Application of ML in Text Analysis: Sentiment Analysis of Twitter Data Using Logistic Regression

Antim Antim
Frankfurt University of Applied Sciences

Application of ML in Text Analysis: Sentiment Analysis of Twitter Data Using Logistic Regression

Antim Antim
Masters in High Integrity Systems
Frankfurt University of Applied Sciences
Frankfurt, Germany
antim.antim@stud.fra-uas.de

***Abstract*—This paper is based on a machine-learning project that uses sentiment analysis and logistic regression to detect racist and sexist content in tweets. It consists of three stages: data preprocessing, feature extraction, and model building and evaluation. The project shows moderate results to detect racist/sexist tweets and aims to reduce harmful content on social media platforms. Using logistic regression on bag of words and tf-idf features the f1 score is obtained 0.5303 and 0. 5451 respectively.The project goals are achieved by different machine-learning python libraries. It also shows a comparative study of results and future work opportunities based on the problem statement.**

***Index Terms*—sentiment analysis, logistic regression, tweeter, text analysis, machine learning, learning from data, bag of words, tf-idf**

I. Introduction

Machine learning (ML) is a subfield of artificial intelligence (AI) that can be used to automatically classify, cluster, or generate text based on patterns and relationships found in data.

Application of machine learning in text analysis is a broad field and there are a lot of applications available. The following are the widely used application of machine learning in text analysis:

1. Sentiment analysis 2. Topic modeling 3. Text classifications 4. Text clustering 5. Named entity recognition 6. Machine translation and much more

In this paper, only sentiment analysis using machine learning algorithms (logistic regression) is presented. The project is available to download on Github. The project is developed in python with the help of various machine-learning libraries. Download the whole project through this Link

This paper presents a detailed approach to a project to detect racist and sexist content in tweets using sentiment analysis using logistic regression. The primary objective is to classify tweets into two categories: racist/sexist and non-racist/non-sexist, thereby helping in identifying and mitigating the spread of harmful content on social media platforms. The project comes in three main stages: data preprocessing, feature extraction, and classification.

In the data preprocessing stage, we clean and preprocess the tweets by removing tweeter handles, numbers, special characters, URLs, and stop words, followed by tokenization and stemming. This step ensures that the data is prepared adequately for further processing.

For feature extraction, two popular techniques are used: Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), to convert the preprocessed textual data into numerical representations. These features are then used to train a logistic regression model, which serves as the classifier for the sentiment analysis task.

To visualize the dataset, the wordcloud visualization tool is employed to display the word density in the dataset. This provides a qualitative view of the most frequent terms associated with racist and sexist tweets.

The approach taken in the project demonstrates moderate results in accurately classifying tweets based on their sentiment, which can help to reduce the prevalence of harmful content on social media platforms. Future work could explore the integration of additional machine learning algorithms and techniques to improve the overall performance of the sentiment analysis model.

II. Related Works

In recent years, the detection of racist and sexist hate speech on social media platforms has become a growing area of research. Several studies have been conducted to develop techniques and models to address this challenge effectively. Successfully detecting racist/sexist tweets will help reduce hate speech and find out individuals accountable for their actions. There has been a lot of research has been done in this field and shows an impactful result. In the following section, several related works on sentiment analysis are presented in brief.

Istaiteh et al. [1] provide a comprehensive literature review of racist and sexist hate speech detection approaches, focusing on three main aspects: available datasets, exploited features, and machine learning models [1]. The authors highlight various methods used in the field, emphasizing the importance of selecting appropriate features and models for effective hate speech detection. This study serves as a foundation for understanding the current state of research in detecting racist and sexist content on social media platforms.

Lee et al. [2] propose a novel approach for racism detection in tweets using a stacked ensemble deep learning model called Gated Convolutional Recurrent-Neural Networks (GCR-NN). The GCR-NN model combines gated recurrent units (GRUs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) to perform sentiment analysis on tweets. The authors report superior performance with an accuracy of 0.98 and a detection rate of 97 percent for tweets containing racist comments. This study demonstrates the potential of deep learning techniques in the domain of hate speech detection.

