FACULTY OF COMPUTING AND INFORMATICS

CDS6344 : SOCIAL MEDIA COMPUTING

TRIMESTER III, Session 2024/2025

Cyberbullying Classification & Sentiment Analysis

Group 5

Name	Matrics ID	Lecture Section	Tutorial Section
Muhammad Daniyal Bin Mohd Razman	1221304153	TC1L	TT1L
Muhammad Noor Hakimi Bin Sabri	1221305307	TC1L	TT1L

Table Of Content

1. Introduction

The existing data from social media networks shows that companies like Twitter have become important channels for talking and expressing oneself, because they allow users to communicate their thoughts, feelings and perceptions with people all over the world. Although this promotes fellowship and interaction, it also provides space for negative behaviours such as cyberbullying. Cyberbullying refers to the process of using electronic means to mistreat someone by mocking them or making threatening remarks. This has since morphed into one of the serious societal challenges especially among young ones. Online interactions can give people the confidence to engage in harmful behaviours that they might not display in person because of the anonymity and distance they provide. Because of this, cyberbullying victims frequently go through an immense amount of emotional distress, which can have long-term psychological effects like anxiety, depression, and in the worst situations, suicidal thoughts.

Social media companies like Twitter have spent a lot of resources on algorithms and training models that are able to detect hateful speech on their platform. This approach is to analyse hate crime escalation in social media, giving researchers solutions on how to resolve them appropriately and more efficiently. While the website offers an open space for people to discuss and share thoughts, There will always be the few that gravitate toward hateful speeches especially when someone is having a different opinion about a certain subject. For example, someone with different backgrounds, culture and beliefs. King and Suton [1] reported that hate crimes with an anti-Islamic motive occurred in the year following 9/11. However in recent time, with the rapid growth of Online Social Network (OSN), more conflicts are taking place, following each big event or other

Because of the harmful impact of cyberbullying, it is important to come up with effective strategies for identification and prevention. In business, hateful speech could affect the company's performance. Losing the amount of users, furthermore getting a negative reputation. Traditional methods of monitoring and reporting are mostly not enough due to the huge amount of content generated on social media platforms. Therefore, there is a promising solution in

advanced technologies like natural language processing (NLP) and machine learning that can help to counter this offence. Automation of the detection process will enable timely interventions and support for victims. This project aims to fight against cyberbullying by using these technologies to classify malicious tweets about it. Burnap and Williams [2] claimed that collecting and analysing temporal data allows decision makers to study the escalation of hate crimes following "trigger" events.

1.1. Project Overview

This project's objective is to create a system for sentiment analysis which could possibly classify tweets into the categories; positive, negative, neutral as well as harmful content such as racism or sexism. The increase in popularity of social media platforms like Twitter has made it possible for a lot of user generated content that often contains opinions and emotions. These sentiments once understood may lead to better understanding in a variety of situations including content moderation, public opinion evaluation, and identification of hate speech. The stages involved in this project range from data collection and clean-up, feature extraction model training allowing us to accurately identify sentiment. Using a dataset of tweets, this project seeks to resolve numerous challenges present at natural language processing (NLP) level including negation handling; text tokenization as well as use of machine learning algorithms for classification.

2. Problem Statement

Social media platforms like Twitter have become essential for individuals to share their opinions, emotions, and attitudes in today's hyper-connected world. However, with millions of daily posts, it can be challenging to analyze and interpret the vast amount of data due to several factors. First, social media text is often unstructured, lacking grammatical consistency and formality, which makes it difficult for traditional sentiment analysis techniques to process. Additionally, the presence of noise—such as slang, abbreviations, emojis, and informal

language—further complicates the extraction of meaningful insights. Moreover, the subtleties of language, including the use of sarcasm, irony, or negations, pose additional challenges in accurately determining sentiment.

These challenges are particularly critical when addressing harmful content, such as racist or sexist remarks that contribute to online abuse, cyberbullying, and toxic environments. Identifying and analysing such harmful speech requires robust techniques that can handle the nuance and variability of social media text while efficiently processing large volumes of data. This study aims to address these challenges by leveraging sentiment analysis and machine learning models to detect and categorise harmful speech on Twitter, specifically focusing on sexism and racism, in order to contribute to safer online communities.

