TWEET CENSOR: Automated Detection & Analysis of Hate Speech on Twitter

MAJOR PROJECT/DISSERTATION REPORT

Submitted by

FAHAD RAFIQUE
2019-310-042

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY
(COMPUTER SCIENCE AND ENGINEERING)

Under the supervision of

Ms. SHABINA GHAFIR

Department of Computer Science & Engineering
School of Engineering Sciences & Technology

JAMIA HAMDARD
(Deemed to be University)
New Delhi-110062

2023

DECLARATION

I, **Mr Fahad Rafique** a student of **Bachelor of Technology in Computer Science & Engineering (B.Tech CSE), Enrolment No: 2019-310-042** hereby declare that the Major Project/Dissertation entitled **"TWEET CENSOR: Automated Detection & Analysis of Hate Speech on Twitter "**which is being submitted by me to the Department of Computer Science and Engineering, School of Engineering Sciences and Technology, Jamia Hamdard, New Delhi in partial fulfilment of the requirement for the award of thedegree of **Bachelors of Technology (Computer Science & Engineering)** is my original work and has not been submitted anywhere else for the awardof any Degree, Diploma, Associateship, Fellowship or other similar title or recognition.

Fahad Rafique
2019-310-042

Date: May 2023
Place: New Delhi

ACKNOWLEDGEMENT

I express my sincere thanks to Ms Shabina Ghafir (Assistant Professor, Dept. Computer Science and Engineering), my project in charge, who guided me through the project with her valuable suggestions and guidance for completing the project. This project has been a success only because of her guidance.

I deeply express my sincere thanks to our Head of Department Dr. Farheen Siddiqui for encouraging me and allowing me to present the project on the topic **"TWEET CENSOR: Automated Detection & Analysis of Hate Speech on Twitter"** at the department premises for the partial fulfilment of the requirements leading to the award of B. Tech degree.

I am also thankful to the whole computer science and engineering department for providing the technical support to carry out the project work, letting us utilize all the necessary facilities of the institute and providing guidance at each& every step during the project work.

INDEX

INDEX OF FIGURES/ DIAGRAM

TWEET CENSOR:
AUTOMATED DETECTION
&
ANALYSIS
OF HATE SPEECH
ON
TWITTER

OBJECTIVE

This project aims to combine multiple services and open-source tools to make an automated content moderation system that will identify and detect whether the speech is recognized as “Hate Speech” or “Non – Hate Speech” using machine learning algorithms and with the help of sentimental analysis it will further detect the compound score of the speech using VADER.

INTRODUCTION

In recent years, hate speech on social media, particularly on Twitter, has become a major issue. The platform, which allows users to post short messages known as tweets, has been used to spread hateful and harmful messages based on race, religion, gender, sexual orientation, and other characteristics. As a result, there is growing concern about the impact of hate speech on social media on both individuals and society as a whole.

Hate speech on social media can have a number of negative impacts on individuals. For those who are targeted by hate speech, it can be extremely distressing and damaging to their mental health. It can also lead to feelings of isolation and fear, as well as physical harm in extreme cases. For those who witness hate speech, it can lead to a sense of powerlessness and frustration, as well as a feeling of being unsafe in online spaces. Hate speech on social media can also have broader societal impacts.

It can contribute to the normalization of hateful attitudes and beliefs, as well as the spread of misinformation and propaganda. It can also contribute to the polarization of society, making it more difficult to engage in meaningful dialogue and compromise.

One of the challenges of addressing hate speech on social media is the tension between free speech and the need to protect individuals from harm. While free speech is an important value, it is not absolute, and there are limits to what can be said without causing harm. Hate speech, by definition, is speech that seeks to denigrate or harm individuals or groups based on their characteristics. As such, it is not protected by free speech laws.

In recent years, the rise of hate speech on Twitter and other social media has become a major problem. Although it is difficult to measure the exact extent of the increase in hate speech on social media, there is evidence that hate speech has become more visible in recent years.

Anti-Semitic tweets on Twitter increased by 30% between 2016 and 2018, according to the Anti-Defamation League report. Similarly, the number of

white supremacist tweets on the platform increased by 30% over the same period. The report also found a significant increase in the use of Twitter by extremist groups and individuals, with many using the platform to spread hatred and incite violence.

A Pew Research Centre study found that 41% of Americans have experienced online harassment, and many cite hate speech as one of the biggest forms of harassment. The study also found that social media platforms such as Twitter are the most common venue for online harassment.

Similarly, in Europe, the rise of far-right movements and the spread of hateful speeches using social media has become a major problem. A study by the Institute for Strategic Dialogue found that the use of anti-Semitic language on social media platforms has increased significantly in Europe, with Twitter being one of the most commonly used platforms for this purpose. It's one. Overall, statistics suggest that the rise of hate speech on social media, including Twitter, is a significant problem worldwide. Social media platforms have taken some steps to address this issue, but more needs to be done to prevent these platforms from being used as vehicles for hate and intolerance.

CONTENT MODERATION: -

Content moderation refers to the screening of inappropriate content that users post on a platform.

The objective of content balance is to make a secure and aware environment for all clients, whereas moreover protecting freedom of expression and advancing sound wrangle about and talk. This could include expelling substance that abuses the platform's rules, such as despise discourse, badgering, or realistic viciousness, as well as taking activity against clients who abuse these rules, such as by suspending or forbidding their accounts. Content Moderation can be performed by both people and algorithms, depending on the stage and the sort of substance being directed. Human arbitrators are frequently mindful for looking into and evacuating more complex or touchy substances, such as abhor discourse or deception, whereas calculations can be utilized to identify and expel spam or low-level offences, such as copy posts or foulness.

The goal of content moderation is to ensure the platform is safe to use and upholds the brand’s Trust and Safety program.

Content Moderation is an imperative viewpoint of online community administration, and stages contribute critical assets in creating arrangements, apparatuses, and workflows to guarantee that it is performed successfully and effectively. Be that as it may, substance control is additionally a complex and always advancing field, and stages must adjust the requirements for security and free expression with the challenges of scale, setting, and client protection.

WHY CONTENT MODERATION IS ESSENTIAL?

Some arguments claim that any form of content moderation violates the fundamentals of free speech. But it's a very black-and-white discussion in a very grey area. Simply put, content moderation is essential to keep the platform safe from offensive, threatening, violent, illegal and pornographic posts. Illegal content such as child pornography and videos depicting unjustified violence are routinely uploaded, and this is clearly an area where content moderation is essential.

On the other side of the scale, there is a ton of visual content that is illegal for copyright reasons. With the sheer amount of content uploaded every day, content moderation is essential. Leaving this "free for everyone" results in social media feeds filled with horrific content, personal bullying, and the internet flooded with copyright-infringing videos and images.

CONTENT MODERATION CASE EXAMPLE – FACEBOOK: -

Looking at how a platform like Facebook deals with the issue of content moderation gives a good idea of the problems that all "mega-platforms" face when trying to keep content within their guidelines. According to NYU Stern report, Facebook's "human moderators" review more than 3 million photos, videos and messages every day. This may not seem like a bad thing, but keep in mind that many of these posts contain videos that could be longer. This simply means that moderators often have a few seconds to rate other posts. When you work under such pressure, you often make mistakes. Facebook CEO Mark Zuckerberg admitted in a recent white paper that moderators made the wrong decision about 10% of the time. This means that 300,000 messages a day are incorrectly tracked.

WHAT EXACTLY CONSTITUTES UNSUITABLE CONTENT?

The moderation of content can be conducted manually by human moderators or through automated techniques, such as machine learning algorithms. Manual content moderation entails the individual inspection and appraisal of each uploaded content piece to make a decision on whether it adheres to the guidelines or should be removed. Automated content moderation entails the use of algorithms to scan and flag content that infringes platform policies.

Content moderation is a critical aspect of ensuring a secure and wholesome online environment for users. It assists in preventing the dissemination of dangerous or inappropriate content and ensures that the platform is utilized for its intended purpose.

PROBLEM WITH THE HUMAN CONTENT MODERATION: -

- **The Human cost of moderation:** Moderators have been put under a lot of pressure to make ends meet, which has caused them to suffer from mental health issues like burnout and PTSD. It's bad enough that they have to work so hard, but now there are even more reports of them suffering from PTSD. It's easier to understand this when you look at the kind of content they have to deal with every day. According to a report from NYU Stern, in the first three months of 2020, there were a lot of posts that were removed for various reasons.

- Every Major Tech company uses third-party contractors for content moderation. This increases the overall cost with not efficient in removing unsuitable content.

