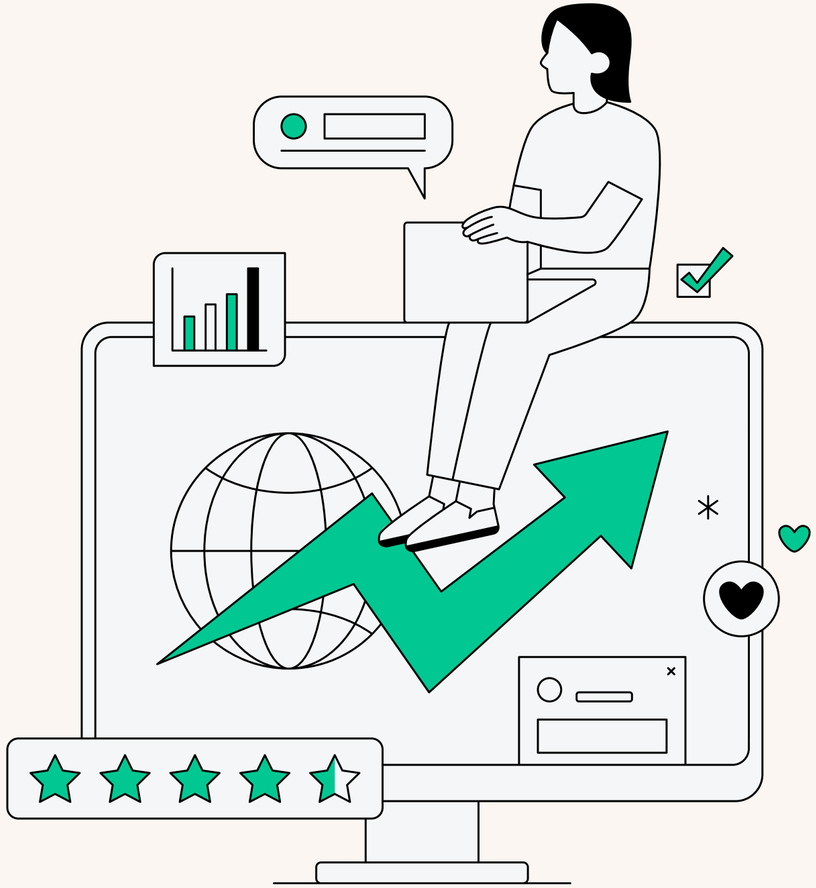
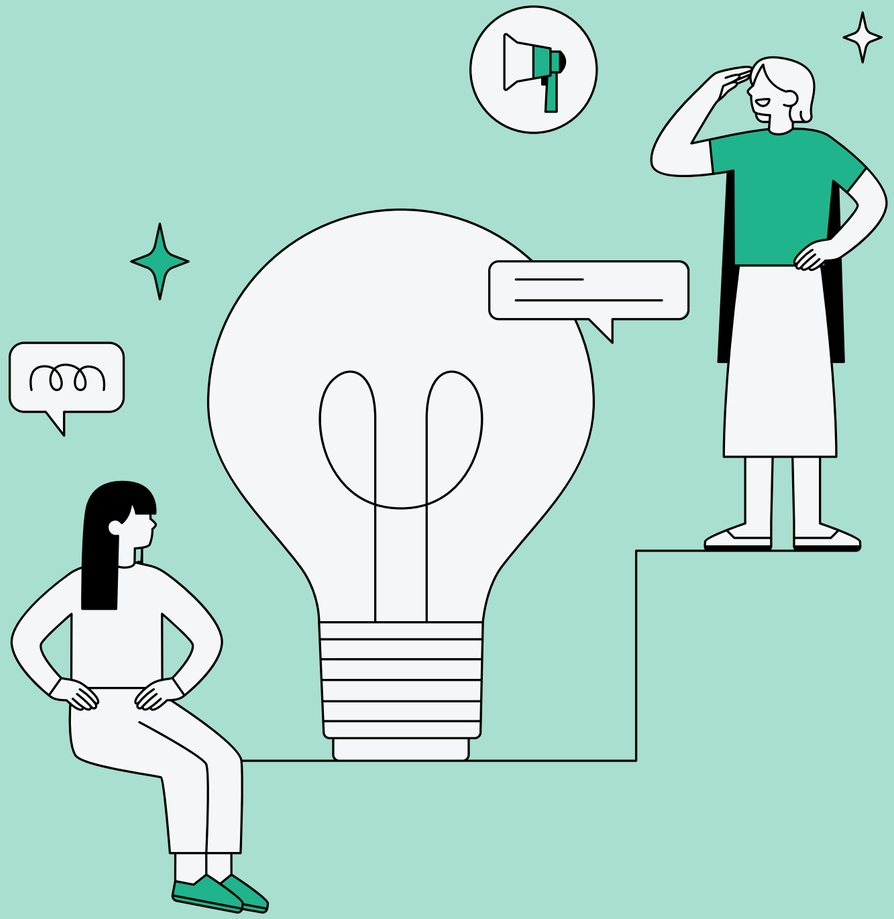
Cyberbullying Classification & Sentiment Analysis



