RESEARCH ARTICLE OPEN ACCESS

Exploring Twitter Sentiment Analysis: A Comparative Study of Machine Learning Models

Tanuja Pramod Gotekar*

*(UG student, Department of Information Technology, B. K. Birla College, Kalyan

Email: gotekartanuja9137@gmail.com)

Abstract:

Social media is afflicted with sexism, creating difficulties for feminist and female users who navigate through these platforms. However, the current natural language processing techniques used to categorize sexism are insufficient. This paper suggests a more detailed way to categorize online harassment on platforms like Twitter. Along with sharing our classification findings, and also talk about the difficulties in labeling these categories on training and testing datasets. The Twitter dataset used in this research paper, which was collected from Kaggle, first distinguishes phases by discussing different levels of NLP and components of Natural Language Understanding and Natural Language Generation. The work is applied by preparatory machine learning to know the concept of sexism and differentiate itself by looking at more precise categories of sexism in social media.This research provides valuable insights into sentiment analysis in the realm of social media data, highlighting the importance of robust tools for tackling these tasks.

Keywords — Natural Language Processing, Natural Language Understanding, Natural Language Generation, Twitter, Naïve Bayes, LSTM, XGBoost, LightGBM

---------------------------------------*************************--------------------------------

I. INTRODUCTION

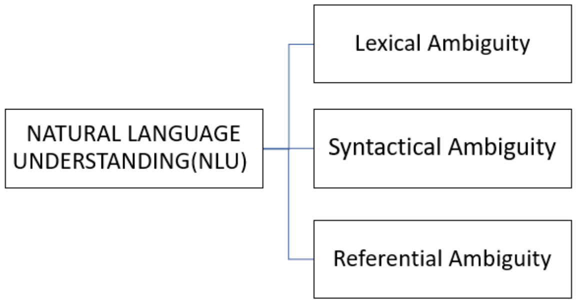
Natural Language Processing (NLP) refers to a branch of computer science and artificial intelligence concerned with the interaction between computers and humans in natural language. It involves the development of models and algorithms that permit computers to understand, interpret, and generate a human language in a way meaningful and useful. The ultimate aim of NLP is to help computers understand language as well as do. It is the driving force behind things like virtual assistance, speech recognition, sentiment analysis, automatic text summarization, machine translation, and plentiful. Sentiment analysis also known as opinion mining is one of the many applications of Natural Language Processing. It is a set of methods and techniques used to extract subjective information from text or speech, such as opinions or attitudes. In simple terms, it involves classifying a text as positive, negative, or neutral. On online platforms there is gender-based violence is a gender rights tragedy. The social media affects different genders differently, whether it's through fostering toxic masculinity or transphobia. It is inherently discriminatory written or spoken language that implies an unjustified sexual bias against a group or an individual, usually women.

Natural language processing, also referred to as text analytics, plays a key role in today's era because of the abrupt volume of text data that users generate around the world on digital channels such as e-commerce websites, blog posts, social media apps, etc. The work of natural language processing is conducted by computer scientists while other professionals have also had interest such as linguistics, phycologists, philosophers, etc. One of the most important aspects of NLP is that it adds to the knowledge of human language. The NLP field is related to the different theories and techniques that deal with the problem of the natural language of communicating with computers.

A. *COMPONENTS OF NLP*

NLP is classified into two parts Natural Language Understanding and Natural Language Generation which emerge the task to understand and generate a text. The above diagram shows the broad classification of NLP. The main purpose of this section is to discuss the Natural Language (Linguistic) (NLU) and the Natural Language Generation (NLG).

NLU: NLU understands the human language and converts it into data. Natural Language Understanding is a subset of NLP. It is more oriented towards the reading aspect of NLP. Examples could be Profanity detection, Sentiment analysis, and Classification of Topics (Like, what's this tweet all about?). It enables machines to understand natural language and analyze it by extracting concepts, entities, emotions keywords, etc.



Lexical Ambiguity: Lexical ambiguity refers to a phenomenon in a language where a word or phrase has multiple meanings or interpretations. The word or phrase pertains to having more than one meaning in the language in which the word can be understood in different ways, leading to confusion or miscommunication. When homonyms can occur in the same position in an assertion, the result is lexical ambiguity.Example Give me the bat!