DATA UNDERSTANDING: -

❖ **What are the linguistic variations between hate speech and offensive language?**

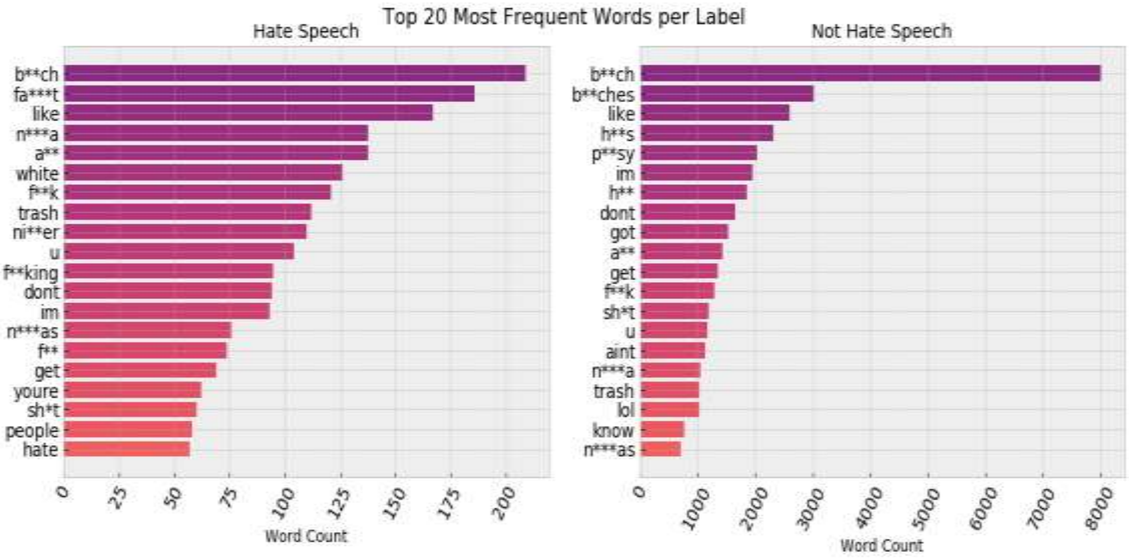


Fig (1)

Linguistically, the distinction between hate speech and offensive language frequently comes down to how it targets oppressed communities, often in threatening ways.

Despite the fact that the labels contain extremely similar commonly recurring words, just 20% of the "hate speech" label is unique in total.

For example, hate speech frequently contains the n-word with a hard 'r.'

The use of this slur may suggest malicious intent, which goes beyond the possibility of the word being used as cultural slang.

Such examples highlight the complexities of English slang as well as the narrow boundary between hate speech and inappropriate language. because the vocabulary of each label is so similar; machine learning algorithms may struggle to distinguish between them and establish what constitutes hate speech.

- **What are the popular hashtags in the dataset of each tweet type?**

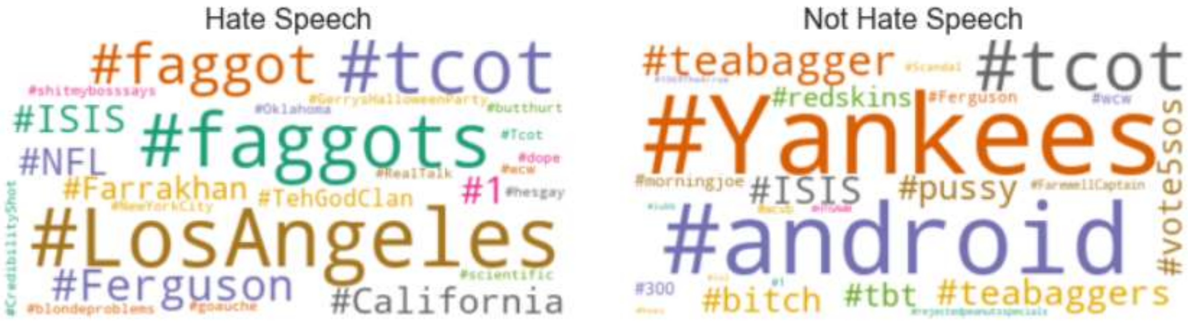


Fig (2)

Some most popular hashtags used in the tweets by the users.

- **The overall polarity of the tweets: -**

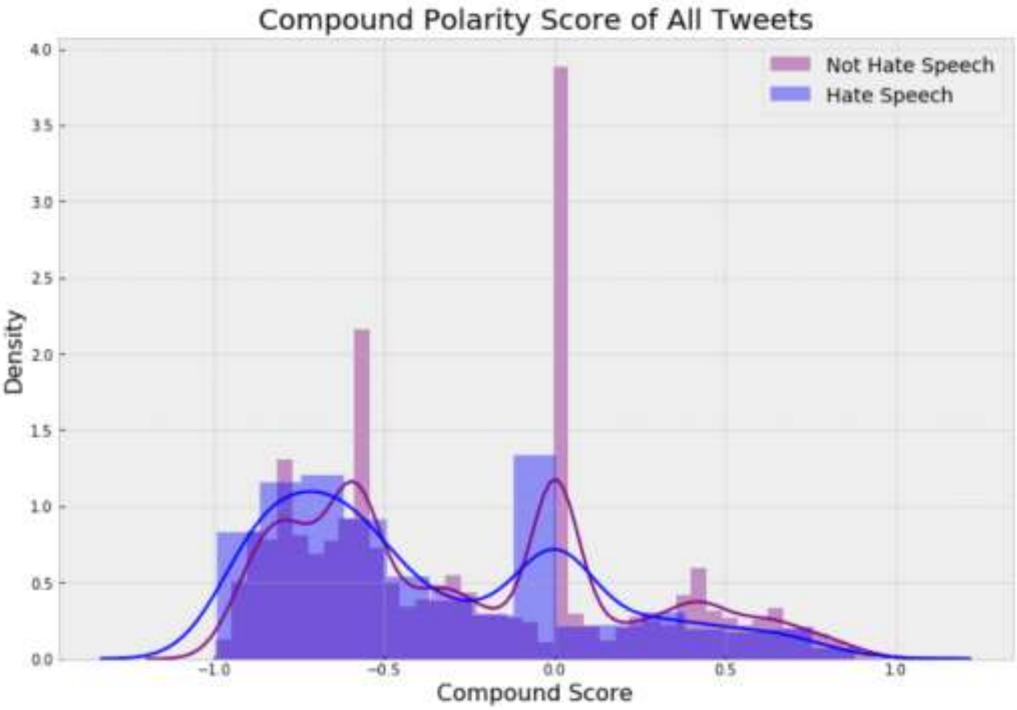


Fig (3)

- **CLASS IMBALANCE: -**

  The main obstacle of this dataset is the extreme class imbalance which basically affects the whole model. The dataset contains 5.77% which is labelled as hate speech. This could present challenges during the modelling process.

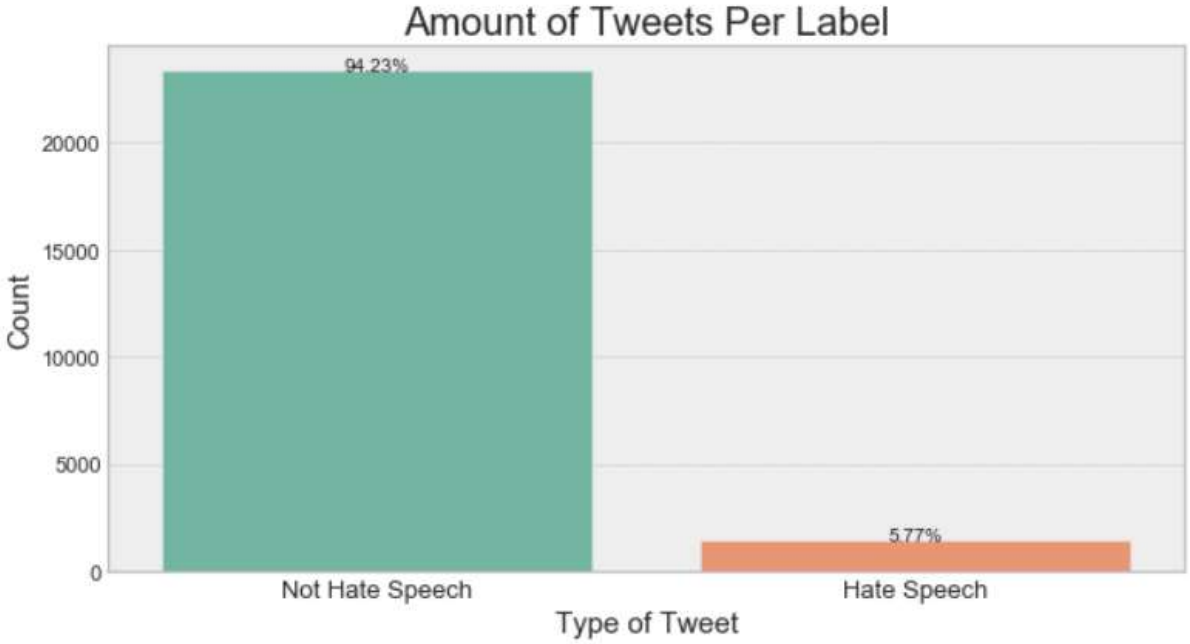


Fig (4)