2.1. Methodology

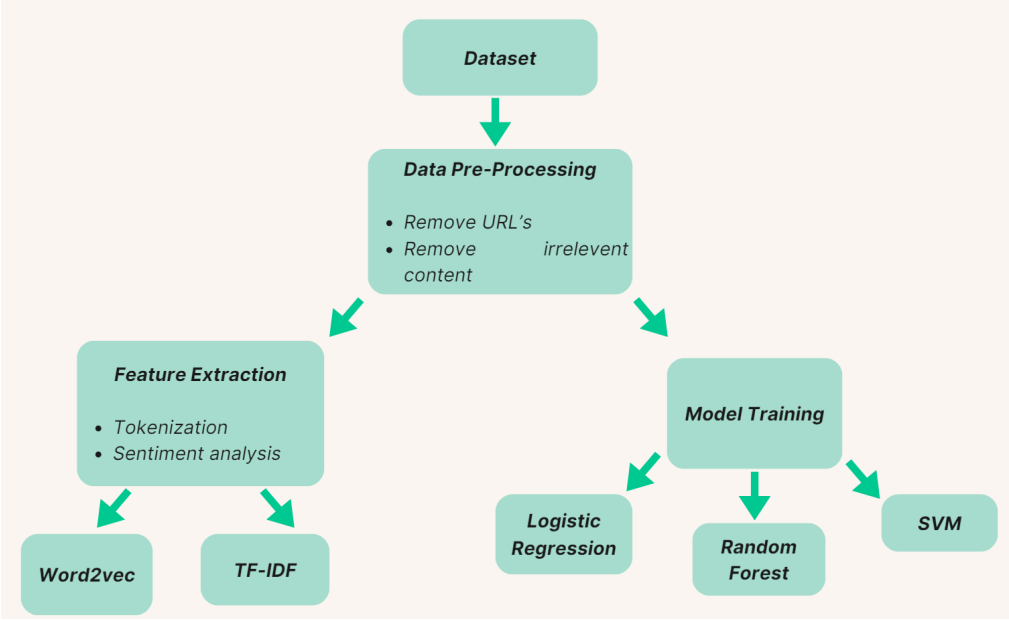

Figure 2.1 Map of the process throughout the project

2.1.1. Dataset

We have gathered a dataset for this particular project from Kaggle. This dataset consists of 14,881 historical tweets. Tweets on this dataset have been classified into one of the following three classes: “None,” “Sexism,” and “Racism”. The last two, referring to a specific form of hate speech, have been included as a part of “oh label” “1”, whereas the tweets of the class “0” have been discarded because there is no indication whether they are clean or offensive but rather a neutral one.

2.1.2. Data Pre-Processing

First step, we clean up the tweets by removing URLs and irrelevant expressions. This is because these do not add any information on whether the tweet might express hate or not. In particular for the case of hashtags and mentions, they could help to find the root cause and the person who starts the escalation of hate speeches.

Second step, we perform tokenization. Tweets are unconstructed data consisting of a continuous string of words and letters. Tokenizing will split the words into meaningful units. In this case, we implement Tokenization using the Natural Language Toolkit (NLTK). NLTK is a very powerful library for NLP and at the same time really easy to navigate through.

Afterwards, we clarified what can qualify as a negation vector: we detect the position of negation words and detect the coverage of these words. In the work of Das and Chen [3], negation words cover all that follows it until the next punctuation mark or the occurrence of a constraint word (e.g., “but,” “however,” etc).

2.1.3. Feature Extraction

In this part, we choose two systems that we think are compatible with our particular project. H. Watanabe et al [4] explained, Hate is basically a sentiment among others, a negative

sentiment to be precise. Relying on just sentiment polarity of the tweet is an important indicator of whether or not it can be a potential hateful tweet.

Word2Vec are simple and easy to be used, it also helps to capture the frequency of words, which can be very useful for sentiment analysis

Term Frequency-Inverse Document Frequency (TF-IDF) however can capture more meaningful distinction between tweets better than Word of Back.

