting misspled words
NLP pretext library to extract text from emojis
Feature Extraction:
Test representation (word bag)
POS (Point wise mutula information)
Word to Vec
TFIDF
Linguistic inquiry word count
Conceptual number batch embedding
Character based embedding
Feature generation:
PCA
CNN
NB
LSTM
KNN
BERT
Voting system
Text similarity:
Edit distance, Cosine similarity

Fig. 1. Basic flow diagram of our approach

The data for our project is available on the website https://hatespeechdata.com/, the website has a huge amount of data about the topic hate speech, also the website provides us with the labeled dataset. The size of the data is about 300 thousands of samples, and it will take about 300 MB in space. Basically, the dataset has five columns, index, Text, ed_label_0, ed_label_1, oh_label. The column oh_label contains two values either 0 or 1 which means hate speech or not an hate speech for a given text. Below given is the snapshot of the dataset.



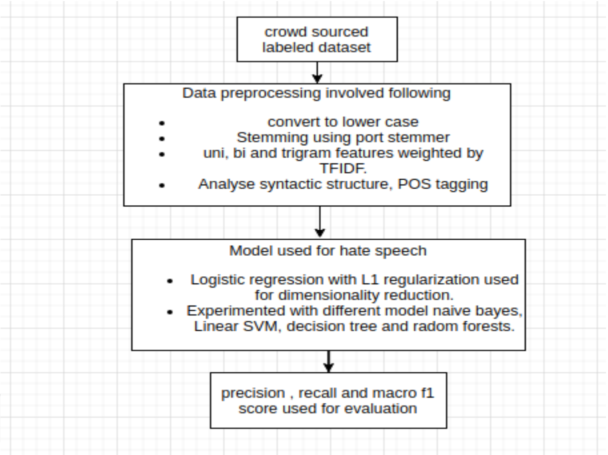
	index	Text	ed_label_0	ed_label_1	oh_label
0	0	`- This is not ``creative``. Those are the di...	0.900000	0.100000	0
1	1	`:: the term ``standard model`` is itself le...	1.000000	0.000000	0
2	2	True or false, the situation as of March 200...	1.000000	0.000000	0
3	3	Next, maybe you could work on being less cond...	0.555556	0.444444	0
4	4	This page will need disambiguation.	1.000000	0.000000	0

Fig. 2. Basic flow diagram of our approach

II. Literature Review

Following is the summary of research papers surveyed for literature review some of their methods will be incorporated in our research work with additional functionalities and additional analysis based on our extended research work.

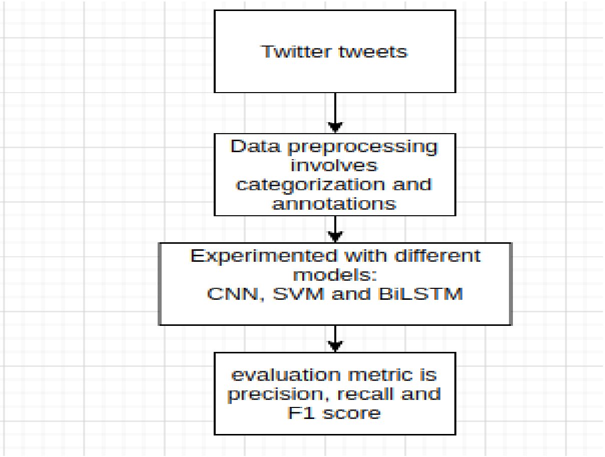
- The aim of this paper [1] is to build a detection system for hate speech, we can conclude this paper by three parts: A) Data, the data is from Twitter, using the Twitter API, they build a speech lexicon with words and phrases, and then the sample will manually coded by CrowdFlower (CF) workers, workers were asked to label each tweet as one of three categories: hate speech, offensive but not hate speech, or neither offensive nor hate speech. B) Features, firstly, they lowcased the sample, and then the data were stemmed by using the Porter stemmer, then they created bigram, unigram, and trigram features, each one was done by the TFIDF. C) Model, they run many different algorithms on this paper, which include, logistic regression, naive bayes, decision trees, random forests, and linear SVMs. On those above models, Logistic Regression and Linear SVM are much better than other algorithms. In this process, they use a one versus rest framework on the SVM, I think we also can use one versus one in our approach. Below figure describes the approach used by the author.
- This paper [2] uses the machine learning method to detect hate speech with the context. The models of this paper used are logistic regression and neural networks. Firstly, they consider many different features, but we need to clear that the maximum length of sequence they consider is 125. This paper used Word level and Character level Ngram Features, LIWC Feature, NRC Emotion Lexicon Feature. And then based on those features, they run the two algorithms, Logistic regression and neural networks.