- **Unique words comparison in each corpus label: -**

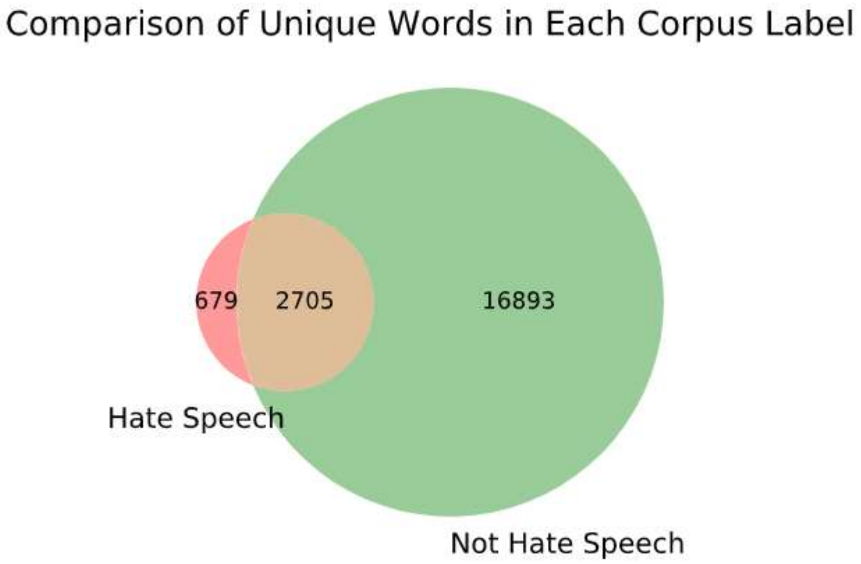


Fig (5)

STEPS INVOLVED IN BUILDING THE PROJECT

1) DATA SOURCING
2) DATA CLEANING
3) DATA PREPROCESSING
4) EXPLORATORY DATA ANALYSIS
5) FEATURE ENGINEERING
6) MODEL PROCESS
7) WORKING ON BASELINE MODELS
8) MODELS EVALUATION ON THE BASIS OF THEIR ACCURACY
9) FINALIZATION OF THE MODEL
10) WEBSITE

1. **Data Sourcing:** The dataset for this project was sourced from a study named “Automated Hate Speech Detection and the Problem of Offensive Language” which was conducted by Thomas Davidson and a team at Cornell University in the year of 2017.

The dataset is a .csv file with 24,802 text posts from Twitter where almost 6% of the tweets are labelled as “Hate Speech”.

2. **Data Cleaning:** This is the first step that involves the cleaning of the sourced dataset. The aim of information cleaning right here is to find the proper manner to rectify high-satisfactory issues like eliminating bad data, and filling in missing data to form the information efficient for the model.

3. **TEXT PREPROCESSING:** This step involves the following transformations.

- Lowercase
- Tokenization
- Removing special characters
- Removing stop words and punctuation

4. **EXPLORATORY DATA ANALYSIS:** EDA is performed on the cleaned dataset to figure out the features and uniqueness of the data.

5. **FEATURE ENGINEERING:** The goal of feature engineering in Natural Language Processing is to convert tokenized text data into numerical vectors that the machine learning algorithm can "understand." We'll be iterating the dataset with the three different feature engineering techniques in our notebook: **count vectorization, Doc2Vec and TF-IDF vectorization.** Trying these techniques on the same four baseline models could result in vastly different metrics.

6. **MODEL PROCESS**:

   Metrics for Evaluation: We will use the F1 score as the primary metric for this business problem, while also considering Precision and Recall. The F1 score calculates the harmonic mean of Precision and Recall, which is useful for data with a high-class imbalance.

   Overall, we want as much hate speech as possible to be flagged and efficiently removed. This entails optimising the True Positive Rate, also known as Recall.

   - **F1 Weighted Score:** We will also consider the weighted F1 score, which can account for label class imbalance by calculating metrics for each label.

- **Cross Validation:** In order to check out whether a model is overfitting or underfitting, we can use K-Fold Cross Validation to generate an F1 Score for the training set.

7. **WORKING WITH BASELINE MODELS**: Used multiple baseline machine learning models (Random Forest, Logistic Regression, Naïve Bayes, Support Vector Machine.
Out of the above models, the best model with the most accuracy and best performance will be selected to iterate the further procedure of the model.

Random Forest: Random Forest: A random forest is a classification algorithm which consists of a variety of decision trees on different subsets of a given data set and takes the mean to improve the prediction performance of that data set. The greater the number of trees, the greater the accuracy and equipment risk. the lower the surplus.

- Random Forest is able to execute both classification and regression tasks.
- It improves model accuracy and avoids the problem of overfitting.

Support Vector Machine: SVM is primarily utilised in classification tasks. Each data item is plotted as a point in an n-dimensional space by the algorithm (where n is the number of features it possesses). where the value of each feature is the value of a single coordinate. The classification is then carried out by locating hyperplanes that clearly distinguish the two classes. The support vector machine is the best boundary between the two classes (hyperplane and line). SVM can employ kernels if your data demands nonlinear categorization. This is a function that turns a low-dimensional input space into a high-dimensional space. H. They separate problems that were previously intertwined.

Naïve Bayes: In classification problems, the goal is to assign a label (or class) to an input data point based on its features. The naive Bayes algorithm assumes that the features are independent of each other, which simplifies the computation of the probability of the data given the class (or likelihood) and the probability of the class (or prior).

The naive Bayes algorithm calculates the probability of each class for a given data point using the Bayes' theorem and then selects the class with the highest probability as the predicted class. The probability of each class is computed as the product of the likelihood of the features given the class and the prior probability of the class, divided by the probability of the data.

The algorithm is called "naive" because it assumes that all features are independent and equally important, which is often not true in practice. However, despite this simplifying assumption, naive Bayes can perform surprisingly well in many real-world applications, especially when the number of features is large.

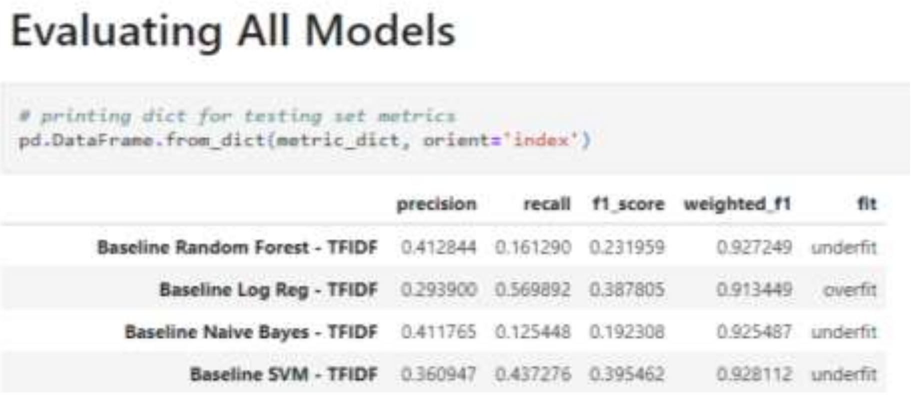
Overall, Naive Bayes is a popular and effective algorithm for classification tasks in machine learning, especially for text classification and spam filtering.

Logistics Regression: Logistic regression is a popular statistical method in machine learning for binary classification applications. Based on a collection of input features, the purpose of binary classification is to predict a binary output, such as "yes" or "no." Logistic regression employs a logistic function to represent the relationship between the input features and the output variable, yielding a probability estimate for each class.

Logistic regression is a simple yet powerful technique for binary classification problems that is widely utilised in a wide range of applications, including customer churn prediction, fraud detection, and medical diagnosis. It is interpretable, which means that the learned

parameters can be used to comprehend the link between the input features and the output variable

8. MODEL EVALUATION ON THE BASIS OF THEIR ACCURACY:



Evaluating All Models

```
# printing dict for testing set metrics
pd.DataFrame.from_dict(metric_dict, orient='index')
```

	precision	recall	f1_score	weighted_f1	fit
Baseline Random Forest - TFIDF	0.412844	0.161290	0.231959	0.927249	underfit
Baseline Log Reg - TFIDF	0.293900	0.569892	0.387805	0.913449	overfit
Baseline Naive Bayes - TFIDF	0.411765	0.125448	0.192308	0.925487	underfit
Baseline SVM - TFIDF	0.360947	0.437276	0.395462	0.928112	underfit

Fig (6)

9. FINALIZATION OF THE MODEL: The best performing model with the best accuracy among all the algorithms is logistics regression model using Count Vectorizer for feature engineering. The hyperparameters were penalty = 12 and class weight = 'balanced'.

10. DEPLOYMENT OF THE MODEL OF THE WEB PAGE:

After successfully performing the steps involves from data cleaning to finalization of the model.

The model is deployed on the web page with the help of Streamlit framework.

PROBLEM STATEMENT

The use of human content moderation in social media platforms has been a subject of scrutiny due to reports of the exploitation of workers. In 2019, an article published by The Verge uncovered the extensive list of inhumane working conditions that employees faced at Cognizant, which was Facebook's former moderation contractor. The article shed light on how content moderators were subjected to traumatic and distressing content without adequate support or counselling and were paid low wages despite the highly demanding and psychologically taxing nature of their work. However, it is important to note that the use of human moderators is still prevalent among major tech companies, including Twitter, both domestically and overseas.

Hate speech, which is defined as abusive or threatening language that expresses prejudice against a particular group, based on factors such as race, religion, or sexual orientation, is a major concern for social media platforms. However, detecting and addressing hate speech is a complex challenge for social media platforms, as it requires a nuanced understanding of language and context, as well as ethical and legal considerations. It is worth noting that datasets used to train hate speech recognition systems can sometimes contain offensive language, which may be triggering or harmful to some individuals.

SOFTWARE REQUIREMENT SPECIFICATIONS

1. PyCharm
2. Jupyter Notebook
3. Streamlit (Frontend)
4. Some Preinstalled & External libraries (Pickle, Pandas, Scikits Learn etc)

PyCharm: -

PyCharm is designed to increase productivity and efficiency by offering a variety of features and functionalities, such as code completion, syntax highlighting, and project management tools.