In addition to the previously discussed works, another relevant study is by Krupalija et al., who propose a new algorithm for calculating a user hate speech index based on user post history [3]. This algorithm is aimed at improving hate speech detection in Twitter posts. The authors preprocess and tokenize the text, remove outliers, and balance classes in the collected dataset. They then use the algorithm to determine the hate speech index of users who posted tweets from the dataset.

Krupalija et al. experimented with multiple machine learning models, including k-means clustering, naïve Bayes, decision trees, and random forests, using four different feature subsets for training and testing [3]. They employ anomaly detection, data transformation, and parameter tuning to enhance classification accuracy. Their results indicate that using the user hate speech index, alone or in combination with other user features, improves the accuracy of hate speech detection. This study demonstrates the potential of incorporating user-specific features to enhance the performance of hate speech detection models.

Another study of interest is by Citation and Abstract, which focuses on Twitter sentiment analysis using supervised machine learning techniques for detecting hate speech [4]. The authors analyze a dataset of 5,000 tweets using Weka software and apply two filters—Tweet to Sparse Feature Vector and Tweet to Lexicon Feature Vector—to improve the model's accuracy. Their research examines various machine learning techniques and finds that the Random Forest method yields the highest accuracy of 93 percent in both cases.

This study emphasizes the importance of selecting appropriate feature extraction techniques and machine learning algorithms to effectively detect hate speech in social media content [4]. Although the authors do not specifically target racist or sexist tweets, their work demonstrates the general potential of supervised machine learning in sentiment analysis for hate speech detection.

In contrast to these studies, our work focuses on using logistic regression for sentiment analysis and classification of tweets as racist/sexist or not. Our approach incorporates Bag of Words and TF-IDF techniques for feature extraction, with word cloud visualization to display word density.

While our study emphasizes logistic regression, it could benefit from incorporating user-specific features and exploring deep learning techniques to further improve the performance of the sentiment While our focus lies on logistic regression, our work could benefit from the exploration of additional supervised machine learning techniques, such as Random Forest, to improve the overall performance of the sentiment analysis model.

III. Methodology

The primary objective of our research is to detect hate speech in tweets by identifying those containing racist or sexist sentiments. In this context, a tweet to contain hate speech if it exhibits racist or sexist undertones. Our task, therefore, involves classifying tweets into two categories: racist/sexist (label '1') and non-racist/non-sexist (label '0').

The trained dataset [5] contains 31,000 tweets where 29.000 tweets are marked as "label 0" (non-racist/sexist) tweets and 2000 tweets are marked as "label 1" (racist/sexist) tweets.

To achieve this, the project involves training a logistic regression model on a given dataset of labeled tweets, with the aim of accurately predicting the labels on a test dataset. The evaluation metric for this task is the F1-Score, which balances precision and recall to assess the performance of our classification model.

In the following sections, steps involved in the project such as methodology, including data preprocessing, feature extraction, and model training and evaluation are presented below.

A. Data Processing

To remove noisy and inconsistent data, preprocessing the dataset is crucial for accurate sentiment analysis. The primary goal of preprocessing is to eliminate elements that contribute little to sentiment identification, such as punctuation, special characters, numbers, and terms with minimal contextual relevance.

The followings are the steps taken for data processing to get a clean and tidy dataset.

The Twitter dataset used has 3 columns 'id', 'label', and 'tweet' respectively. No processing has been done on the id(unique id of the record) and label(label 0 means no racist/sexist and label 1 means racist/sexist tweet) column. The thi

B. Eliminating Twitter handles

In the dataset, the Twitter handles have been anonymized as @user to address privacy concerns. Consequently, these masked handles provide minimal insight into the content or sentiment of the tweet.

C. Excluding punctuation, numerals, and special characters

Exclude punctuation, numbers, and special characters, as this meta information is unlikely to contribute very less or none in terms of sentiment in the tweet text.

D. Discarding brief words

Many shorter words, such as "hmm", "ok", 'his', and 'all', typically offer limited value in sentiment analysis. Therefore, it is advisable to remove them from the dataset to enhance the overall accuracy.