crowd sourced labeled dataset
Data preprocessing involved following
convert to lower case
Stemming using port stemmer
uni, bi and trigram features weighted by TFIDF.
Analyse syntactic structure, POS tagging
Model used for hate speech
Logistic regression with L1 regularization used for dimensionality reduction.
Experimented with different model naive bayes, Linear SVM, decision tree and radom forests.
precision , recall and macro f1 score used for evaluation

Fig. 3. Flow chart of research approach.

- The objective of this paper [3] can be concluded as three things: a)Offensive Language Detection ; B)Categorization of Offensive Language; C)Offensive Language Target Identification. To achieve that objectives, this paper mainly do two things, firstly, create dataset with annotation, they name dataset as “Offensive Language Identification Dataset ”(OLID), which is a new large scale dataset of tweets with annotations, the annotations in include keywords(for example, “medical marijuana”, “you are”) and so on. Step 2, Run 3 different algorithms based on the annotation dataset. The algorithms are: SVM, BiLSTM, CNN. Below figure describes the approach used by the author.



Twitter tweets
Data preprocessing involves categorization and annotations
Experimented with different models:
CNN, SVM and BiLSTM
evaluation metric is precision, recall and F1 score

Fig. 4. Flow chart of research approach.

- The idea of this paper [5] is to design a system which can prevent hate speech, the way how they did this can be concluded as 3 steps. Firstly, they collect the data from Reddit and Gab, which is different from most papers, the data will be a full conversation, so the researcher will have a context for samples. Step 2, the samples are manually labeled as hate or non hate speech by Mechanical Turk workers, and then all samples will be replied, the reply also will be recorded. Step 3 Generative Intervention, the method shown on our figure where c is the conversation, r is the corresponding intervention response, and D is the dataset.(Obj is the objective of generating response) Then based on the samples and replies, they run four different machine learning algorithms to learn how to pick responses, the best performance is from CNN algorithm.

$$Obj = \max \sum_{(c,r)\in D} \log p(r|c)$$

Fig. 5. Generative intervention method

- The idea of this paper [6] is combine the text in image and text to rich our samples, and then the author will convert image and text information to different features, the image will generate 2048 features, text will generate 150 feature, but there two dataset for text, one is from image samples, another one is from Tweet text. Therefore, the system will get 2348 feature, the author plan to run a CNN model for those features, the model of design show as figure below.