Thebat can be illuminated in two ways: -A baseball bat or a cricket bat (used for sports)
-A flying nocturnal animal.

Without context, it isn't clear even if the speaker is referring to the first or second of these meanings. This can be a lexical ambiguity.

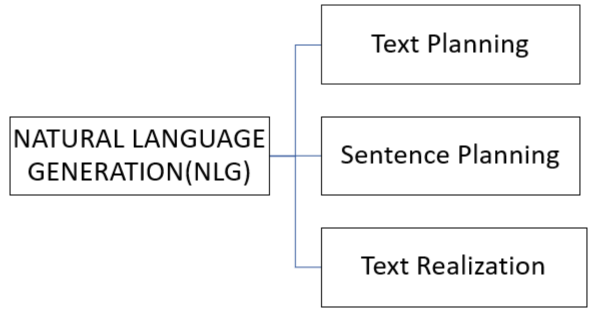
Syntactical Ambiguity:Syntactic ambiguity is the potential for multiple interpretations of a piece of written or spoken language because of the way words or phrases are structured.Example
- Canberra is the capital of Australia.
- Is Canberra Australia's capital?

Both sentences have the same set of words, but only the first one issystematicallycorrect, and comprehensible. Basic lexical processing techniques are required to understand the relationship between individual words in these sentences.

Referential Ambiguity: Referential ambiguity occurs when a word or phrase, in the context of a specific sentence, could refer to two or more properties or things. It is something clear from the context that meaning is deliberate, but not always.

Example – The boy told his father about the theft.He was referentially ambiguous because it can comment on both the boy and the father.

NLG: NLG stands for Natural Language Generation it uses structured data and generates meaningful narratives out of it. This is intelligent. Computers write here, meaningfully, in fact, yes, they generate. Structured data-to-text conversion.



Text Planning: In the text planning general content is formulated and ordered logically. It includes retrieving the pertinent content from the knowledge base. Text Generation is the task of generating text with the main goal of appearing indistinguishable from human-written text. Text generation can be consigned with Markov processes or deep generative models like LSTMs.

Sentence Planning: Sentence Planning considers punctuation and text flow, breaking out the content into paragraphs and sentences and incorporating pronouns or conjunctions where appropriate. It includes choosing required words, forming meaningful phrases, and setting the tone of the sentence.

Text Realization: In text realization, it is mapping sentence plan into sentence structure. Creating a text which should be correct according to the rules of syntax, morphology, and orthography. The realizer uses various types of techniques like grammar rules, templates, and language models to generate a natural language text that is coherent, grammatically correct, and semantically appropriate. It is used in various applications such as text generation, automated writing, chatbots, and many more.

II. LITERATURE REVIEW

The main purpose of NLP includes simplifying, analysing, and manipulating natural language data using various algorithms, models, and methods. However, there are many challenges involved in human language which may depend upon the natural language data. Therefore, there are different models and methods in the field of NLP. The sexism classification task was first introduced by Waseem, 2016[1]. He elucidates sixteen thousand tweets and sorts them as racist, sexist, and neither. His dataset is shared in GitHub with the tweet IDs. On his dataset, Waseem,2016[1] created a hashtag helpful but after an adjacent look, we saw that the corpus had a slant and did not contain different types of sexism. Waseem, 2016[1] tried different methods such as character level grams and word grams, and employed logistic regression with 10-fold model-free

In 2017, Akshita and Radhika [2] introduced benevolent sexism in a new category. They possess one of the three qualities features: "protective paternalism", “complementary gender differentiation,” and “heterosexual intimacy. Akshita and Radhika [2] used different types of classifiers and reported the results along with the other polarity. Alessandro et al. [3] focus on automatically categorizing vulgar and offensive speech on social media, especially Facebook. The study uses a computational approach to process taboo content found in 31,749 Facebook comments. The primary annotation and classification are lexicon-based, aiming to effectively detect and categorize the taboo expressions in the comments.