PyCharm simplifies the process of developing Python applications by providing a user-friendly interface that facilitates coding and debugging. It supports a range of Python frameworks and libraries and includes built-in tools for testing and debugging code. PyCharm also offers integrations with other development tools, such as version control systems and package managers, to streamline the development process.

Jupyter Notebook: -

Jupyter notebooks are particularly excellent for "presenting the work" done by your data team by combining code, markdown, links, and photos. They're simple to use and may be run cell by cell to see how the code works.

Streamlit (Frontend): -

Streamlit is a fantastic new tool that allows engineers to quickly create highly dynamic online applications based on their data, machine learning models, or anything else.

NumPy:-

It is used for working with arrays and linear algebra.

Pandas: -

It is used for data analysis and data pre-processing, CSV file I/O
(e.g. *pd.read_csv)*

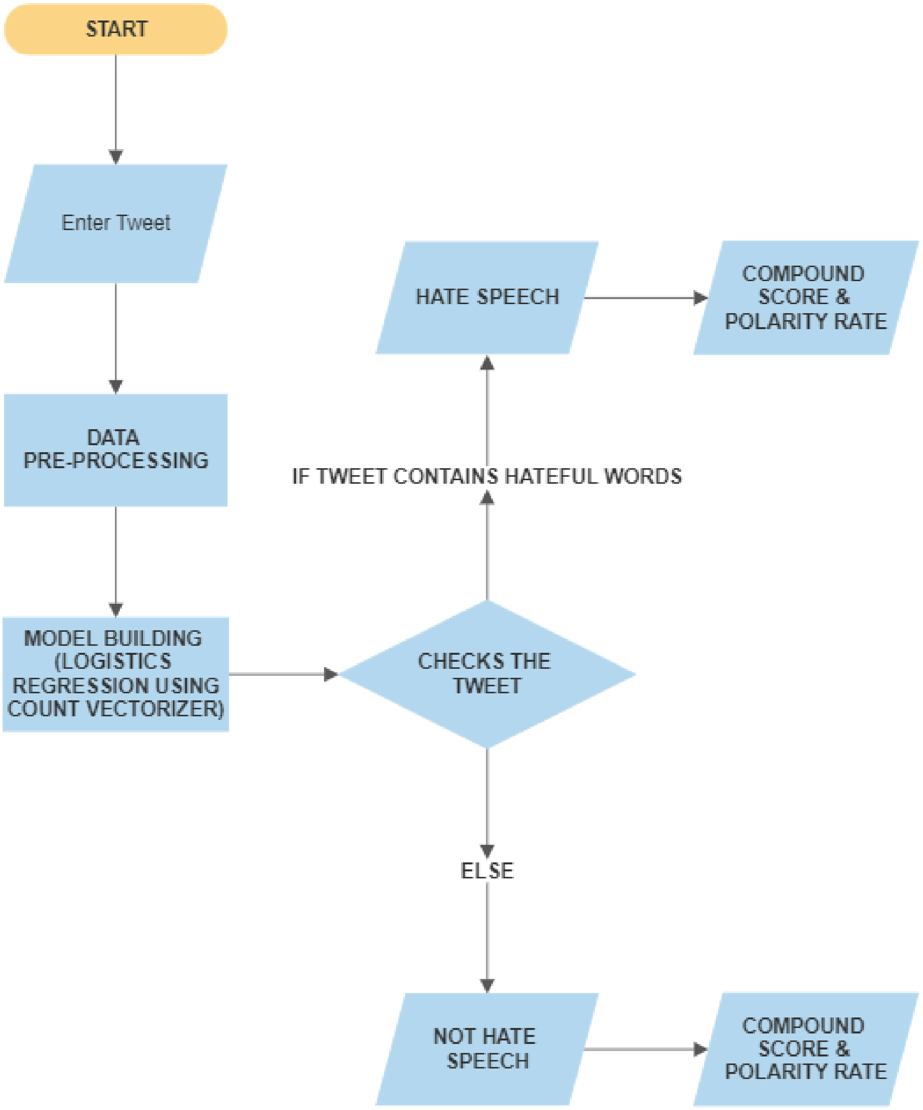
Sklearn: -

It is used for making use of Machine learning tools. In this project, we have used cosine similarity for recommendations.

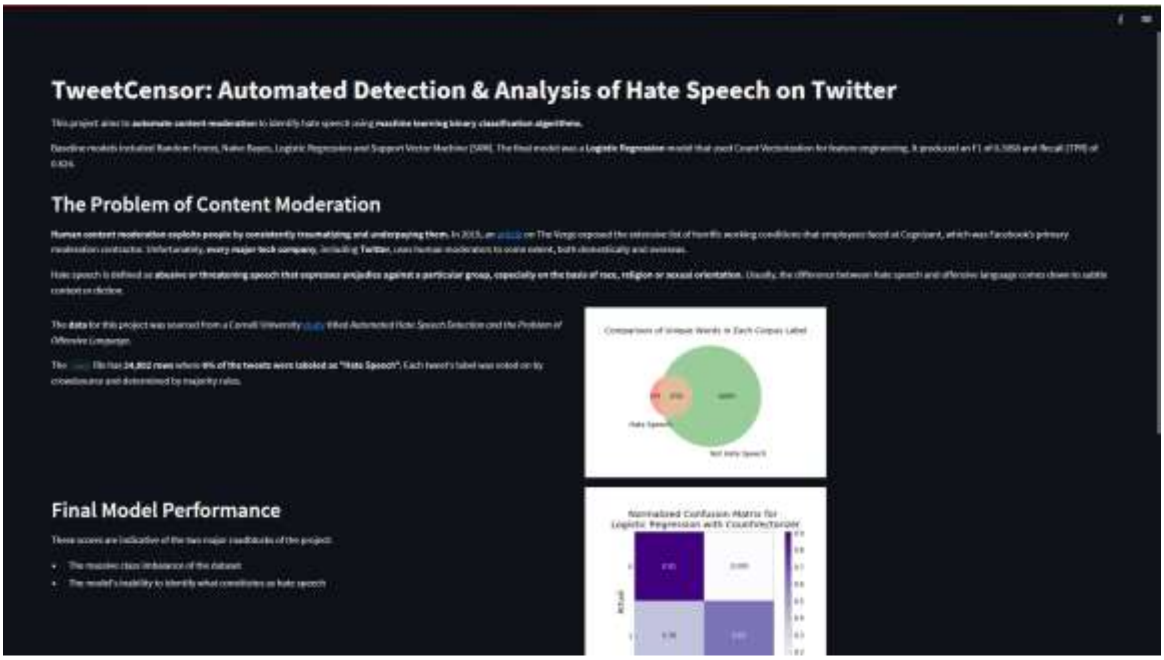
NLTK: -

NLTK is a toolkit built for working Python with NLP. Various word processing libraries are provided with many test datasets. NLTK allows you to perform various tasks such as B. Tokenization, parse tree visualization, etc.

ENTITY RELATION DIAGRAM


START
Enter Tweet
DATA PRE-PROCESSING
MODEL BUILDING (LOGISTICS REGRESSION USING COUNT VECTORIZER)
CHECKS THE TWEET
IF TWEET CONTAINS HATEFUL WORDS
HATE SPEECH
COMPOUND SCORE & POLARITY RATE
ELSE
NOT HATE SPEECH
COMPOUND SCORE & POLARITY RATE

SNAPSHOTS OF THE DIFFERENT INPUT AND OUTPUT SCREENS



TweetCensor: Automated Detection & Analysis of Hate Speech on Twitter

This project aims to **automate content moderation** to identify hate speech using **machine learning binary classification algorithms.**

Baseline models included Random Forest, Naive Bayes, Logistic Regression and Support Vector Machine (SVM). The final model was a **Logistic Regression** model that used Count Vectorization for feature engineering. It produced an F1 of 0.3958 and Recall (TPR) of 0.624.

The Problem of Content Moderation

Human content moderation exploits people by consistently traumatizing and underpaying them. In 2019, an article on The Verge exposed the extensive list of horrific working conditions that employees faced at Cognizant, which was Facebook's primary moderation contractor. Unfortunately, **every major tech company**, including **Twitter**, uses human moderators to some extent, both domestically and overseas.

Hate speech is defined as **abusive or threatening speech that expresses prejudice against a particular group, especially on the basis of race, religion or sexual orientation.** Usually, the difference between hate speech and offensive language comes down to subtle context or diction.

The **data** for this project was sourced from a Cornell University study titled *Automated Hate Speech Detection and the Problem of Offensive Language*.

The [illegible] file has **24,802 rows** where **6% of the tweets were labeled as "Hate Speech"**. Each tweet's label was voted on by crowdsource and determined by majority rules.