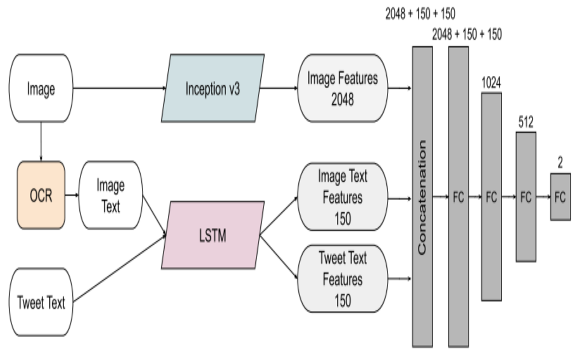


Fig. 6. Multimodel

- This paper [8] aims to detect hate speech and the type of hate speech from different languages, it has three different sub tasks on this paper. The first one is to simply check whether it is a hate speech content or non-offensive content, the second task is based on the task 1, if the content is belong to hate speech, the task 2 will identify the type of hate speech, and then this paper want to target(mark) the post of hate speech. The algorithm they used is LSTM, but on the method of evaluations, this paper introduces two ways: weighted F1 and Macro F1. The weighted F1 score calculates the F1 score for each class independently. When it adds them, it uses a weight based on the number of true labels of each class.The ’macro’ calculates the F1 separately for each class but does not use weights for the aggregation.
- This paper [12] is a kind survey report, it takes the data from Twitter, and throws out a comparative research topic: what is the difference between hate speech instigators and hate speech targets? To figure out the answer to this question, they extract some information from data: account characteristics(Gender, TimeZone, Invalid image), Personality Traits (Agreeableness, Conscientiousness, Extraversion, Emotional range, and Openness). The measure for account characteristics is simply count, but for the second one, the author of this paper used a pre trained model which is IBM Watson Personality API, this API describes scores [0,1] that reflect the normalized percentile score for the characteristic.
- This paper [14] aims to build a hate speech detection based on the lexicon sentiment. To get the lexicon sentiment, this paper describes this as three main steps, the first step is subjectivity detection, which will be done by isolating sentences that have subjective expressions from those normal sentences. In the second step, this paper builds a lexicon of hate related words with a rule based method using subjective features identified from the sentences and semantic features learned directly from the corpus. In the end this paper will build a classifier that uses the features from lexicon and use it to test the new document.
- This paper [15] aims to detect hate speech in social media like Twitter, the approach of this paper is about structure of sentence. Most existing efforts to measure hate speech require knowing the hate words or hate targets apriori. Compared to other measures, this paper provides a new way based on logic structure of sentences, which will consider a sentence from I, ¡intensity¿, ¡userintent¿, ¡hate target¿, for example I really hate Asian people. After this design they count the all used hate word from Twitter(this is because Twitter provides some related API). Finally, they will analyze the targets of hate speech on the internet.
- The core idea of this paper [16] is about how to build features from text for detecting hate speech content. Specially, this paper annotates samples by two methods: firstly, they ask Amazon Mechanical Turk (MTurk) workers to annotate these posts to cover three facets. In addition to classifying each post into hate, offensive, or normal speech, and they also select the target’s communities(like victims of speech); secondly, they use “rationales” to highlight parts of the text that could justify their classification decision. Then the author named this kind of sample as HateXplain. For algorithms, this paper runs the CNNGRU, BiRNN, BiRnnAttention (the

difference with BiRNN is that BiRnnAttention includes the attention layer) and BERT pre trained model. Below figure describes the approach used by the author.

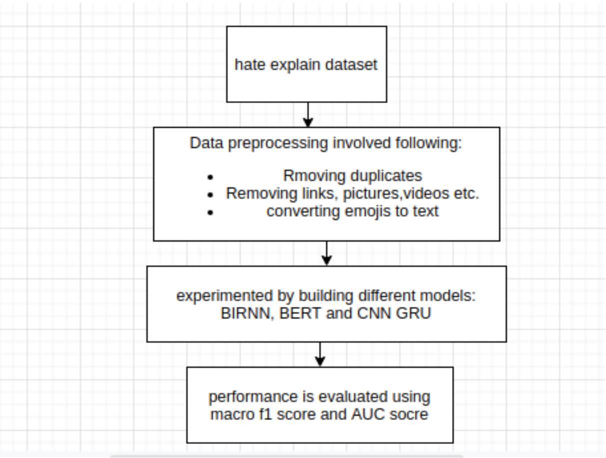

hate explain dataset
Data preprocessing involved following:
Rmoving duplicates
Removing links, pictures,videos etc.
converting emojis to text
experimented by building different models: BIRNN, BERT and CNN GRU
performance is evaluated using macro f1 score and AUC socre

Fig. 7. Flow chart of research approach.

III. Discussion

Most of the researchers have used LSTM, SVM, random forests, BERT and other similar models for hate speech detection. Data preprocessing steps involved converting to lower case, POS tagging, stemming/Lemmetization, stop words removal. Also various approach like unigram, bigram and trigram based with TFIDF was used. Most of the researchers created their own data sets by annotations. The common problems faced by the authors during their research work are Lack of external context such as profile bio, user gender, history of posts etc., which might be helpful in the classification task, building a multi language model, context analysis and subjective analysis.