Group 5

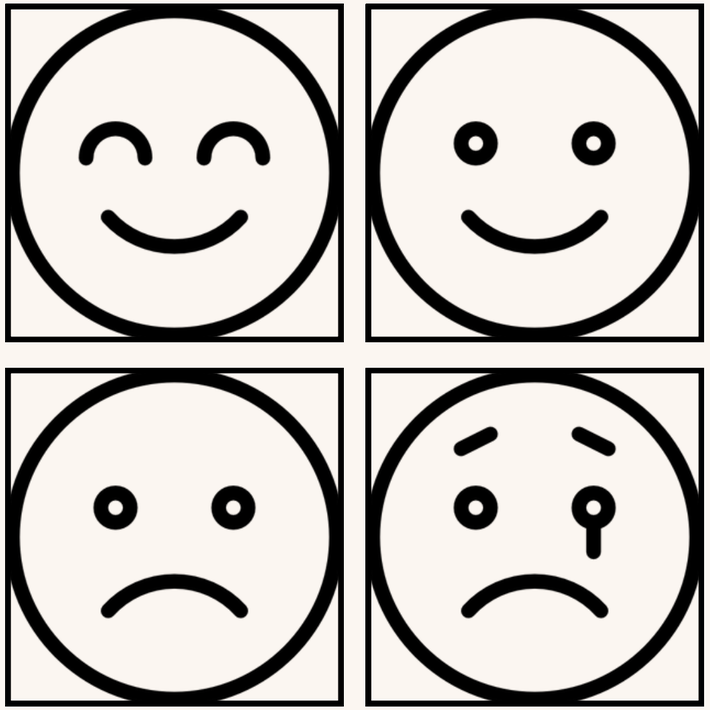
1. Introduction

Twitter has grown to be a popular self-expression platform that promotes global interaction but also allows harmful activities like cyberbullying. Bullying in this case has the potential to inflict long-term mental health conditions as well as great emotional suffering. Because there is so much content created online, standard monitoring techniques are not enough to stop hate speech and enhance user safety. As a result, businesses have invested in technology like natural language processing.