E. Tokenization

Tokenization refers to extracting words known as tokens from a sentence[6]. In this context, tokenization is performed on the processed tweets After completing the aforementioned three steps, each tweet can be divided into individual words or tokens, a fundamental process in any natural language processing (NLP) task. For example “Black life matters” will be tokenized as “black”, “life”, “matters”.

F. Stemming

In the dataset, there might be multiple versions of the same words like love, loves, and loving. These tokens are usually referred to in the same context. To get a tidy dataset only the root variant of the word is considered in the dataset which is “love” in this context. As the word “play” is considered a token for multiple versions of this word like plays, played, playing, etc.[7]

Now we have cleaned and tidied the dataset after the fifth step of operations. In the following section, the tidy dataset is visualized for a clear understanding of the dataset.

IV. Data Visualization

To visualize the clean and tidy dataset, a library called wordcloud is used in the project. Wordcloud library provides a graphical representation in which the most common words are displayed in larger font sizes, while less frequent words are shown in smaller sizes [8].

The following figure shows the wordlcoud visualization of words from whole data including non racist/sexist tweets. Wordcloud can help extract the most important words and tags. After tokenization we have two groups of tokens, the first group contains tokens from the tweet itself and the second group contained only the hashtag of the tweet. Both groups are visualized using the wordcloud library.

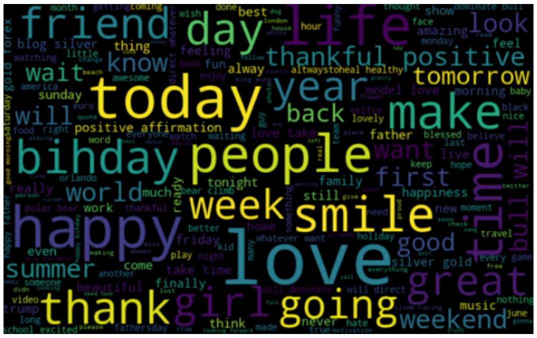


Fig. 1: Wordcloud visualization of whole dataset

A. Wordcloud visualization of words in non racist/sexist tweets

In the dataset tweet that are labeled 0 is non-racist/sexist tweets. After visualizing the words from these groups of tweets using the word cloud the following figure 1 is extracted. The majority of the words observed are positive or neutral in nature, with ”happy,” ”smile,” and ”love” being the most common. As a result, the frequently used words align well with the sentiment of non-racist and non-sexist tweets.

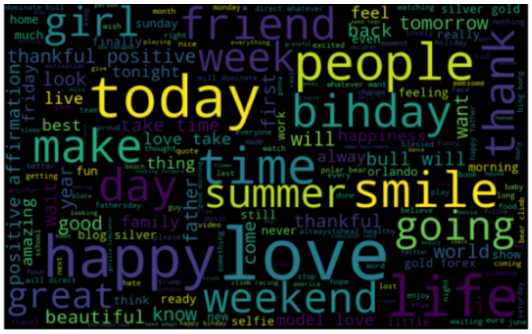


Fig. 2: Words in racist/sexist tweets

B. Wordcloud visualization of words in racist/sexist tweets

In the dataset tweet that are labeled 1 are racist/sexist tweets. After visualizing the words from these groups of tweets using wordlcloud the following figure 2 is extracted. The majority of the words observed are negative or racist in nature, with ”hate,” ”black,” and ”racist” being the most common. As a result, the frequently used words align well with the sentiment of racist and non-sexist tweets.

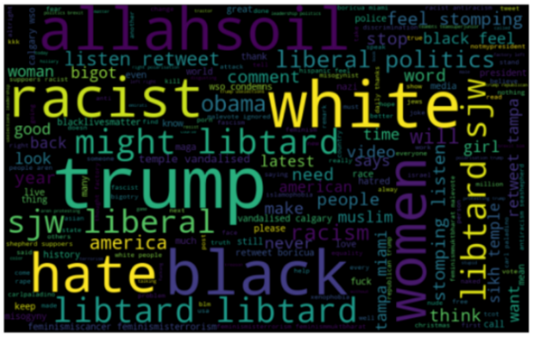


Fig. 3: Words in racist/sexist tweets

C. Wordcloud visualization of hashtags in non-racist/sexist tweets

In the dataset tweets that are labeled 0 are non-racist/sexist tweets. After visualizing the words from these groups of tweets using wordlcloup the following figure 1 is extracted. The