IV. Conclusion

Henceforth we would like to carry out an voluminous research analysis to detect hate speech in text by understanding the sentiment and the context of a text, tweet and other short sentences. Furthermore numerous patterns in text data needs to be explored to understand and decode hate speech. Additionally also our project can be used as a preprocessing step to remove biases like hate speech before developing other projects like text summarization, text generation and other projects.

References

[1] Thomas Davidson, Dana Warmsley, Michael Macy and Ingmar Weber, ”Automated Hate Speech Detection and the Problem of Offensive Language”, arXiv:1703.04009v1 [cs.CL] 11 Mar 2017.

[2] Lei Gao, Ruihong Huang, ”Detecting Online Hate Speech Using Context Aware Models”, arXiv:1710.07395v2 [cs.CL] 22 May 2018.

[3] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra and Ritesh Kumar, ”Predicting the Type and Target of Offensive Posts in Social Media”, Proceedings of NAACL-HLT 2019.

[4] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, Dit-Yan Yeung, ”Multilingual and Multi-Aspect Hate Speech Analysis”, arXiv:1908.11049v1 [cs.CL] 29 Aug 2019.

[5] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding and William Yang Wang, ”A Benchmark Dataset for Learning to Intervene in Online Hate Speech”, arXiv:1909.04251v1 [cs.CL] 10 Sep 2019.

[6] Raul Gomez, Jaume Gibert, Lluis Gomez and Dimosthenis Karatzas, ”Exploring Hate Speech Detection in Multimodal Publications”, Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV (2020) 1459-1467.

[7] Nina Bauwelinck, Gilles Jacobs, Veronique Hoste and Els Lefever, ”LT3 at SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (hatEval)”, Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages (436–440).

[8] Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia and Aditya Patel, ”Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages”, FIRE ’19: Proceedings of the 11th Forum for Information Retrieval Evaluation December (2019) Pages (14–17).

[9] Ona de Gibert, Naiara Perez, Aitor Garcıa-Pablos and Montse Cuadros, ”Hate Speech Dataset from a White Supremacy Forum”, arXiv:1809.04444 [cs.CL] 12 Sep 2018.

[10] Zeerak Waseem and Dirk Hovy, ”Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter”, Proceedings of NAACL-HLT (2016), pages (88–93).

[11] Elisabetta Fersini, Debora Nozza and Paolo Rosso, ”Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI)”, CEUR Workshop Proceedings (2018).

[12] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna and Elizabeth Belding, ”Peer to Peer Hate: Hate Speech Instigators and Their Targets”, arXiv:1804.04649v1 [cs.SI] 12 Apr 2018.

[13] Poletto F, Basile V, SanguBasileinetti M and Bosco C, ”Resources and benchmark corpora for hate speech detection: a systematic review”, Lang Resources Evaluation (2021).

[14] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien and Jun Long, ”A Lexicon-based Approach for Hate Speech Detection”, International Journal of Multimedia and Ubiquitous Engineering (2015) 10(4) 215-230.

[15] Silva, LMondal, MCorrea, DBenevenuto and FWeber, ”Analyzing the Targets of Hate in Online Social Media”, Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016 (2016) 687-690.

[16] Mathew B, Saha P, Yimam S, Biemann C, Goyal P and Mukherjee A, ”HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection”, ArXiv: 2012.10289 18 Dec 2020.

[17] Cao, RLee R and Hoang T, ”DeepHate: Hate Speech Detection via Multi-Faceted Text Representations”, WebSci 2020 - Proceedings of the 12th ACM Conference on Web Science (2020) 11-20.

[18] Marzieh Mozafari, Reza Farahbakhsh and Noël Crespi, ”Hate speech detection and racial bias mitigation in social media based on BERT model”, PLoS ONE 15(8): e0237861.

[19] Walker, Samuel. Hate speech: The history of an American controversy. U of Nebraska Press, 1994.

[20] Vedeler, Janikke Solstad, Terje Olsen, and John Eriksen. ”Hate speech harms: a social justice discussion of disabled Norwegians’ experiences.” Disability Society 34.3 (2019): 368-383.