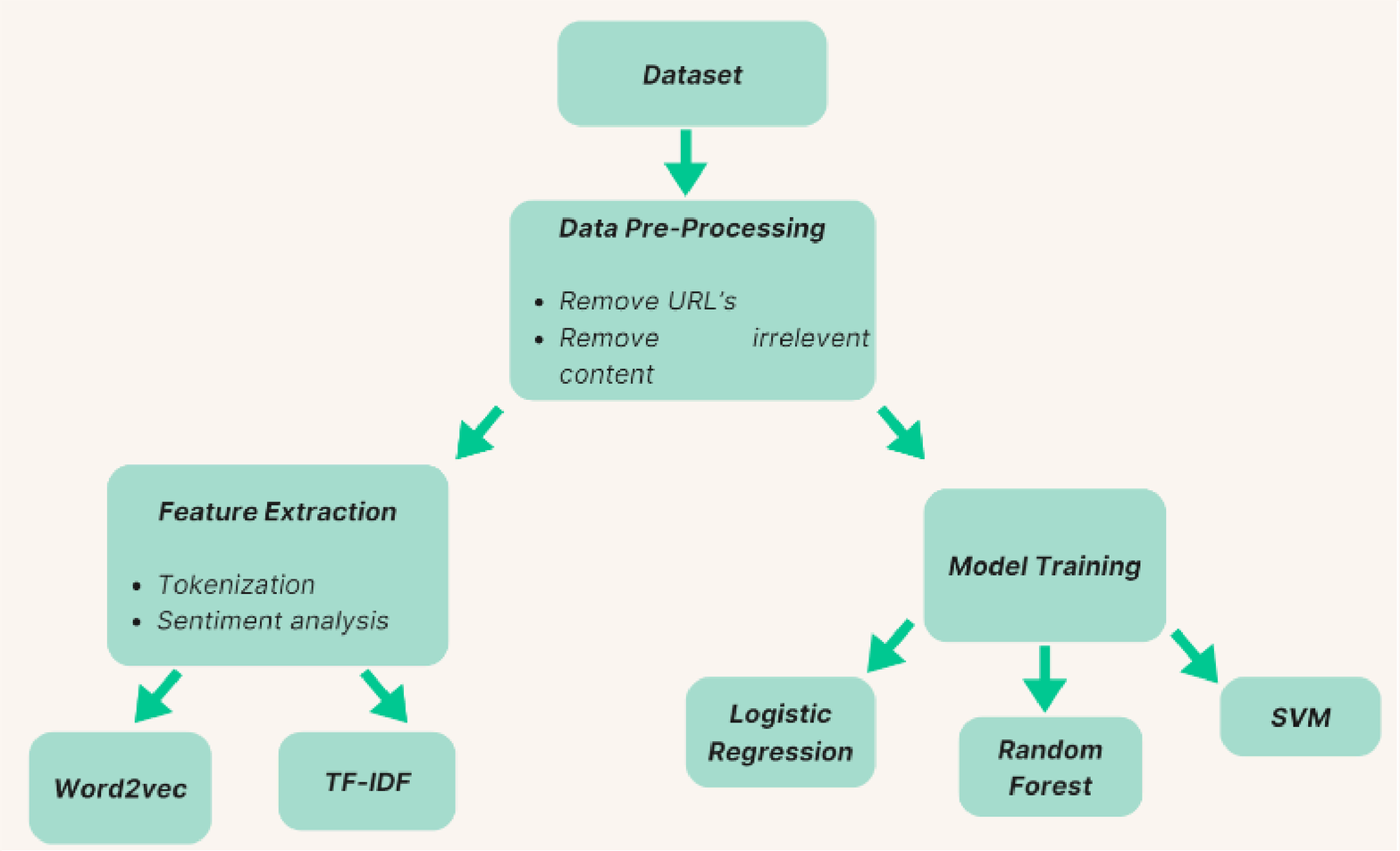


2. Project Overview

The objective of this project is to create a sentiment analysis system that will categorize tweets as positive, negative, neutral, or harmful (e.g., racism, sexism). With social media's rise, content moderation and hate speech identification can be improved by having a better understanding of user feelings. The project aims to address natural language processing challenges through phases including feature extraction, data cleaning, collection, and machine learning model training.



3. Methodology



3.1 Dataset

The dataset for this project, sourced from Kaggle, includes 16,848 historical tweets categorized into three classes: “None,” “Sexism,” and “Racism.” Tweets labeled as "1" denote hate speech, while those marked "0" are neutral and discarded due to uncertainty about their content.

0	**11501**
1	**5347**
Total = 16848	

3.2 Data Pre-Processing

The process begins with cleaning tweets by removing URLs and irrelevant expressions, as they don't indicate hate speech. Mentions and hashtags assist in identifying the origin of hate speech. After that, we divide the continuous text into meaningful parts by tokenizing the tweets using the Natural Language Toolkit (NLTK). Lastly, we locate negation words and their coverage, including all text up until the following punctuation or constraint word (such as "but," "however"), in order to create a negation vector.

3.3 Feature Extraction

Hate is a negative sentiment, making sentiment polarity crucial for identifying potentially hateful tweets. We selected two systems for our project :

- Word2Vec
- & TF-IDF

Word2Vec is easy to be used since it effectively utilizes the frequency of words. Nonetheless, the Term Frequency-Inverse Document Frequency (TF-IDF) model provides better differentiation between the tweets making the sentiment analysis more useful.