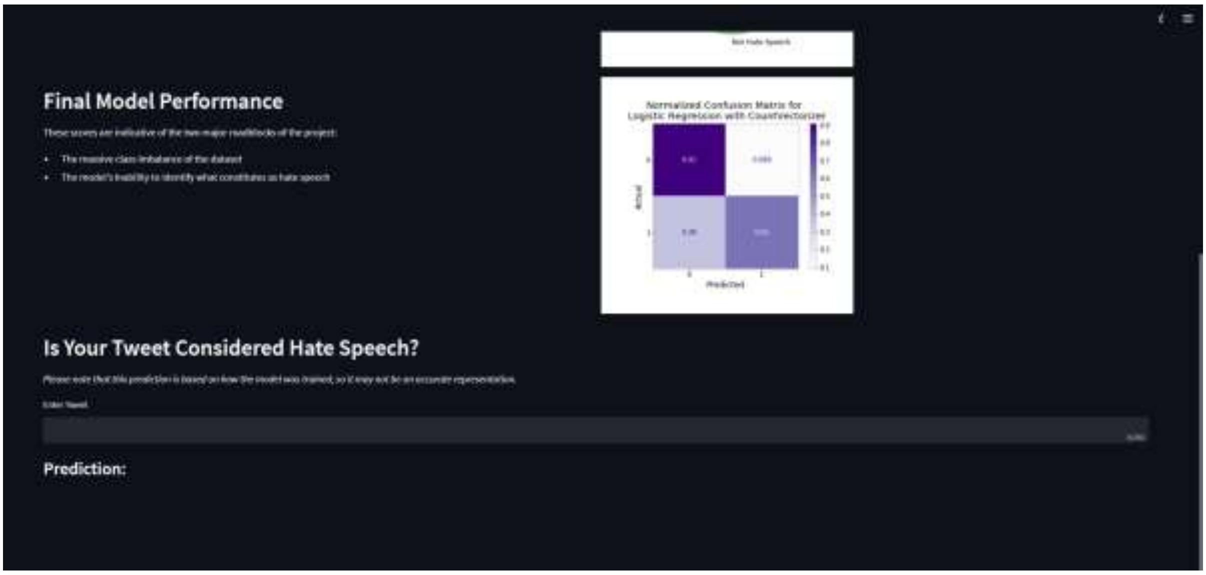
Final Model Performance

These scores are indicative of the two major roadblocks of the project:

- The massive class imbalance of the dataset
- The model's inability to identify what constitutes as hate speech

Normalized Confusion Matrix for Logistic Regression with CountVectorizer
Actual
0
1

Fig (8) Interface of the web application



Final Model Performance

These scores are indicative of the two major roadblocks of the project:

- The massive class imbalance of the dataset
- The model's inability to identify what constitutes as hate speech

Normalized Confusion Matrix for Logistic Regression with CountVectorizer
Actual
Predicted
0
1

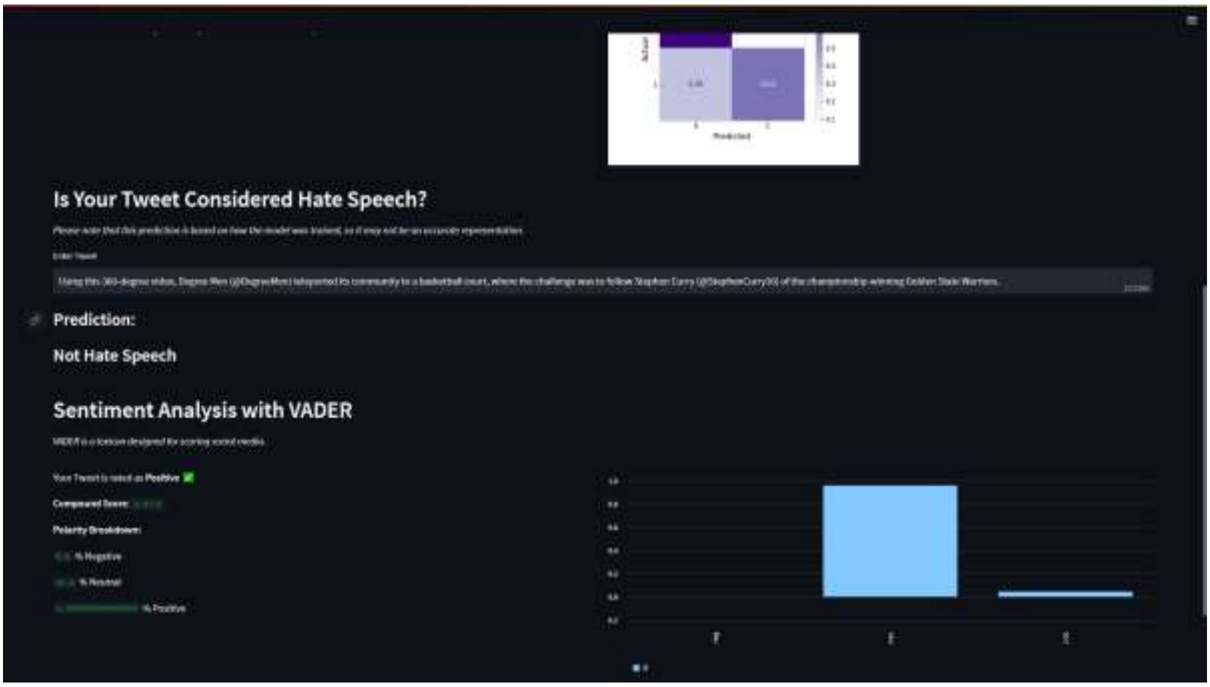
Is Your Tweet Considered Hate Speech?

Please note that this prediction is based on how the model was trained, so it may not be an accurate representation.

Enter Tweet

Prediction:

Fig (9) Interface of the web application



Is Your Tweet Considered Hate Speech?

Please note that this prediction is based on how the model was trained, so it may not be an accurate representation.

Enter Tweet

Using this 360-degree video, Degree Men (@DegreeMen) teleported its community to a basketball court, where the challenge was to follow Stephen Curry (@StephenCurry30) of the championship-winning Golden State Warriors.

Prediction:

Not Hate Speech

Sentiment Analysis with VADER

VADER is a lexicon designed for scoring social media.

Your Tweet is rated as **Positive** ✅

Compound Score: [illegible]

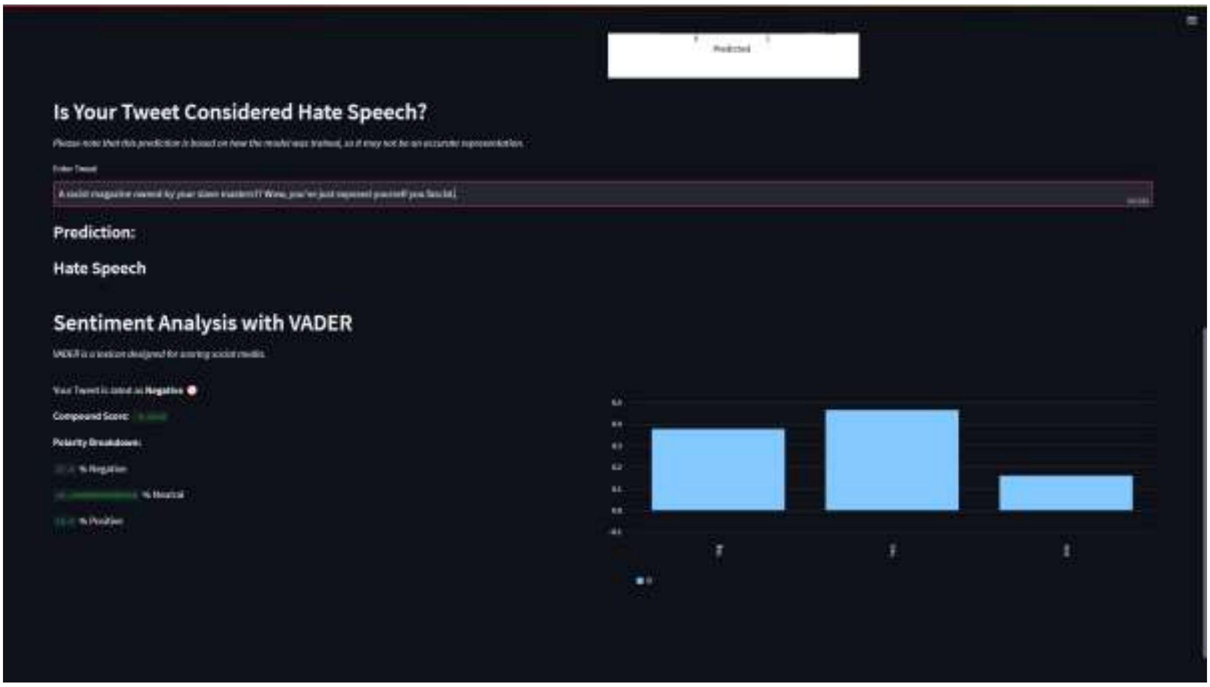
Polarity Breakdown:

[illegible] % Negative

[illegible] % Neutral

[illegible] % Positive

Fig (10) User entered a tweet and the model predicted it as “Not Hate Speech” Content along with the Compound Score & Polarity Breakdown



Is Your Tweet Considered Hate Speech?

Please note that this prediction is based on how the model was trained, so it may not be an accurate representation.

Enter Tweet

A racist magazine owned by your slave masters?? Wow, you've just exposed yourself you fascist.

Prediction:

Hate Speech

Sentiment Analysis with VADER

VADER is a lexicon designed for scoring social media.

Your Tweet is rated as **Negative** ●

Compound Score: [illegible]

Polarity Breakdown:

[illegible] % Negative

[illegible] % Neutral

[illegible] % Positive

Fig (11) User entered a tweet and the model predicted it as “Hate Speech” Content along with the Compound Score & Polarity Breakdown.