III. METHODS

Data collection

In our research, the Twitter datasetwas collected from Kaggle, and on that dataset, there was a wide range of used tweets and hashtags. In the dataset, many tweets show the feeling of ignorance or lack of being sexy or ignorant. In the dataset, an assorted range of tweets exhibit varying sentiment orientations and circumscribe positive, negative, and neutral sentiments. The training dataset consists of three columns: “ID”, “Label”, and “Tweet”, while the test dataset consists of two columns, “ID”, and “Tweet”. There are many hashtags in the dataset including:
“#motivation” “#friday!” “#tgif” “#media” “#fuck” “#love” “#model”

A. Naïve Bayes

In this study for sentiment analysis on Twitter, the dataset involved a systematic approach that desegregated the Naïve Bayes algorithm. The Naïve Bayes algorithm is a well-established technique in natural language processing and text classification. Sentiment analysis on the Twitter dataset was performed using the Naïve Bayes algorithm. The Natural Language Toolkit is the leading platform for building a Python program to work with human language data. It contains many libraries for tokenization, parsing, classification, stemming, tagging, semantics, etc. There is a seaborn library for visualization purposes. Our train dataset has 3 columns and 31962 rows, and our test dataset has 2 columns and 17197 rows. The number of values labeled ‘0’ contains 29720, and ‘1’ contains 2242 values of count. Then combine the train and test datasets into one single dataset and combine the dataset with the 49159 rows and 3 columns. "Gensim," which means generate similarly, is designed to process raw, unstructured digital texts using an unsupervised machine learning algorithm. It is a popular open-source natural language processing library. Using gensim, analyze the plain text or semantic structure.The bag of words model is used to preprocess text by converting it into a bag of words which keeps a count of the total occurrence of the most frequently used words. Term Frequency (TF) using term frequency we find the ratio of the number of times that word appears in the document to the total number of words. We use Multinomial Naïve Bayes classifier object from sklearn naïve bayes. Lastly, found the accuracy using the naïve Bayes model is 0.94 % accurate.

IV. RESULTS

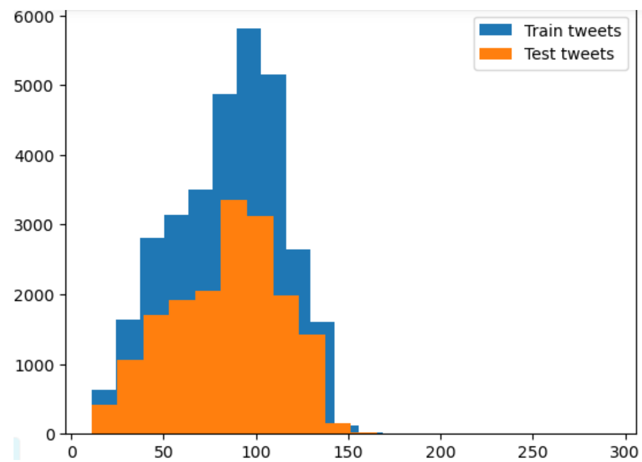
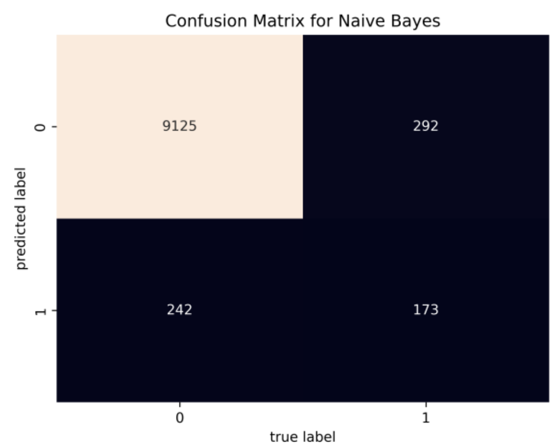


Fig.[1a]



Confusion Matrix for Naive Bayes
Fig.[1b]

B. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a type of recurrent Neural network (RNN) architecture is commonlyusedfor sequentialanalysis. Applying LSTM to our Twitter dataset to d o a task Data Preparation: first, we need to preprocess a Twitter dataset. This includestasks likedata cleaning, tokenization, and converting data into numerical form. Word embeddings represent words in a numerical format. Theembed ding captures semantic relations between words, which can be crucial for understanding sentiment analysis. using one-hot encoding to find categorical variables in models that required numeral input. Fig.[2a]