3.4 Model Training

For classification, we used several types of supervised learning models, such as Support Vector Machine (SVM), Random Forest, and Logistic Regression. We are able to compare their performance as a result. Key performance indicators (KPIs) such as true positives, precision, recall, and F1-score were used to assess them .

$$\text{F1 Score} = 2 \times \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

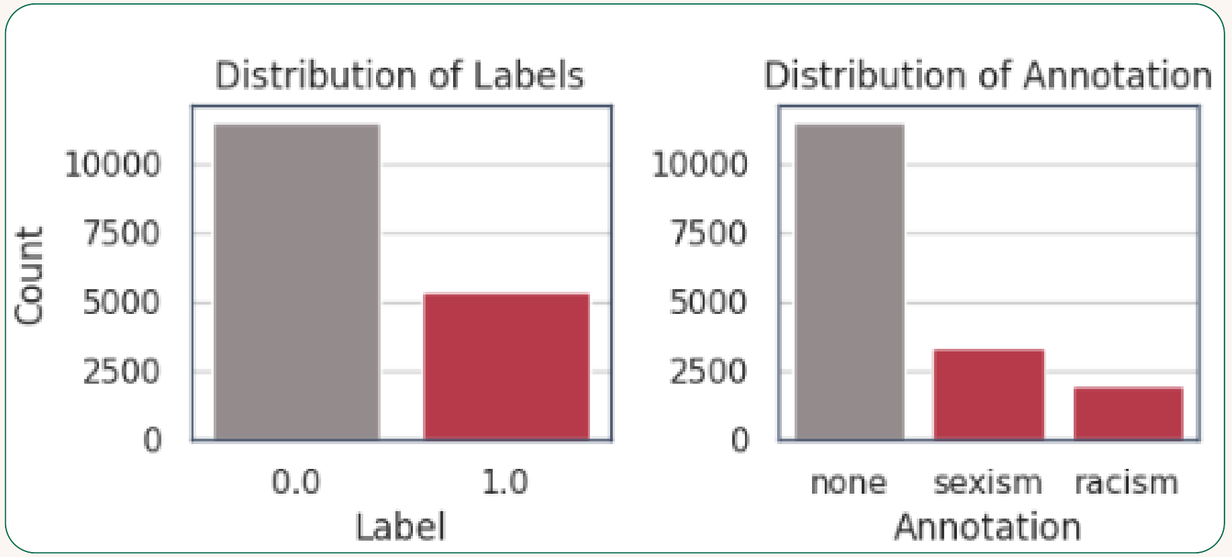
	0	1	Accuracy
SVM	0.89	0.73	0.84
Random Forest	0.89	0.67	0.83
Logistic Regression	0.88	0.66	0.82

All models showed good accuracy, however Random Forest and SVM had the best F1-scores. SVM performs well with high-dimensional text data and offers a thorough method of sentiment identification; Random Forest manages non-linear correlations, while Logistic Regression is more straightforward and comprehensible.