CONCLUSION

It is critical to comprehend why the model performed poorly and how this connects to the business issue. The "Hate Speech" label forecasts lowered the F1 score. The algorithm successfully predicted 91% of the "Not Hate Speech" classifications, but only 62% of the other labels.

Its performance reflects the project's two key roadblocks:

- The dataset's massive class imbalance Model's incapacity to "understand" the intricacies of hate speech

The problem of class imbalance can be solved using pre-processing and oversampling/under sampling approaches. Yet, recognising hate speech is a broader issue that many large technology companies, such as Twitter, Facebook, and Instagram, are still grappling with.

This is due to the fine line that exists between Hate Speech and plain inflammatory words. Hate speech is described as abusive or threatening speech that conveys prejudice against a certain group, particularly based on race, religion, or sexual orientation. The distinction between hate speech and offensive language is usually determined by small differences in context or diction.

We can observe that the "Hate Speech" label has 679 terms that are unique to it. Some of these terms are illogical or meaningless, but others are particularly offensive. This label, for example, comprises phrases like "sp-cs", "mo-kies", "ov-nj-w", "fa—ot" and many more. All of these are racial or homophobic slurs. More importantly, we saw in earlier EDA that this label contains the N-word with the hard "R" excessively. By specifically targeting underprivileged communities, language like this obviously reveals hate speech.

Unfortunately, it is difficult for a machine learning model to recognise the intricacies of this derogatory language. **Despite the fact that these words are unique to the "Hate Speech" label corpus, there is simply insufficient data for that label.**

As a result, a critical next step would be to collect more data that CrowFlower voters have recognised as hate speech.

Furthermore, for the time being, we can recommend that Twitter manually regulate tweets containing terms, much as it could with the top hashtags discovered for each category in previous EDA.

Furthermore, due of the intricacies in English slang and slurs, automated hate speech identification is an incredibly challenging task. This effort was able to start that process, but much more work needs to be done to keep this stuff off of public-facing venues like Twitter.

LIMITATIONS

There are several challenges that are faced in hate speech recognition systems for social media platforms such as Twitter. Some of the key challenges include:

- Contextual understanding: Hate speech can be highly contextual, and what constitutes hate speech may vary based on factors such as the target group, the historical and cultural context, and the intention of the speaker. Therefore, developing algorithms that can accurately detect hate speech in different contexts can be a challenge.

- Language diversity: Social media platforms such as Twitter have a diverse user base, and people use different languages and dialects to communicate. Recognizing hate speech in multiple languages requires a large amount of training data and the ability to handle variations in language structure and expression.

- Sarcasm and irony: Hate speech is not always direct and can be expressed through sarcasm, irony, and other forms of indirect language. Recognizing such language requires understanding of the underlying context and the ability to identify patterns and connections between different phrases.

- New trends and terminology: The language and terminology used to express hate speech can evolve quickly and vary across different communities. Developing algorithms that can recognize new trends and terminology requires continuous monitoring and updates to the training data.

- Legal and ethical considerations: Hate speech detection systems need to balance the need to protect individuals from harmful speech with the need to preserve freedom of expression. Developing algorithms that can accurately identify hate speech while avoiding false positives and respecting the diversity of opinions can be a challenging task.

- Overall, developing effective hate speech recognition systems for social media platforms such as Twitter requires overcoming these and other challenges, and requires continuous refinement and improvement as new trends emerge and language evolves.

FUTURE SCOPE

The future scope of hate speech recognition systems for social media platforms such as Twitter is vast and promising. Here are some potential areas of growth and improvement:

- Multilingual support: Hate speech recognition systems can expand their language support beyond English to include a wider range of languages, as social media users from all over the world communicate in different languages.
- Fine-grained classification: Hate speech recognition systems can aim for more nuanced classification of hate speech, such as identifying specific types of hate speech, or distinguishing between hateful speech and speech that is simply controversial or offensive.
- Contextual understanding: Hate speech recognition systems can improve their contextual understanding capabilities to better recognize the intent behind language usage, by analysing the broader conversation or the history of interactions between the speakers.
- Adapting to new trends: The systems can keep up with new trends in hate speech and update their models accordingly, such as recognizing new vocabulary and identifying the shifting nuances of meaning and context.
- Combating algorithmic bias: To prevent discrimination and ensure fairness, hate speech recognition systems can strive to reduce algorithmic bias, by improving data quality, diversifying data sources, and including a broader range of perspectives in the development process.

Overall, the future scope of hate speech recognition systems for social media platforms such as Twitter involves improving the accuracy and effectiveness of hate speech detection, while also considering ethical and legal implications, and taking steps towards creating a more inclusive and respectful online environment.

If work on data as for training the model accurately, massive amount of data is required in order to train the working of the model more efficiently.

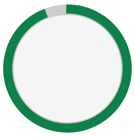
BIBLIOGRAPHY

- https://docs.streamlit.io/library/get-started
- https://www.nltk.org/book/
- https://stackoverflow.com/
- https://www.quora.com/What-are-the-advantages-of-using-a-naive-Bayes-for-classification
- https://towardsdatascience.com/document-feature-extraction-and-classification-53f0e813d2d3

Dupli Checker

PLAGIARISM SCAN REPORT

6%
Plagiarised



94%
Unique

Date	2023-05-20
Words	837
Characters	5433

Content Checked For Plagiarism

OBJECTIVE

This project aims to combine multiple services and open-source tools to make an automated content moderation system that will identify and detect whether the speech is recognized as “Hate Speech” or “Non – Hate Speech” using machine learning algorithms and with the help of sentimental analysis it will further detect the compound score of the speech using VADER.

INTRODUCTION

In recent years, hate speech on social media, particularly on Twitter, has become a major issue. The platform, which allows users to post short messages known as tweets, has been used to spread hateful and harmful messages based on race, religion, gender, sexual orientation, and other characteristics. As a result, there is growing concern about the impact of hate speech on social media on both individuals and society as a whole.

Hate speech on social media can have a number of negative impacts on individuals. For those who are targeted by hate speech, it can be extremely distressing and damaging to their mental health. It can also lead to feelings of isolation and fear, as well as physical harm in extreme cases. For those who witness hate speech, it can lead to a sense of powerlessness and frustration, as well as a feeling of being unsafe in online spaces. Hate speech on social media can also have broader societal impacts.

It can contribute to the normalization of hateful attitudes and beliefs, as well as the spread of misinformation and propaganda. It can also contribute to the polarization of society, making it more difficult to engage in meaningful dialogue and compromise.

One of the challenges of addressing hate speech on social media is the tension between free speech and the need to protect individuals from harm. While free speech is an important value, it is not absolute, and there are limits to what can be said without causing harm. Hate speech, by definition, is speech that seeks to denigrate or harm individuals or groups based on their characteristics. As such, it is not protected by free speech laws.

In recent years, the rise of hate speech on Twitter and other social media has become a major problem. Although it is difficult to measure the exact extent of the increase in hate speech on social media, there is evidence that hate speech has become more visible in recent years.

Anti-Semitic tweets on Twitter increased by 30% between 2016 and 2018, according to the Anti-Defamation League report. Similarly, the number of

white supremacist tweets on the platform increased by 30% over the same period. The report also found a significant increase in the use of Twitter by extremist groups and individuals, with many using the platform to spread hatred and incite violence.

A Pew Research Centre study found that 41% of Americans have experienced online harassment, and many cite hate speech as one of the biggest forms of harassment. The study also found that social media platforms such as Twitter are the most common venue for online harassment.

Similarly, in Europe, the rise of far-right movements and the spread of hateful speeches using social media has become a major problem. A study by the Institute for Strategic Dialogue found that the use of anti-Semitic language on social media platforms has increased significantly in Europe, with Twitter being one of the most commonly used platforms for this purpose. It's one. Overall, statistics suggest that the rise of hate speech on social media, including Twitter, is a significant problem worldwide. Social media platforms have taken some steps to address this issue, but more needs to be done to prevent these platforms from being used as vehicles for hate and intolerance.

CONTENT MODERATION: -

Content moderation refers to the screening of inappropriate content that users post on a platform.

The objective of content balance is to make a secure and aware environment for all clients, whereas moreover protecting freedom of expression and advancing sound wrangle about and talk. This could include expelling substance that abuses the platform's rules, such as despise discourse, badgering, or realistic viciousness, as well as taking activity against clients who abuse these rules, such as by suspending or forbidding their accounts. Content Moderation can be performed by both people and algorithms, depending on the stage and the sort of substance being directed. Human arbitrators are frequently mindful for looking into and evacuating more complex or touchy substances, such as abhor discourse or deception, whereas calculations can be utilized to identify and expel spam or low-level offences, such as copy posts or foulness.

The goal of content moderation is to ensure the platform is safe to use and upholds the brand’s Trust and Safety program.

Content Moderation is an imperative viewpoint of online community administration, and stages contribute critical assets in creating arrangements, apparatuses, and workflows to guarantee that it is performed successfully and effectively. Be that as it may, substance control is additionally a complex and always advancing field, and stages must adjust the requirements for security and free expression with the challenges of scale, setting, and client protection.

