Negative Sentiment Analysis: Cyber Bullying and Hate Speech Detection

Vedant Bhardwaj
Department of Information Technology
Xavier Institute of Engineering
Mumbai, India

Bhavin Jain
Department of Information Technology
Xavier Institute of Engineering
Mumbai, India

Baljot Singh Kohli
Department of Information Technology
Xavier Institute of Engineering
Mumbai, India

ABSTRACT

Sentiment Analysis, often known as opinion mining, is the technique of determining whether text data is positive, neutral, or neg-ative using natural language processing, computer linguistics, and related technology. Hate Speech, on the other hand, is abusive or threatening speech or writing that conveys prejudice, hatred, and/or urges violence against a person or group of people because of their race, religion, sex, or sexual orientation.With the exponential rise in the number of individuals using social media, where lots of stuff is shared everyday that is ostensibly harmless, hate speech has also increased dramatically. The necessity for identifying and detecting hate speech on social media platforms, particularly Twitter, has grown considerably, as large corporations work to establish mech-anisms to combat hate speech online. The goal of this project is to meet the demand for recognising unfavourable tweets that promote hate speech. 3

Keywords

Negative Sentiment Analysis, Cyber Bullying

1. INTRODUCTION

True social media started on May 24, 1844, when a series of electronic dots and dashes were tapped out by hand on a telegraph machine. Fast forward to 1997, when Andrew Weinreich launched ”SixDegrees,” the first true social networking site, and then came giants like MySpace, Facebook, and Twitter.

Since the introduction of social media to the world, the number of users joining such sites has exploded, and within days, millions of individuals were using Facebook, Twitter, Instagram, and other similar platforms. With this rapid growth came a huge problem: the inability to monitor what is tweeted / posted on such platforms, in this case Twitter, which resulted in a massive increase in hate speech, cyber-bullying, cyber-harassment, and the use of hate on such platforms to incite violence against individuals belonging to a specific racial or linguistic group.

Hate speech is defined differently around the world, but the most basic definition is ”communication that advocates violence, prej-udice, or discrimination against a particular group of individuals based on their race, ethnicity, sexuality, or religious affiliation.” Before even we begin, one has to understand the following three :

(1) Hate Speech: Hate speech is defined by the Cambridge Dictionary as ”public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation”. Hate speech is ”usually thought to include communications of animosity or disparagement of an individual or a group on account of a group characteristic such as race, colour, national origin, sex, disability, religion, or sexual orientation”. A legal definition of hate speech varies from country to country.

(2) Cyber Bullying: Cyber-bullying or Cyber-harassment is a form of bullying or harassment using electronic means. Cyber-bullying and Cyber-harassment are also known as online bullying. It has become increasingly common, especially among teenagers, as the digital sphere has expanded and technology has advanced.

(3) Offensive Language: Language that is unambiguous in its po-tential to be abusive, for example language that contains racial or homophobic slurs. The use of this kind of language doesn’t imply hate speech, although there is a clear correlation.

The datasets used are labeled as:
—0: Hate Speech

—1: Abusive Language

—2: Neither

The paper proposes a system which takes in a custom input from a user in the jupyter notebook, and run it to check whether the input string is hateful or not, using the models trained using the annotated data.

2. ANALYSIS OF RELATED WORK

In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets

In this particular paper, the author explains briefly about how text can be classified into different apsect based on their innate forms and also gives the idea whether the datasets is balanced or not and if not then how can we manage imbalanced datasets. The author presents paper which sets an perfect example for selecting benchmark dataset which has consistent train-test-validation split,accessible data format and has less bias data.

Characterizing and Detecting Hateful Users on Twitter

The author of this technical paper tries to classify and detect hateful Twitter users in this research, as defined by Twitter’s hateful behaviour guidelines. With a random-walk-based crawler on Twitter’s retweet graph, we generate a dataset of 100, 386 individuals, each with up to 200 tweets. This paper finds users who used words from a hate speech-related lexicon and create a sub-sample of users who are at varied distances from these users. Through crowd-sourcing, these are manually tagged as hateful or not.This research used Crowdflower, a crowd-sourcing platform, to manually annotate 4, 988 users, 544 (11 percent) of whom were deemed to be nasty. We argue that this methodology addresses two flaws in previous research: it allows the researcher to strike a balance between a generic sample and a sample prejudiced toward a set of words in a vocabulary, and it provides annotators with re-alistic context, which is sometimes required to identify hate speech.