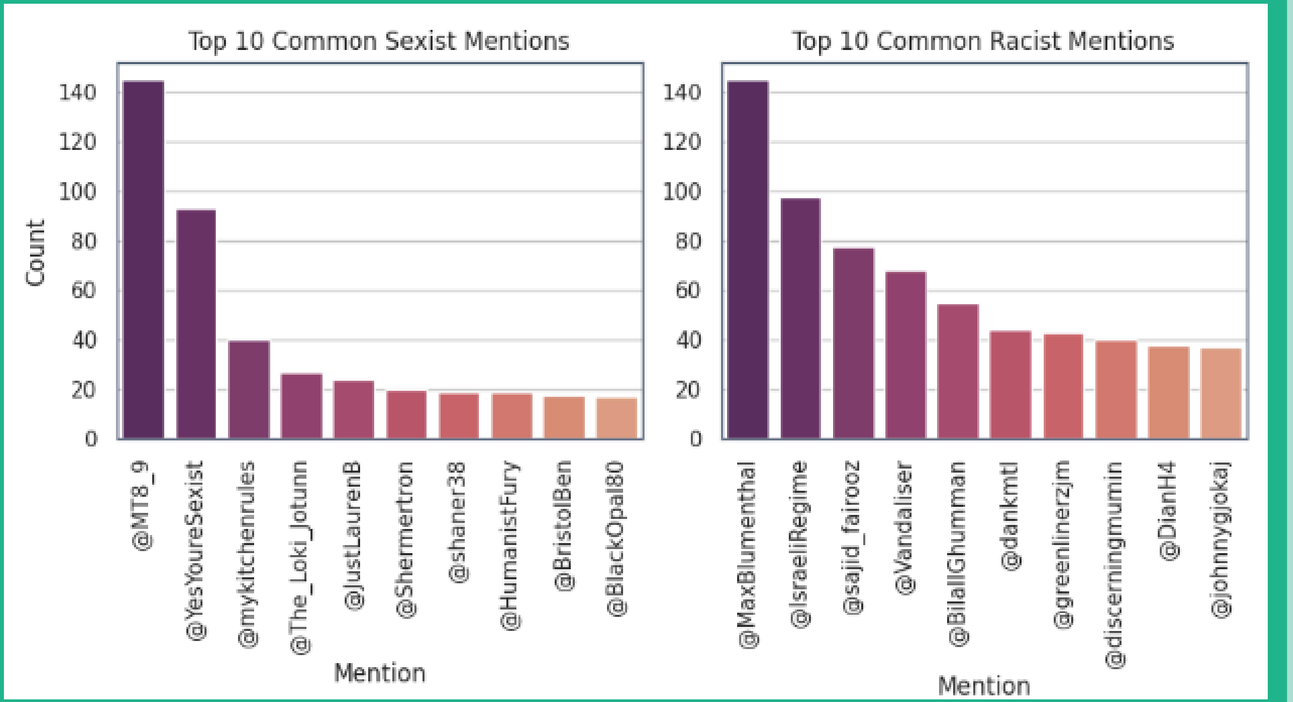
4. Sentiment Analysis

4.1 Label and Annotation

In our dataset, sentiment category 1 is divided into sexism (3,377 tweets) and racism (1,970 tweets), with sexism being more common. This may be due to stronger social penalties for racist content and better monitoring algorithms on platforms like Twitter. We chose not to balance the categories to reflect a more realistic environment, even if it affects accuracy.



4.2 Mentions



The dataset includes mentions to many accounts in racist and sexist tweets. Interestingly, @YesYoureSexist is frequently mentioned in sexist tweets, yet instead of encouraging sexism, it draws attention to it. Accounts such as @JustLaurenB and @MT8_9, on the other hand, promote sexist sentiments. When it comes to racist tweets, @MaxBlumenthal's controversial opinions lead to discussions that become heated over racism. The difficulty in determining meaning in sentiment analysis is shown by this complexity.

4.3 Wordcloud

Words that are linked to women, such as "feminism," "women," and "female," dominate the dataset's sexism category, while references to "men" are less common. In the racism category, phrases such as "Islam," "Muslim," and "Quran" are frequent, indicating that Muslims are a primary target. Religious groups encounter greater hate speech than specific ethnicities, highlighting a need for future research to improve hate speech identification by examining high-frequency terms as triggers.

5. Discussion

Three models are compared in the discussion: SVM, Random Forest, and Logistic Regression. For non-offensive tweets, Logistic Regression had high precision (0.82) and recall (0.94), but it struggled with offensive tweets, showing a recall of 0.56. Random Forest achieved a recall of 0.96 for non-offensive content but had issues detecting offensive tweets (recall 0.55). The best model for precisely recognizing offensive content was SVM, which had the highest accuracy (84%) and balanced performance, with a recall of 0.66 for offensive tweets and a better F1-score (0.73).

Models	Precision	Recall	F1-Score	Support
Logistic Regression	0.82	0.94	0.88	1726
Random Forest	0.82	0.96	0.89	1726
SVM	0.81	0.66	0.73	802

Table 4.1 – Performance for Label 0

Models	Precision	Recall	F1-Score	Support
Logistic Regression	0.81	0.56	0.66	802
Random Forest	0.87	0.55	0.67	802
SVM	0.81	0.66	0.73	802

Table 4.1 – Performance for Label 1

6. Conclusion and Future Work

In order to evaluate sentiment and categorize cyberbullying, we employed three well-known machine learning models in this study: Support Vector Machines (SVM), Random Forest, and Logistic Regression. By investigating advanced algorithms like LSTM and BERT, adjusting hyperparameters, and enhancing feature extraction, future research may increase model precision. Developing sentiment lexicons related to hate speech and integrating contextual information, such as user behavior and tweet timing, may lead to enhanced detection of hate speech and cyberbullying, eventually improving classification accuracy.

Thank you very much!