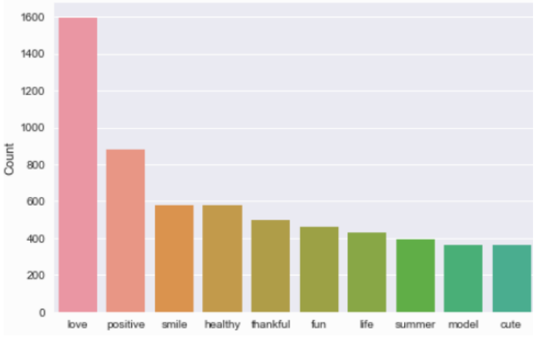


Fig. 4: Hashtags in non racist/sexist tweets

majority of the words observed are positive or neutral in nature, with ”love,” ”positive,” and ”smile” being the most common. As a result, the frequently used words align well with the sentiment of non-racist and non-sexist tweets.

D. Wordcloud visualization of hashtags in racist/sexist tweets

In the dataset tweets that are labeled 1 are racist/sexist tweets. After visualizing the hashtags from these groups of tweets using wordlcloup the following figure 4 is extracted. The majority of the hashtags observed are negative or racist in nature, with ”trump,” ”black,” and ”libtard” being the most common.

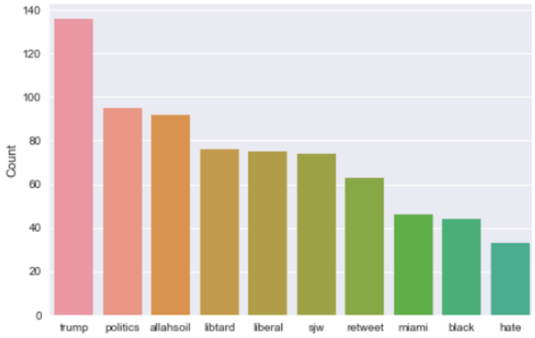


Fig. 5: Hashtags in racist/sexist tweets

As a result, the frequently used hashtags align well with the sentiment of racist and non-sexist tweets. Now, the visualization of the dataset is done.

The visualization shows that the dataset is right in place to extract features from the data and train the model based on the data.

V. Feature Extraction

The dataset is processed for feature extraction. Feature extraction is necessary for model building. This project uses Bag-of-Words and TF-IDF methods for feature extraction from the dataset.

A. Bag of Words Features

In this sentiment analysis project, the Bag of Words (BoW) technique is used for feature extraction in the context of sentiment analysis. The BoW model is a widely-used, simple, yet effective method for transforming textual data into a numerical representation that can be fed into machine learning algorithms. To implement the BoW approach, we utilized the CountVectorizer function from the popular Python library, sklearn. The CountVectorizer function creates a matrix where each row represents a document and each column corresponds to a term. The matrix entries are the frequency counts of terms within each document

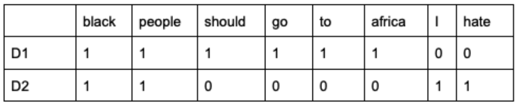
.

D1: Black people should go to Africa
D2: I hate black people

Combining tokens and unique tokens in the corpus are = [”black”, ”people”, ”should”, ”go”, ”to”, ”africa”, ”i”, ”hate”]

The matrix M can be represented now as:
D1: [1, 1, 1, 1, 1, 1, 0, 0]
D2: [1, 1, 0, 0, 0, 0, 1, 1]

To manage the computational complexity and reduce potential noise in our analysis, max features parameter is set to 1000, which restricted the feature set to the top 1000 terms ordered by their term frequency across the entire corpus.



	black	people	should	go	to	africa	I	hate
D1	1	1	1	1	1	1	0	0
D2	1	1	0	0	0	0	1	1

Fig. 6: Bag of Word Feature

This decision enabled us to focus on the most relevant and frequently occurring terms, while simultaneously filtering out less significant terms that might contribute little to our sentiment analysis. In doing so, we effectively leveraged the power of the Bag of Words technique for sentiment analysis, providing valuable insights into the underlying patterns and trends within the textual data.