2.1.4. Model Training

After the extraction of features, we continue to the final part of the project by using a multiple training model. The classification was done by using supervised learning models such as Logistic Regression, Random Forest, and Support Vector Machine (SVM). The use of multiple training model are so that we can compare between these models that perform the best

To evaluate the performance of classification, we use key performances indicators (KPIs) which are the percentage of true positives, the precision, the recall and the F1-score defined as shown below:

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

Figure 2.2 – F1 Score formula

Random Forest and SVM perform the best F1-score compared to Logistic Regression. However, all three of these models performed at such a high accuracy and have very well performed F1-score. It made sense since Logistic Regression is a simpler and interpretable model. Random Forest adds complexity and robustness by considering non-linear relationships and averaging multiple trees to prevent overfitting. SVM is particularly powerful for high-dimensional text data and offers flexibility with kernel functions to handle non-linear decision boundaries.

By using all three models, we will have a well-rounded approach to the problem, each with its own strengths in handling text classification and sentiment detection.

3. Sentiment Analysis

3.1. Label and Annotation

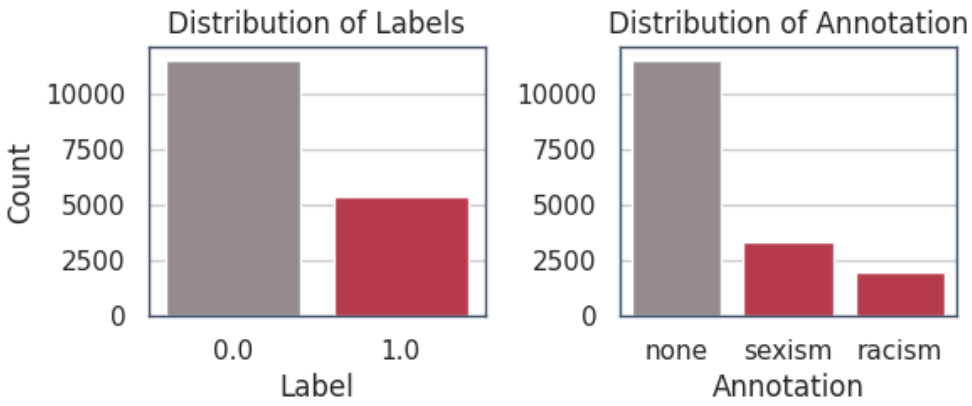

Figure 3.1 – Distribution of Labels and Distribution of Annotation bar chart

The label representing sentiment category 1 has been further subdivided into two distinct categories: sexism and racism. Upon analysis, it is evident that sexist content (3,377 tweets) is more prevalent in the dataset compared to racist content (1,970 tweets). This disparity could be attributed to a number of factors. There might be one reason. And that might be that people who are often critical of racist content are more likely to criticise them so there can be tougher social penalties against it and more notifications about such content. Furthermore, we can also think that monitoring algorithms on Twitter work better compared to Facebook hence reducing the chances of reading racist tweets. Basically, it means that although all types of hate speech (including racist one) are bad, their perception can be shaped within different media platforms depending on public knowledge or policy guidelines followed by the site management.

In most scenarios, having an evenly distributed category on the dataset will give a fairer result when we are going to give it to the training model. H. Watanabe et al [4] on his approach to sentiment analysis project, they take the least number of tweets data categories and limit the rest of the class number of tweets so each class will have a balanced amount.

We decided not to balance out the distribution between the classes, this to ensure that the model can predict and be trained in a more realistic environment. In return, the result may not be as accurate but may be more organic theoratically.