Hate Speech Dataset from a White Supremacy Forum

In this paper, the author presents the first public dataset of hate speech annotated at the level of the sentence on Internet forum posts in English. Storm-front, the largest online group of white nationalists, is the source forum, which is renowned for pseudo-rational talks on race featuring different degrees of offensiveness. (Schafer, 2002) Storm-front is credited as being the first hate website. The generated dataset contains 10,000 statements that have been classified as hate speech or not. Several features of the generated dataset have also been investigated, such as the annotators' need for extra context in order to make a decision, or the distribution of the vocabulary used in the dataset.

Detecting Online Hate Speech Using Context Aware Models Through this paper, the author briefs us the detailed information about hate-speech detection models and explains how hatespeech is used in different forms.The author showed how important it is to use context information when detecting hate speech online. Initially this paper started by presenting a corpus of hate speech made up of entire threads of internet discussion topics. The author tried introduced two types of models, feature-based logistic regres-sion models and neutral network models, for incorporating context information into hate speech detection performance. Furthermore, it ensemble models that combine the capabilities of both types of models get the greatest results for detecting hate speech online automatically.

Application of Sentiment Analysis Using Machine Learning Techniques

Through this paper, the author explains that this paper deals more of applications of Sentiment analysis and also gives detailed anal-ysis about algorithmic approaches.The author makes sure that this paper anticipates that sentiment analysis applications will continue to expand in the future, and that sentiment analytical approaches will be standardised across diverse systems and services. Future study will concentrate on three distinct characteristics that will be used to analyse diverse data sets using a combination of logistic regression and SVM methods.

A Study on Sentiment Analysis Techniques of Twitter Data Through this paper, the author wants to present the current methods for sentiment analysis of twitter data and provide in-depth study on these methods through thorough comparisons. Different kinds of approaches are used by the author for this detailed study. At the start of the paper, the author define about what is sentiment analysis and different classification methods used in machine learning. The author further concentrates on extensive study on document level and 4 types of sentence level sentiment analysis approaches of twitter data: supervised machine learning approaches, ensemble approaches, lexicon based approaches (unsupervised methods) and hybrid approaches. Lastly, comparisons have been done of all these approaches to provide a detailed outlook on sentiment analysis techniques.

Twitter Sentimental Analysis

Through this paper, the author uses sentiment analysis as a method of analysing a human's opinions and polarity of thoughts. The data gives different types of polarity indications such as positive, negative, or unbiased values. It mainly focuses on the person's tweets and hash tags to have an idea about the situations in every aspect of the existing criteria. As per the author, the goal of this paper is to see and analyse renowned people's twitter id's or hashtags and get an idea of the thinking of the people in a situation when the respective person has tweeted on it. In this paper, the system analyses the sentiments of people using Python, Twitter API and Text Blob. Lastly, in the paper various types of visualisation techniques are implemented and used for further analysis and to also get it done more accurately.

HATECHECK: Functional Tests for Hate Speech Detection Models

Through this paper, the author detects online hate using HATE-CHECK, a set of functional tests for hate speech detection models instead of typical use of metrics like accuracy and F1 score due to their inability to detect weak points in the data. HATECHECK consists of 29 model functionalities wherein test cases are made to check the quality of the models through an extensive and structured annotation process. The functional tests were selected on the basis previous research data of hate speech and also through civil society stakeholders. Since usually models are examined using held-out test data, it becomes quite difficult to assess them properly and hence through this paper, the author shows HATECHECK's targeted insights make the understanding of the model limits better which in turn allows developments of stronger models in the future.

Are You a Racist or Am I Seeing Things?