After implementing the bag of words into the dataset, we get the BoW matrix (49159, 1000).

B. TF-IDF Features

In this sentiment analysis project, to explore the importance of individual words within a given corpus by employing the term frequency-inverse document frequency (TF-IDF) approach. The TF-IDF method allows us to weigh the relevance of terms within a document, accounting for their frequency of occurrence while penalizing terms that are commonly found across multiple documents.

To illustrate the application of this technique, two sample documents are considered, D1: ”Black people should go to Africa” and D2: ”I hate black people”.

Initially, we constructed an array of unique words present in both documents: [”black”, ”people”, ”should”, ”go”, ”to”, ”africa”, ”i”, ”hate”]. Next, we calculated the term frequency (TF) for each unique word in D1 and D2, obtaining the following frequency distributions: [1, 1, 1, 1, 1, 1, 0, 0] for D1, and [1, 1, 0, 0, 0, 0, 1, 1] for D2.

To compute the inverse document frequency (IDF), determined the total number of documents (in this case, 2) and divided it by the number of documents containing each unique word. Lastly, we applied the logarithmic scale to the obtained quotient.

By leveraging the TF-IDF approach, it shows the significance of specific words in the context of sentiment analysis, such as ”hate” and ”should,” which carry strong sentiment connotations.

VI. Model Building

All necessary pre-modeling stages are completed to prepare the data. The project constructs predictive models using two different feature sets: Bag-of-Words and TF-IDF[09].

The feature is already extracted and BoW and TF-IDF matrix is already generated. Implementing logistic regression on these extracted features will train the model and will get f1 score.

Logistic regression will be employed to develop these models, as it is a statistical method that estimates the probability of an event occurring by fitting the data to a logit function.

Logistic regression is a supervised machine learning algorithm that predicts a binary outcome (e.g., true/false, yes/no) based on one or more input features [10]

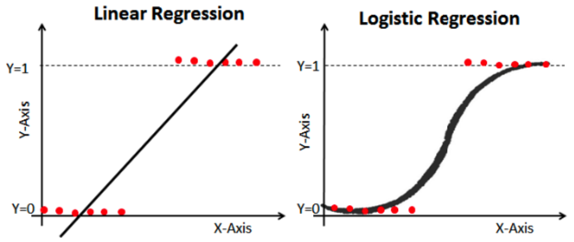


Fig. 7: Linear vs Logistic Regression

In model building, logistic regression is chosen over linear regression logistic because regression is suitable for predicting binary outcomes, while linear regression is appropriate for predicting continuous values.

The following equation is used in logistic regression to predict binary outcomes.

$$\log\left(\frac{p}{1-p}\right) = \beta_o + \beta(Age)$$

A. *Building a model using Bag-of-Words features*

After preprocessing the data and extracting relevant features using the Bag of Words, The project trained the logistic regression model on the dataset. To implement logistic regression as our machine learning algorithm using the sklearn library.

Upon evaluating the model's performance, we obtained an F1 score of 0.5303. The F1 score is a widely-used metric for binary classification problems, as it represents the harmonic mean of precision and recall, thus capturing a balanced view of the model's performance.

The f1 score of 0.5303 shows the moderate performance of the model, with 1 being an excellent score indicating perfect precision and recall, while 0 signifies poor performance with either no true positives or a complete lack of both precision and recall.

B. *Building a model using TF-IDF features*

After preprocessing the data and utilizing the TF-IDF technique for feature extraction, we trained the logistic regression model on the dataset. Upon evaluating the model's performance, we obtained an F1 score of 0.5451.

The F1 score is a valuable metric for binary classification problems as it represents the harmonic mean of precision and recall, thus providing a balanced view of the model's performance.

The f1 score of 0.5451 shows moderate accuracy of the model. But it's slightly improved compared to the bag of words feature f1 score.