3.2. Mentions

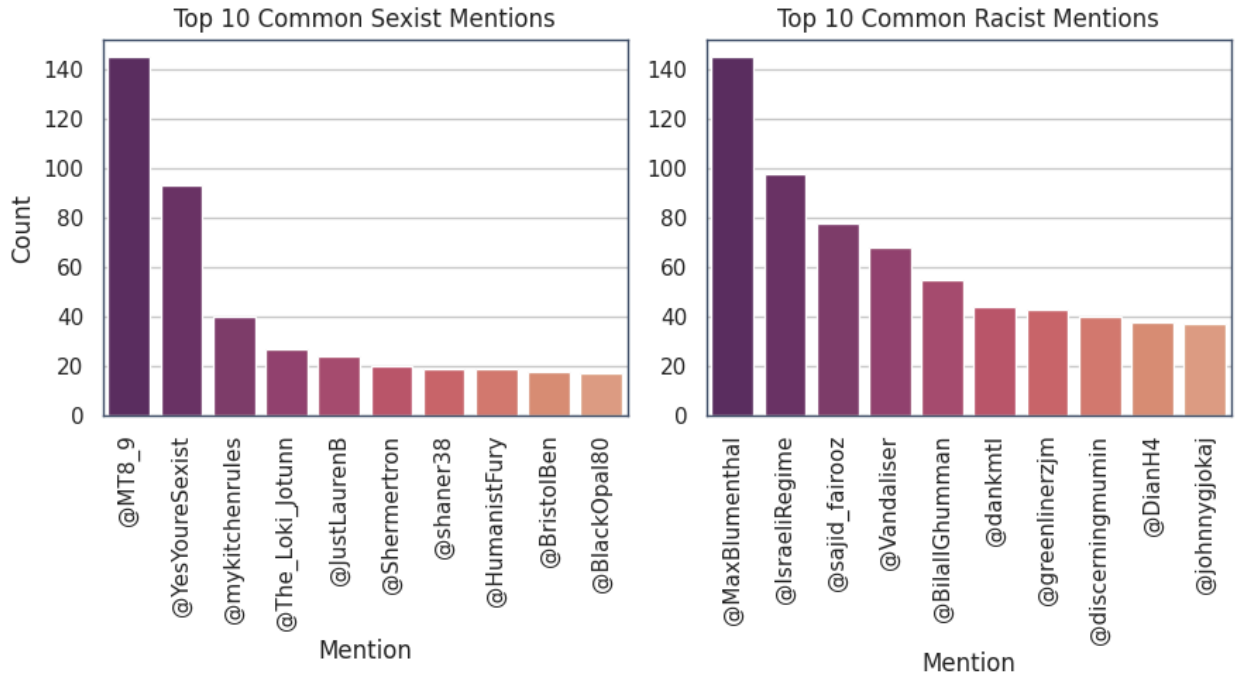

Figure 3.2 – Top 10 Common Sexist Mentions and Racist Mentions

Due to this dataset being almost a decade old, some of the accounts that are mentioned are no longer available on Twitter. So we picked a few mentions in the top 10 of both categories that are still on the platform. Further analysis of each category will be explained below.

3.2.1. Sexist Tweets

"@YesYoureSexist" ranks among the top mentions in tweets classified under the "sexist" category. However, a closer look reveals an interesting nuance: the account itself is not promoting sexist views. In fact, @YesYoureSexist is known for highlighting and calling out instances of sexism, often exposing discriminatory comments and behaviour. The irony lies in the fact that the provocative handle, which seems confrontational, has led to its frequent association with tweets that carry sexist content. The nature of social media means that many of the tweets mentioning @YesYoureSexist are often replies or discussions that engage with sexist attitudes, either by criticising them or inadvertently propagating sexist rhetoric through discussion

Meanwhile, accounts like @JustLaurenB and @MT8_9, which have been identified as spreading sexist comments, further complicate the landscape of Twitter mentions. These accounts actively contribute to the propagation of sexist rhetoric, and the tweets in which they are mentioned tend to align more explicitly with harmful and discriminatory viewpoints. This highlights a key point: not all mentions of an account are created equal, and discerning the intent behind a mention (whether it is calling out sexism, or promoting it) becomes a challenge in sentiment analysis. Future work could explore more on this part.

3.2.2. Racist Tweets

@MaxBlumenthal, a controversial American author and blogger, has garnered attention for his outspoken views on various political and social issues. Blumenthal's work and opinions often spark heated debates, especially when touching on sensitive topics like U.S. foreign policy, the Israel-Palestine conflict, and race relations. His mention in racist tweets could arise from criticisms or accusations levelled at him, or from the amplification of his comments by others in the context of racially charged discourse.