Through this paper, the author investigates the impact of annotator knowledge of hate speech on classification models by comparison of results of classification obtained through extensive training on expert and amateur annotations. The author gives evaluation through his own data set using the Waseem and Hovy dataset (2016) on which the models are run. By this paper the author also reveals that amateur annotators are more likely to categorise items as hate speech than expert annotators, and also the systems trained on amateur annotations are most likely to lose out to the systems trained on expert annotations. Lastly in the paper, tables of different metrics are shown for accurate realisation of the purpose of the paper.

How Will Your Tweet Be Received?

Through this paper, the author predicts the dominating sentiment among tweet replies (first-order) to an English source tweet. The author uses a large dataset called RETWEET which contains tweets and responses with sentiment labels manually added. The author proposes a Deep Learning approach as a starting point for solving this problem. The author first predicted the overall polarity of the tweets, i.e., whether they are received positively, negatively, or neutrally.The author then creates automatic labels for replies, trained a network which predicts the reaction of the twitter audience. Through this method, the author shows that it makes an upper-bound baseline for the polarity of the overall first-reaction of the respective tweet.

3. EVALUATION

This project makes use of three different classification algorithms, one neural net architecture and four evaluation metrics.

3.1 Understanding The Data

For this project, two data sets were used, namely:

(1) Hate Speech and Offensive Language Dataset

(2) Twitter Sentiment Analysis Dataset

Both of the datasets are two publicly available datasets, however both of them have different labels. For the ”Hate Speech and Offensive Language” dataset, there were three labels, namely:

—0: Hate Speech

—1: Abusive Language

—2: Neither

While the other dataset, ”Twitter Sentiment Analysis” had only the ”0” label. Both datasets had different columns and needed to be merged into one common dataset.

From the ”Hate Speech and Offensive Language Dataset”, we take labels 1 and 0 and copy the labels of ’0’ on to ’1’ and then rename the label ’1’ to label ’0’ and the label ’2’ to label ’1’.

The second dataset, had only one label, ’0’, and thus it didn’t need any such changes. In the end both the datasets were concate-nated into one dataset, and data cleansing / cleaning was performed, which involved removal of stop words, symbols,

3.2 Analysis of Algorithms

Classification Algorithm

A Classification Algorithm is a supervised learning technique that uses training data to predict the category of future observations. Simply put, this technique can be used to forecast which category an observation will fall into, such as whether the answer will be Yes or No. In this case, simply put, whether the user input tweet will be hateful in nature or not. The classification algorithms utilised for this project are:

(1) Multinomial Naive-Bayes

The Multinomial Naive Bayes algorithm is a probabilistic learning approach popular in Natural Language Processing (NLP). The programme guesses the tag of a text, such as an email or a newspaper story, using the Bayes theorem. It calculates each tag’s likelihood for a given sample and outputs the tag with the highest probability.

The Multinomial Naive Bayes algorithm is a probabilistic learning approach popular in Natural Language Processing (NLP). The programme guesses the tag of a text, such as an email or a newspaper story, using the Bayes theorem. It calculates each tag’s likelihood for a given sample and outputs the tag with the highest probability.

(2) Support Vector Machines

It’s a representation of the training data as points in space divided into groups by as large a gap as possible. New examples are then mapped into the same space and classified according to which side of the gap they fall on.

It is particularly effective in high-dimensional spaces and makes decisions using only a fraction of training points, making it memory efficient. The SVM approach is not recommended for huge data sets. For SVM, we must choose an optimal kernel, which is a challenging task.

Furthermore, SVM works badly when the number of training data samples is smaller than the number of features in each data set. Because the support vector machine is not a proba-bilistic model, we are unable to explain the classification in terms of probability.

(3) Random Forest Classifier

A meta-estimator that fits a number of decision trees on different sub-samples of datasets and utilises the average to improve the model’s predicted accuracy and control over-fitting. The size of the sub-sample is always the same as the size of the original input sample, but the samples are generated with replacement. It is adaptable to both classification and regression issues, and it works well with both categorical and continuous variables, as well as missing values. It’s a type of ensemble learning that can provide more accurate predictions than most other machine learning algorithms.