Matched Source

Similarity 6%

Title:Content Moderation - Services - GDC

https://www.gdc-services.com/services/business-process-outsourcing/content-moderation/

Similarity 5%

Title:imagga.com › blog › what-is-content-moderationWhat Is Content Moderation? | Types of Content Moderation ...

Sep 8, 2021 · The goal of content moderation is to ensure the platform is safe to use and upholds the brand’s Trust and Safety program. Content moderation is widely used by social media, dating websites and apps, marketplaces, forums, and similar platforms. Why Is Content Moderation Important?

https://imagga.com/blog/what-is-content-moderation/

Dupli Checker

PLAGIARISM SCAN REPORT

0%
Plagiarised

100%
Unique

Date	2023-05-20
Words	980
Characters	6520

Content Checked For Plagiarism

WHY CONTENT MODERATION IS ESSENTIAL?

Some arguments claim that any form of content moderation violates the fundamentals of free speech. But it's a very black-and-white discussion in a very grey area. Simply put, content moderation is essential to keep the platform safe from offensive, threatening, violent, illegal and pornographic posts. Illegal content such as child pornography and videos depicting unjustified violence are routinely uploaded, and this is clearly an area where content moderation is essential.

On the other side of the scale, there is a ton of visual content that is illegal for copyright reasons. With the sheer amount of content uploaded every day, content moderation is essential. Leaving this "free for everyone" results in social media feeds filled with horrific content, personal bullying, and the internet flooded with copyright-infringing videos and images.

CONTENT MODERATION CASE EXAMPLE – FACEBOOK: -

Looking at how a platform like Facebook deals with the issue of content moderation gives a good idea of the problems that all "mega-platforms" face when trying to keep content within their guidelines. According to NYU Stern report, Facebook's "human moderators" review more than 3 million photos, videos and messages every day. This may not seem like a bad thing, but keep in mind that many of these posts contain videos that could be longer. This simply means that moderators often have a few seconds to rate other posts. When you work under such pressure, you often make mistakes. Facebook CEO Mark Zuckerberg admitted in a recent white paper that moderators made the wrong decision about 10% of the time. This means that 300,000 messages a day are incorrectly tracked.

WHAT EXACTLY CONSTITUTES UNSUITABLE CONTENT?

The moderation of content can be conducted manually by human moderators or through automated techniques, such as machine learning algorithms. Manual content moderation entails the individual inspection and appraisal of each uploaded content piece to make a decision on whether it adheres to the guidelines or should be removed. Automated content moderation entails the use of algorithms to scan and flag content that infringes platform policies.

Content moderation is a critical aspect of ensuring a secure and wholesome online environment for users. It assists in preventing the dissemination of dangerous or inappropriate content and ensures that the platform is utilized for its intended purpose.

PROBLEM WITH THE HUMAN CONTENT MODERATION: -

- The Human cost of moderation: Moderators have been put under a lot of pressure to make ends meet, which has caused them to suffer from mental health issues like burnout and PTSD. It's bad enough that they have to work so hard, but now there are even more reports of them suffering from PTSD. It's easier to understand this when you look at the kind of content they have to deal with every day. According to a report from NYU Stern, in the first three months of 2020, there were a lot of posts that were removed for various reasons.

- Every Major Tech company uses third-party contractors for content moderation. This increases the overall cost with not efficient in removing unsuitable content.

What are the linguistic variations between hate speech and offensive language?
Linguistically, the distinction between hate speech and offensive language frequently comes down to how it targets oppressed communities, often in threatening ways.
Despite the fact that the labels contain extremely similar commonly recurring words, just 20% of the "hate speech" label is unique in total.
For example, hate speech frequently contains the n-word with a hard 'r.'
The use of this slur may suggest malicious intent, which goes beyond the possibility of the word being used as cultural slang.
Such examples highlight the complexities of English slang as well as the narrow boundary between hate speech and inappropriate language. because the vocabulary of each label is so similar; machine learning algorithms may struggle to distinguish between them and establish what constitutes hate speech.

What are the popular hashtags in the dataset of each tweet type?
Some most popular hashtags used in the tweets by the users.
The overall polarity of the tweets: -

CLASS IMBALANCE: -
The main obstacle of this dataset is the extreme class imbalance which basically affects the whole model. The dataset contains 5.77% which is labelled as hate speech. This could present challenges during the modelling process.

STEPS INVOLVING IN BUILIDING THE PROJECT

1) DATA SOURCING
2) DATA CLEANING
3) DATA PREPROCESSING
4) EXPLORATORY DATA ANALYSIS
5) FEATURE ENGINEERING
6) MODEL PROCESS
7) WORKING ON BASELINE MODELS
8) MODELS EVALUTAION ON THE BASIS OF THEIR ACCURACY
9) FINALIZATION OF THE MODEL
10) WEBSITE

1. Data Sourcing: The dataset for this project was sourced from a study named “Automated Hate Speech Detection and the Problem of Offensive Language” which was conducted by the Thomas Davidson and a team at Cornell University in the year of 2017.
Dataset is a .csv file with 24,802 text posts from twitter where almost 6% of the tweets are labelled as “Hate Speech”.

2. Data Cleaning: This is the first step that involves the cleaning of the sourced dataset. The aim of information cleaning right here is to find the proper manner to rectify high-satisfactory issues like eliminating bad data, and filling in missing data to form the information efficient for the model.

3. TEXT PREPROCESSING: This step involves the following transformations.

- Lowercase
- Tokenization
- Removing special characters
- Removing stop words and punctuation

4. EXPLORATORY DATA ANALYSIS: EDA is performed on the cleaned dataset to figure out the features and uniqueness of the data.

5. FEATURE ENGINEERING: The goal of feature engineering in Natural Language Processing is to convert tokenized text data into numerical vectors that the machine learning algorithm can "understand." We'll be iterating the dataset with the three different feature engineering techniques in our notebook: count vectorization, Doc2Vec and TF-IDF vectorization. Trying these techniques on the same four baseline models could result in vastly different metrics.

Matched Source

No plagiarism found

Dupli Checker

PLAGIARISM SCAN REPORT

0%
Plagiarised

100%
Unique

Date	2023-05-20
Words	1000
Characters	6549

Content Checked For Plagiarism

6. MODEL PROCESS:

Metrics for Evaluation: We will use the F1 score as the primary metric for this business problem, while also considering Precision and Recall. The F1 score calculates the harmonic mean of Precision and Recall, which is useful for data with a high-class imbalance.

Overall, we want as much hate speech as possible to be flagged and efficiently removed. This entails optimising the True Positive Rate, also known as Recall.

- F1 Weighted Score: We will also consider the weighted F1 score, which can account for label class imbalance by calculating metrics for each label.

• Cross Validation: In order to check out whether a model is overfitting or underfitting, we can use K-Fold Cross Validation to generate an F1 Score for the training set.

7. WORKING WITH BASELINE MODELS: Used multiple baseline machine learning models (Random Forest, Logistic Regression, Naïve Bayes, Support Vector Machine.

Out of the above models, the best model with the most accuracy and best performance will be selected to iterate the further procedure of the model.

Random Forest: Random Forest: A random forest is a classification algorithm which consists of a variety of decision trees on different subsets of a given data set and takes the mean to improve the prediction performance of that data set. The greater the number of trees, the greater the accuracy and equipment risk. the lower the surplus.

- Random Forest is able to execute both classification and regression tasks.
- It improves model accuracy and avoids the problem of overfitting.

Support Vector Machine: SVM is primarily utilised in classification tasks. Each data item is plotted as a point in an n-dimensional space by the algorithm (where n is the number of features it possesses). where the value of each feature is the value of a single coordinate. The classification is then carried out by locating hyperplanes that clearly distinguish the two classes. The support vector machine is the best boundary between the two classes (hyperplane and line). SVM can employ kernels if your data demands nonlinear categorization. This is a function that turns a low-dimensional input space into a high-dimensional space. H. They separate problems that were previously intertwined.

Naïve Bayes: Naive Bayes is a simple yet powerful probabilistic algorithm that is widely used in machine learning for classification tasks. The algorithm is based on Bayes' theorem, which describes the probability of a hypothesis given evidence.

In classification problems, the goal is to assign a label (or class) to an input data point based on its features. The naive Bayes algorithm assumes that the features are independent of each other, which simplifies the computation of the probability of the data given the class (or likelihood) and the probability of the class (or prior).

The naive Bayes algorithm calculates the probability of each class for a given data point using the Bayes' theorem and then selects the class with the highest probability as the predicted class. The probability of each class is computed as the product of the likelihood of the features given the class and the prior probability of the class, divided by the probability of the data.

The algorithm is called "naive" because it assumes that all features are independent and equally important, which is often

not true in practice. However, despite this simplifying assumption, naive Bayes can perform surprisingly well in many real-world applications, especially when the number of features is large.
Overall, Naive Bayes is a popular and effective algorithm for classification tasks in machine learning, especially for text classification and spam filtering.
Logistics Regression: Logistic regression is a popular statistical method in machine learning for binary classification applications. Based on a collection of input features, the purpose of binary classification is to predict a binary output, such as "yes" or "no." Logistic regression employs a logistic function to represent the relationship between the input features and the output variable, yielding a probability estimate for each class.