This finding suggests that his online persona has become a focal point in conversations where race-related grievances are voiced, amplifying discussions that may involve allegations of racism or the propagation of controversial views on race.

3.3. Wordcloud

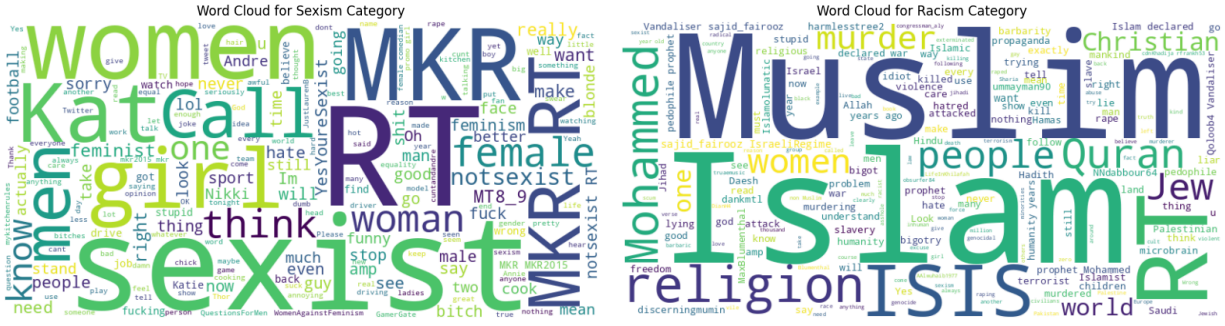

Figure 3.3 – Word Cloud of Sexism and Racism categories

We can see that sexism category are mostly dominated by feminine descriptions. The frequency of the words “women,” “feminism,” and “female” are higher than words like “Men.” My Kitchen Rules (shortened MKR, a cooking show competition) can be seen mentioned a lot of times, this could be a misidentification.

On racism, Muslim has been the most targeted marginalised group with words like “Islam,” “Muslim,” “Quran,” and “ISIS,” being the most prevalent words that appear in this category. Followed by other marginalised groups evidently by the word “Jew,” and “Christian.” Religious groups tend to receive a higher rate of cyberbullying and hate speeches online especially when engaging with a person with a different beliefs and religion. Hate speech towards a specific ethnicity or race are not as significant, that could be explained since religious groups have smaller divisions between sides.

Future work can be done here to further study how can hate speech detection be improve by leveraging on understanding and detecting these high frequency words and turn it into a trigger

4. Discussion

Done with feature extraction and sentiment analysis, we proceed to our final experiments to train models. Model training is a crucial step as it transforms the preprocessed and vectorized text data into predictive outputs. It involves feeding the labelled data (a tweet and its associated sentiment annotations) into machine learning algorithms. 3 training models were deployed which are Logistic Regression, Random Forest, and Support Vector Machine (SVM). These will help to distinguish different sentiment classes like racism, sexism or neutral.

Data that has been through feature extraction will represent the frequency and importance of words in a tweet, enabling the model to learn which words or phrases that can be pinpoint into a specific sentiment.

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

4.1. Logistic Regression

The Logistic Regression model performed well in classifying non-offensive tweets (label 0), showing a precision of 0.82 and a recall of 0.94. This indicates that it could correctly identify most non-offensive content, with only a small proportion of misclassifications. The high recall for non-offensive tweets suggests that Logistic Regression had a low false negative rate in this category, meaning that very few non-offensive tweets were incorrectly classified as offensive.

However, when it came to classifying offensive tweets (label 1), the model struggled more. The recall for offensive tweets was 0.56, meaning that nearly half of the offensive tweets were not identified, and the model produced more false negatives for this category. The F1-score of 0.66 indicates that while the precision was relatively good, the low recall diminished the model's effectiveness in this category. Overall, Logistic Regression provided solid performance for non-offensive content but was limited in its ability to identify offensive tweets.