For the following project, after a trial and error while processing the original 70,000 tweets dataset, the final dataset size was set at 15,000 tweets, after revaluating the model accuracy for different data set sizes. The difference was clearly evident from the results themselves, as there was a significant difference in outputs, with the 15,000 tweets delivering the best possible scores. 7500 tweets belonged to each of the two datasets.

For the following project, Recurrent Neural Networks architecture. Recurrent Neural Network(RNN)

Artificial neural networks (ANNs) and simulated neural networks (SNNs), which are at the heart of deep learning approaches, are a subset of machine learning. Their name and structure are inspired by the human brain, and they function similarly to biological neurons in terms of communication.

RNN (Recurrent Neural Network) is a sort of artificial neural network that works with sequential or time series data. These deep learning algorithms are often used for language translation, natural language processing (NLP), speech recognition, and image captioning, and they’ve been integrated into popular apps like Siri and Google Translate.

Table 1: Initial Results

Algorithm	LABEL	PRECISION	RECALL	ACCURACY	F1-SCORE
Multinomial Na¨ıve Bayes	0	0.84	0.90	0.84	0.88
	1	0.9	0.80		0.76
Support Vector Machines	0	0.80	0.85	0.85	0.90
	1	0.78	0.81		0.81
Random Forest Classifier	0	0.83	0.79	0.88	0.80
	1	0.81	0.75		0.85

They use data from previous inputs to shape the present input and outcome. While typical deep neural networks presume that inputs and outputs are independent of one another, recurrent neural networks’ output is reliant on the sequence’s prior elements.

Evaluation Metrics

(1) Precision

In a dataset, when the classes are imbalanced, accuracy is not a reliable metric for measuring our performance. Precision is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that were retrieved.

P recision = T rueP ositive=P redictedY es (1)

(2) Recall

In information retrieval, recall is the fraction of the relevant documents that are successfully retrieved.

T rueP ositivityRate = T rueP ositive=ActualY es (2)

(3) Accuracy

Accuracy is the proportion of true results among the total number of cases examined.

Accuracy = (T P + T N)=(T P + F P + T N + F N) (3)

Where

—TP is True Positive

—TN is True Negative
—FN is False Negative
—FP is False Positive

(4) F1 Score
F1 Score is a number between 0 and 1 and it is the harmonic mean of precision and recall

F 1 = 2 ((precision recall)=precision + recall) (4)

F1 score sort of maintains a balance between the precision and recall for your classifier.

3.3 Table of Results

a) Multinomial
Naive Baiyes Label ’0’
Scores:

—Precision = 0.94

—Recall = 0.92

—F1 Score = 0.93
Label ’1’ scores:

—Precision 0.90

—Recall 0.92

—F1 Score 0.91

Accuracy in both the cases, was 0.92

b) Support Vector
Machines Label ’0’ Scores:

—Precision = 0.90

—Recall = 0.96

—F1 Score = 0.93

Label ’1’ scores:

—Precision 0.90

—Recall 0.87

—F1 Score 0.91

Accuracy in both the cases, was 0.92

c) Random Forest
Classifier Label ’0’ Scores:

—Precision = 0.83

—Recall = 0.99

—F1 Score = 0.90

Label ’1’ scores:

—Precision 0.98

—Recall 0.75

—F1 Score 0.85

Accuracy in both the cases, was 0.88

TABLE 2. : FINAL RESULTS

Algorithm	LABEL	PRECISION	RECALL	ACCURACY	F1-SCORE
Multinomial Na¨ıve Bayes	0	0.94	0.92	0.92	0.93
	1	0.90	0.92		0.91
Support Vector Machines	0	0.90	0.96	0.92	0.93
	1	0.90	0.87		0.91
Random Forest Classifier	0	0.83	0.99	0.88	0.90
	1	0.98	0.75		0.85

Sample Test Cases

Sample test cases are tested with the algorithm to check their respective nature on the context of the output produced which is a prediction score i.e. if its more than 0.5 then its of a hateful nature and and if its less then 0.5 then otherwise.