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 20, 40)            200000

 lstm (LSTM)                 (None, 100)               56400

 dense (Dense)               (None, 1)                 101

=================================================================
Total params: 256,501
Trainable params: 256,501
Non-trainable params: 0
```

Model Sequential

Fig.[2a]

The Twitter dataset is split into training and testing which 33% of data is tested and 67% is trained. After finally Training dataset epochs 10,Accuracyis 99% with a loss of 28% val_loss 14% val_acurracy 93%.Finally, using LSTM theaccuracy score is 92%

	precision	recall	f1-score	support
0	0.93	1.00	0.96	9806
1	0.00	0.00	0.00	742
accuracy			0.93	10548
macro avg	0.46	0.50	0.48	10548
weighted avg	0.86	0.93	0.90	10548

Classification Report Using LSTM

Fig[2b]

C. XGBoost and LightGBM

In the field of Natural Language Processing (NLP), text classification stands as a fundamental task with a wide variety of practical applications. The two machine learning algorithms used for text classification tasks are XGBoost and LightGBM. First, import the necessary libraries, including XGBoost (Extreme Gradient Boosting), LightGBM (Light Gradient Boosting), and sci-kit-learn, and load a dataset splitting it into text and label components. The two popular machine learning models, XGBoost and LightGBM, are trained and evaluated for text classification, providing insights into their accuracy and performance metrics like precision, recall, and F1-score.fig [3a1 and 3a2]

RandomizedSearchCV, using hyperparametertuning, found the best parameters for both the XGBoost and LightGB M models. Best Hyperparameters for XGBoost: {'n_estimatos': 300,'max_depth': 5, 'learning rate': 0.1} and Best Hyperpara meters for LightGBM: {'n_estimators': 300,'max_depth': 5, 'learning_rate': 0.1}also include visualizationand demonstration. sentiment distribution, and a data processing step to cleantext by removing special characters and emoji. Then create a bar chart to countthe prediction for each sentime nt analysis class define the labels as 'Negative', 'Neutral', and 'Positive' as shown in Fig.[3d]

The VADER sentiment analysis tool from NLTK is used to classify a list of sample texts into categories such as humour, ridicule, or irony. It determines the sentiment analysis, defines a function to classify text-based sentiment scores, and then provides the corresponding reaction category for each text fig.[3c]. Logistic Regression is the machine learning algorithm used to perform text classification to detect sexist content in a label dataset. The dataset consists of tweets, where 'label' is set 1 for sexist tweets and 0 for non-sexist ones. Then start by loading and preparing the data. Finally, evaluate, the model's performance using accuracy and generate a classification report, which finds detailed metrics like precision, recall, and f1-score for both classes (sexist and non-sexist) in fig.[3d1 and3d2]

V. RESULTS

Model: LightGBM
Accuracy: 0.9507273580478649
Classification Report:

	precision	recall	f1-score	support
0	0.96	0.99	0.97	5937
1	0.81	0.40	0.54	456
accuracy			0.95	6393
macro avg	0.88	0.70	0.76	6393
weighted avg	0.95	0.95	0.94	6393

Classification Report Using LightGBM
Fig. [3a1]

Model: XGBoost
Accuracy: 0.9507273580478649
Classification Report:

	precision	recall	f1-score	support
0	0.96	0.99	0.97	5937
1	0.81	0.40	0.54	456
accuracy			0.95	6393
macro avg	0.88	0.70	0.76	6393
weighted avg	0.95	0.95	0.94	6393

Classification Report Using XGBoost

Fig. [3a2]

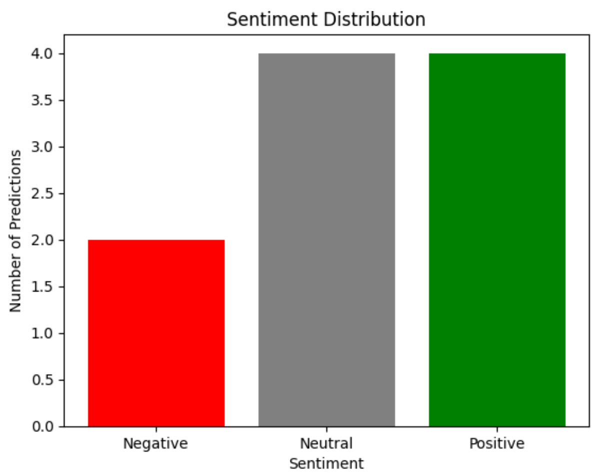


Fig.[3b]

Text: we won!!! love the land!!! #allin #cavs #champions #cleveland #clevelandcavaliers â□¦', 'Another comment with emojis 😄 👍', 'This is a normal comment
Reaction: Positive (e.g., humour)

Text: @user #cnn calls #michigan middle school 'build the wall' chant '' #tcot
Reaction: Neutral

Text: 50 people went to nightclub to have a good night and 1 man's actions means those people are lost to their families forever #rip#orlando
Reaction: Neutral

Text: woohoo!! just over 5 weeks to go!
Reaction: Positive (e.g., humour)

Text: going to la tomorrow!!!
Reaction: Neutral

Text: if you hold open a door for a woman because she's a woman and not because it's a nice thing to do, that's. don't even try to deny it
Reaction: Negative (e.g., ridicule or irony)

Fig.[3c]