Logistic regression is a simple yet powerful technique for binary classification problems that is widely utilised in a wide range of applications, including customer churn prediction, fraud detection, and medical diagnosis. It is interpretable, which means that the learned
parameters can be used to comprehend the link between the input features and the output variable
8. MODEL EVALUATION ON THE BASIS OF THEIR ACCURACY:
9. FINALIZATION OF THE MODEL: The best performing model with the best accuracy among all the algorithms is logistics regression model using Count Vectorizer for feature engineering. The hyperparameters were penalty = 12 and class weight = 'balanced'.

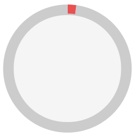
10. DEPLOYMENT OF THE MODEL OF THE WEB PAGE:
After successfully performing the steps involves from data cleaning to finalization of the model.
The model is deployed on the web page with the help of the Streamlit framework.
PROBLEM STATEMENT
The use of human content moderation in social media platforms has been a subject of scrutiny due to reports of the exploitation of workers. In 2019, an article published by The Verge uncovered the extensive list of inhumane working conditions that employees faced at Cognizant, which was Facebook's former moderation contractor. The article shed light on how content moderators were subjected to traumatic and distressing content without adequate support or counselling and were paid low wages despite the highly demanding and psychologically taxing nature of their work. However, it is important to note that the use of human moderators is still prevalent among major tech companies, including Twitter, both domestically and overseas.
Hate speech, which is defined as abusive or threatening language that expresses prejudice against a particular group, based on factors such as race, religion, or sexual orientation, is a major concern for social media platforms. However, detecting and addressing hate speech is a complex challenge for social media platforms, as it requires a nuanced understanding of language and context, as well as ethical and legal considerations. It is worth noting that datasets used to train hate speech recognition systems can sometimes contain offensive language, which may be triggering or harmful to some individuals.

Matched Source

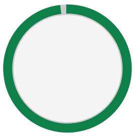
No plagiarism found

Dupli Checker

PLAGIARISM SCAN REPORT



2%
Plagiarised



98%
Unique

Date	2023-05-20
Words	961
Characters	6677

Content Checked For Plagiarism

SOFTWARE REQUIREMENT
SPECIFICATIONS

1. PyCharm
2. Jupyter Notebook
3. Streamlit (Frontend)
4. Some Preinstalled & External libraries (Pickle, Pandas, Scikits Learn etc)

PyCharm: -

PyCharm is designed to increase productivity and efficiency by offering a variety of features and functionalities, such as code completion, syntax highlighting, and project management tools.

PyCharm simplifies the process of developing Python applications by providing a user-friendly interface that facilitates coding and debugging. It supports a range of Python frameworks and libraries and includes built-in tools for testing and debugging code. PyCharm also offers integrations with other development tools, such as version control systems and package managers, to streamline the development process.

Jupyter Notebook: -

Jupyter notebooks are used for exploratory data analysis (EDA), data cleansing and transformation, data visualisation, statistical modelling, machine learning, and deep learning, among other data science tasks.

Jupyter notebooks are particularly excellent for "presenting the work" done by your data team by combining code, markdown, links, and photos. They're simple to use and may be run cell by cell to see how the code works.

Streamlight (Frontend): -

Streamlit is a fantastic new tool that allows engineers to quickly create highly dynamic online applications based on their data, machine learning models, or anything else.

NumPy:-

It is used for working with arrays and linear algebra.

Pandas: -

It is used for data analysis and data pre-processing, CSV file I/O
(e.g. pd.read_csv)

Sklearn: -

It is used for making use of Machine learning tools. In this project, we have used cosine similarity for recommendations.

NLTK: -

NLTK is a toolkit built for working with NLP in Python. Various word processing libraries are provided with many test datasets. NLTK allows you to perform various tasks such as B. Tokenization, parse tree visualization, etc.

CONCLUSION

It is critical to comprehend why the model performed poorly and how this connects to the business issue. The "Hate Speech" label forecasts lowered the F1 score. The algorithm successfully predicted 91% of the "Not Hate Speech" classifications, but only 62% of the other labels.

Its performance reflects the project's two key roadblocks:

- The dataset's massive class imbalance Model's incapacity to "understand" the intricacies of hate speech

The problem of class imbalance can be solved using pre-processing and oversampling/under-sampling approaches. Yet, recognising hate speech is a broader issue that many large technology companies, such as Twitter, Facebook, and Instagram, are still grappling with.

This is due to the fine line that exists between Hate Speech and plain inflammatory words. Hate speech is described as abusive or threatening speech that conveys prejudice against a certain group, particularly based on race, religion, or sexual orientation. The distinction between hate speech and offensive language is usually determined by small differences in context or diction.

We can observe that the "Hate Speech" label has 679 terms that are unique to it. Some of these terms are illogical or meaningless, but others are particularly offensive. This label, for example, comprises phrases like "sp-cs", "mo-kies", "ov-nj-w", "fa—ot" and many more. All of these are racial or homophobic slurs. More importantly, we saw in earlier EDA that this label contains the N-word with the hard "R" excessively. By specifically targeting underprivileged communities, language like this obviously reveals hate speech.

Unfortunately, it is difficult for a machine learning model to recognise the intricacies of this derogatory language. Despite the fact that these words are unique to the "Hate Speech" label corpus, there is simply insufficient data for that label.As a result, a critical next step would be to collect more data that CrowFlower voters have recognised as hate speech.

Furthermore, for the time being, we can recommend that Twitter manually regulate tweets containing terms, much as it could with the top hashtags discovered for each category in previous EDA.

Furthermore, due of the intricacies in English slang and slurs, automated hate speech identification is an incredibly challenging task. This effort was able to start that process, but much more work needs to be done to keep this stuff off of public-facing venues like Twitter.

LIMITATIONS

There are several challenges that are faced in hate speech recognition systems for social media platforms such as Twitter. Some of the key challenges include:

- Contextual understanding: Hate speech can be highly contextual, and what constitutes hate speech may vary based on factors such as the target group, the historical and cultural context, and the intention of the speaker. Therefore, developing algorithms that can accurately detect hate speech in different contexts can be a challenge.
- Language diversity: Social media platforms such as Twitter have a diverse user base, and people use different languages and dialects to communicate. Recognizing hate speech in multiple languages requires a large amount of training data and the ability to handle variations in language structure and expression.
- Sarcasm and irony: Hate speech is not always direct and can be expressed through sarcasm, irony, and other forms of indirect language. Recognizing such language requires understanding of the underlying context and the ability to identify patterns and connections between different phrases.
- New trends and terminology: The language and terminology used to express hate speech can evolve quickly and vary across different communities. Developing algorithms that can recognize new trends and terminology requires continuous monitoring and updates to the training data.
- Legal and ethical considerations: Hate speech detection systems need to balance the need to protect individuals from harmful speech with the need to preserve freedom of expression. Developing algorithms that can accurately identify hate speech while avoiding false positives and respecting the diversity of opinions can be a challenging task.
- Overall, developing effective hate speech recognition systems for social media platforms such as Twitter requires overcoming these and other challenges, and requires continuous refinement and improvement as new trends emerge and language evolves.

Matched Source

Similarity 8%

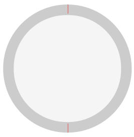
Title:Project Title: Depression Detection Analysis

NLTK is a toolkit built for working with NLP in Python. It provides us with various text processing libraries with a lot of test datasets.

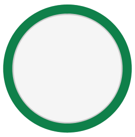
https://voisfortech.com/innovation-marathon-2022-project-reports/Team261.pdf

Dupli Checker

PLAGIARISM SCAN REPORT



0%
Plagiarised



100%
Unique

Date	2023-05-20
Words	291
Characters	2295

Content Checked For Plagiarism

FUTURE SCOPE

The future scope of hate speech recognition systems for social media platforms such as Twitter is vast and promising. Here are some potential areas of growth and improvement:

- Multilingual support: Hate speech recognition systems can expand their language support beyond English to include a wider range of languages, as social media users from all over the world communicate in different languages.
- Fine-grained classification: Hate speech recognition systems can aim for more nuanced classification of hate speech, such as identifying specific types of hate speech, or distinguishing between hateful speech and speech that is simply controversial or offensive.
- Contextual understanding: Hate speech recognition systems can improve their contextual understanding capabilities to better recognize the intent behind language usage, by analysing the broader conversation or the history of interactions between the speakers.
- Adapting to new trends: The systems can keep up with new trends in hate speech and update their models accordingly, such as recognizing new vocabulary and identifying the shifting nuances of meaning and context.
- Combating algorithmic bias: To prevent discrimination and ensure fairness, hate speech recognition systems can strive to reduce algorithmic bias, by improving data quality, diversifying data sources, and including a broader range of perspectives in the development process.

Overall, the future scope of hate speech recognition systems for social media platforms such as Twitter involves improving the accuracy and effectiveness of hate speech detection, while also considering ethical and legal implications, and taking steps towards creating a more inclusive and respectful online environment.

If work on data as for training the model accurately, massive amount of data is required in order to train the working of the model more efficiently.

BIBLIOGRAPHY

- https://docs.streamlit.io/library/get-started
- https://www.nltk.org/book/
- https://stackoverflow.com/
- https://www.quora.com/What-are-the-advantages-of-using-a-naive-Bayes-for-classification
- https://towardsdatascience.com/document-feature-extraction-and-classification-53f0e813d2d3

Matched Source

No plagiarism found

Check By: Dupli Checker