(1) Test Case 1: Input of a Negative Tweet

A sample tweet of the text - ”I hate you and want you dead you filth” is tested with the algorithm. The statement is of a negative nature and is then tested with the algorithm to check the produced classified output. Upon testing a prediction score of 0.6719118 is produced which conveys that this tweet if of a hateful and negative nature.

(2) Test Case 2: Input of a Positive Tweet
Another sample tweet of the text - ”I like middle-eastern food and appreciate the culture as well” is tested with the algo-rithm. The statement is of a fairly positive nature and is then tested

with the algorithm to check the produced classified out-put. Upon testing a prediction score of 0.1969788 is produced which conveys that this tweet if of a positive and fairly non-hateful nature.

4. CONCLUSION

Overall, out of all the social media platforms, Twitter is the most widely used site for discussing issues in all possible contexts. It has evolved into a hub for all digital beings to come together and participate in debates, thereby presenting their opinions and ideals on a certain issue. As a result, Twitter is essentially a big collection of data that may be used for sentiment analysis, which serves as the inspiration for this research. It estimates whether a tweet has a negative or positive sentiment for research and ease of usage. The programme now only tests tweets in English, but future advances and study may allow it to be expanded to other lan-guages.Furthermore, the label imbalance between hate and no hate is lopsided, necessitating the development of a complex manual labelling system that takes learning into account, so benefiting all labelling efforts. With additional advancements, sarcasm handling can also be improved. This document will serve as a baseline for determining the sentiment and true essence of a tweet, which will undoubtedly aid the community by enhancing their thinking on the subject or issue at hand. Thus, data connected to many themes on Twitter, such as white supremacy, Hinduphobia, racism, blasphemy, hate speech, and so on, may be found and used to determine if a tweet is hostile and negative or not utilising this data.

5. REFERENCES

[1] Paul Rottger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem,Helen Margetts, Janet B. Pierrehumbert, University of Oxford, The Alan Turing Institute, Utrecht University, Uni-versity of Sheffield: HATECHECK: Functional Tests for Hate Speech Detection Models, May 2021.

[2] Soroosh Tayebi Arasteh, Mehrpad Monajem, Vincent Christlein, Philipp Heinrich, Anguelos Nicolaou, Hamidreza Naderi Boldaji, Mahshad Lotfinia and Stefan Evert, Friedrich-Alexander-Universitat Erlangen-Nurnberg, Germany, Harvard Medical School, United States, Sharif University of Tech-nology, Iran: How will your Tweets be Recieved? April 2021.

[3] Abdullah Alsaeedi, Mohammad Zubair Khan, A Study on Sentiment Analysis Techniques of Twitter Data - International Journal of Advanced Computer Science and Applica-tions (IJACSA) Volume No. 2, November 2019.

[4] Ona de Gibert, Naiara Perez, Aitor Garc´ıa-Pablos, Montse Cuadros, Hate Speech Dataset from a White Supremacy Fo-rum HSLT Group at Vicomtech, Donos- tia/San Sebasti´an, Spain, September 2018.

[5] Pedro H. Calais, Yuri A. Santos, Virg´ılio A. F. Almeida, Wagner Meira Jr., Universidade Federal de Minas Gerais Belo Horizonte, Minas Gerais, Brazil, Characterizing and Detecting Hateful Users on Twitter - Manoel Horta Ribeiro, March 2018.

[6] Lei Gao, Ruihong Huang, Texas AM Univer- sity, Detecting Online Hate Speech Using Context Aware Models, May 2018.

[7] Kosisochukwu Judith Madukwe, Xiaoying Gao, Bing Xue, School of Engineering and Computer Science, Victoria Uni-versity of Wellington, In Data We Trust: A critical analysis of hatespeech detection datasets, November 2020.

[8] Zeerak Waseem, University of Copenhagen, Copenhagen, Denmark, Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter, 2016.

[9] Shobana G, Vigneshwara B, Maniraj Sai A, Twitter Sentiment Analysis, International Journal of Recent Technology and En-gineering (IJRTE), November, 2018.

[10] Anvar Shathik J and Krishna Prasad K,International Journal of Applied Engineering and Management Letters (IJAEML), A Literature Review on Application of Sentiment Analysis Using Machine Learning Techniques, 2020.