```
Accuracy: 0.9493195682778038
Classification Report:
              precision    recall  f1-score   support

           0       0.95      1.00      0.97      5937
           1       0.92      0.32      0.47       456

    accuracy                           0.95      6393
   macro avg       0.93      0.66      0.72      6393
weighted avg       0.95      0.95      0.94      6393
```

Sexist Classification Report
Fig.[3d1]

```
Accuracy: 0.9530736743312999
Classification Report:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98      5937
           1       0.89      0.39      0.54       456

    accuracy                           0.95      6393
   macro avg       0.92      0.69      0.76      6393
weighted avg       0.95      0.95      0.94      6393
```

Non-Sexist Classification Report
Fig.[3d2]

ACCURACY TABLE

1	Naïve Bayes	94%
2	LSTM	92%
3	XGBoost	95%
4	LightGBM	95%

Table [1]

This research paper presents a detailed comparative analysis using four machine algorithms: Naïve Bayes, LSTM, XGBoost, and LightGBM, in the context of a classification task. This study aimed to identify the best-performing models in terms of accuracy, and it was found that XGBoost and LightGBM consistently outperformed the other algorithms shown in above table [1]

VI. CONCLUSION

In this research, sentiment analysis on the Twitter dataset was conducted using various machine language algorithms, including Naïve Bayes, XGBoost, and LightGBM. The Naïve Bayes algorithm was initially applied to achieve an accuracy of 94%. Then, after careful data preprocessing and training, LSTM was used, and the accuracy was 92%. Finally, the text classification task is extended to XGBoost and LightGBM, which were fine-tuned using RandomizedSearchCV to identify the best hyperparameters. It was observed that XGBoost and LightGBM consistently outperformed the other models with high accuracy. As a result of their consistent performance and accuracy in identifying feelings, XGBoost and LightGBM are suggested to be the two best selections for tasks involving sentiment analysis of Twitter data.

VII. ACKNOWLEDGEMENT

I would like to express my profound gratitude to Professor Swapna Augustine Nikale. I also appreciate B.K. Birla College Arts, Science, and Commerce (Autonomous), Kalyan, for giving us the chance to complete a research paper as part of our academic achievement.

VIII. CHALLENGES OF STUDY

The XGBoost and LightGBM have demonstrated potential through sentiment analysis on Twitter. Future challenges in NLP will focus on improving model adaptability, addressing data imbalances, handling multilingual and cross-cultural contexts, recognizing emotions, real-time analysis, ethical considerations, and privacy concerns to advance the field further.

IX. REFERENCES

1. Sharifirad S, Matwin S, (2019) When a tweet is actually sexist. A more comprehensive classification of different online harassment categories and the challenges in NLP. arXiv preprint arXiv:1902.10584

2. Akshita., J., Radhika, M.: When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science (2017) 7–16

3. Simonetta., M.A.P.S.V., Pierluigi, V.: Mining offensive language on social media. http://ceurws.org/Vol-2006/paper038.pdf.

4. Waseem, Z., Hovy, D.: Hateful symbols or hateful people: Predictive features for hate speech detection on twitter. In Proceedings of NAACL-HLT (2016) 88–93

5. Baclic O, Tunis M, Young K, Doan C, Swerdfeger H, Schonfeld J (2020) Artificial intelligence in public health: challenges and opportunities for public health made possible by advances in natural language processing. Can Commun Dis Rep 46(6):161