VII. Result and Discussion

In our sentiment analysis study, the performance of two feature extraction techniques, Bag of Words (BoW) and TF-IDF is investigated, in conjunction with logistic regression.

BoW is a simple method that focuses on term frequency, while the TF-IDF technique emphasizes the importance of specific words within documents and downplays the influence of commonly occurring terms.

Upon evaluating the models' performances, we found that the BoW-based model achieved an F1 score of 0.5303, while the TF-IDF-based model yielded a slightly better F1 score of 0.5451.

These results indicate that the TF-IDF-based model shows slightly better performances compared to the BoW-based model in terms of classification accuracy.

VIII. Conclusion and Future Scope

In conclusion, this project-based research investigated how sentiment analysis can use machine learning techniques. Along with logistic regression, the effectiveness of the two feature extraction techniques Bag of Words (BoW) and TF-IDF is examined.

The outcomes showed that the sentiment classification model's overall performance is greatly influenced by the feature extraction method used, with the TF-IDF-based model outperforming the BoW-based mode. This emphasizes how crucial it is to choose the right feature extraction methods meticulously when carrying out sentiment analysis tasks.

There are many potential directions for future research and development. Investigating different classification algorithms, such as support vector machines, random forests, or neural networks, to see how well they perform in sentiment analysis tasks and determine which approach is best for this specific issue is one possible direction. In order to record the semantic relationships between words and possibly increase classification accuracy, more advanced feature extraction techniques like word embeddings (such as Word2Vec or GloVe) could be investigated.

Another area of future research is the incorporation of advanced natural languages processing techniques, such as sentiment lexicons or deep learning-based models like transformers, to better understand and capture the nuances within textual data.

Finally, it would be advantageous to assess the model's performance on bigger and more varied datasets, covering a wider variety of sentiment expressions and domains, to increase the generalizability of the model. We intend to advance the functionality of sentiment analysis models and contribute to current developments in the area of natural language processing by investigating these potential future research paths.

IX. References

[1] O. Istaiteh, R. Al-Omoush and S. Tedmori, ”Racist and Sexist Hate Speech Detection: Literature Review,” 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA), Valencia, Spain, 2020, pp. 95-99, doi: 10.1109/IDSTA50958.2020.9264052.

[2] E. Lee, F. Rustam, P. B. Washington, F. E. Barakaz, W. Aljedaani and I. Ashraf, ”Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model,” in IEEE Access, vol. 10, pp. 9717-9728, 2022, doi: 10.1109/ACCESS.2022.3144266.

[3] E. Krupalija, D. Donko and H. supic, ”Usage of user hate speech index for improving hate speech detection in Twitter posts,” 2022 XXVIII International Conference on Information, Communication and Automation Technologies (ICAT), Sarajevo, Bosnia and Herzegovina, 2022, pp. 1-6, doi: 10.1109/ICAT54566.2022.9811159.

[4] M. Dagar, A. Kajal and P. Bhatia, ”Twitter Sentiment Analysis using Supervised Machine Learning Techniques,” 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-7, doi: 10.1109/ISCON52037.2021.9702333.

[5] https://www.kaggle.com/datasets/dv1453/twitter-sentiment-analysis-analytics-vidya Last accessed 2nd February 2023.

[6] N. K. Bolbol and A. Y. Maghari, ”Sentiment Analysis of Arabic Tweets Using Supervised Machine Learning,” 2020 International Conference on Promising Electronic Technologies (ICPET), Jerusalem, Palestine, 2020, pp. 89-93, doi: 10.1109/ICPET51420.2020.00025.

[7] https://www.projectpro.io/article/stemming-in-nlp/780 Last Accessed: 8th February 2023.

[8] https://www.pluralsight.com/guides/natural-language-processing-visualizing-text-data-using-word-cloud Last accessed: 8th February 2023.

[9] K. Bhargava and R. Katarya, ”An improved lexicon using logistic regression for sentiment analysis,” 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), Gurgaon, India, 2017, pp. 332-337, doi: 10.1109/IC3TSN.2017.8284501.

[10] https://www.analyticsvidhya.com/blog/2018/07/hands-on-sentiment-analysis-dataset-python/ Last accessed: 23rd February , 2023.