4.2. Support vector Machine (SVM)

The Random Forest model showed a slightly higher overall accuracy of 83%, with a more significant distinction in precision for offensive tweets (0.87), which was the highest among the models. This suggests that Random Forest was more conservative in labelling tweets as offensive, preferring to avoid false positives. However, the model still had difficulty with offensive tweet recall, which stood at 0.55, comparable to Logistic Regression. This low recall indicates that it was still missing a substantial portion of offensive tweets, resulting in a high false negative rate.

For non-offensive content, Random Forest achieved a recall of 0.96, reflecting its ability to correctly classify most non-offensive tweets. The model’s F1-score of 0.67 for offensive tweets shows that it slightly outperformed Logistic Regression in balancing precision and recall, though it still struggled with detecting all offensive content. The model’s strength lies in its capacity to capture non-linearity in the data, though it still faced limitations in balancing recall and precision for offensive tweets.

4.3. Random Forest (RF)

The SVM model achieved the highest overall accuracy of 84%, performing better than both Logistic Regression and Random Forest. The precision of 0.85 and recall of 0.93 for non-offensive tweets highlights the model's ability to accurately classify the vast majority of non-offensive content, similar to Random Forest. However, SVM excelled in offensive tweet detection, with a recall of 0.66, outperforming the other two models. This indicates that SVM was more effective at identifying offensive tweets, reducing the number of false negatives in this category.

The F1-score of 0.73 for offensive tweets demonstrates that SVM struck a better balance between precision and recall, offering a more balanced performance for both categories. SVM's strength in this project lies in its ability to generalise well and handle cases where the boundary between classes is subtle, making it the most effective model for detecting offensive content without sacrificing precision.

5. Conclusion and Future Work

5.1. Conclusion

In this work, we applied methods that had been established before. Deploying advanced machine learning models like Logistic Regression, Random Forest and Support Vector Machines (SVM). Each model present a very accurate

5.2. Future Work

Regarding our project on sentiment analysis and cyberbullying classification, we identified several directions that could be implemented in the future to improve the precision and efficacy of our models. These include exploring with advanced algorithms, adjusting hyperparameters, and adding contextual information and sentiment lexicons.

Investigating more advanced machine learning methods than the models in use now is one promising way. Deep learning algorithms have demonstrated remarkable performance in a variety of natural language processing tasks. Examples of these algorithms include Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT). These models have the ability to identify minute details and complex trends in the language used in hate speech and cyberbullying. Furthermore, a thorough hyperparameter tuning for the current models—Logistic Regression, Random Forest, and Support Vector Machines (SVM)—may result in additional performance gains. Methods such as grid search or random search can be utilised to determine the best configurations for every model, which could improve the models' accuracy in cyberbullying and sentiment classification.

Improving the features model is another area that needs more attention. Sentiment lexicons designed specifically for hate speech might enhance feature extraction and capture more of the complex elements of language used in cyberbullying. These lexicons might offer more information about the emotional content and tone of the messages, resulting in a more precise classification. Additionally, creating features that take into account tweet context such as user behaviour and tweet timing may result in important insights into the tone and meaning of the messages. With the ability to differentiate between constructive and destructive content, this contextual data may make it easier to identify hate speech and cyberbullying.

6. Reference

1. Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data, King and Suton, The British Journal of Criminology, Volume 56, Issue 2, March 2016, Pages 211–238
https://academic.oup.com/bjc/article/56/2/211/2462519?login=false

2. High Times for Hate Crimes: Explaining the Temporal Clustering of Hate-Motivated Offending, Burnap and Williams, Criminology Volume 51, Issue 2, November 2013 Pages: 871-894
https://www.ojp.gov/ncjrs/virtual-library/abstracts/high-times-hate-crimes-explaining-temporal-clustering-hate

3. Negation Identification and Calculation in Sentiment Analysis, University of Hull, UK & University of Jos, Nigeria, Das and Chen, 2012
https://www.thinkmind.org/articles/immm_2012_1_10_20033.pdf

4. Improving sentiment analysis with multi-task learning of negation, Cambridge University Press, H. Watanabe et al, November 2020
https://www.cambridge.org/core/journals/natural-language-engineering/article/abs/improving-sentiment-analysis-with-multitask-learning-of-negation/14EF2B829EC4B8EC29E7C0C5C77B95B0