CLARK
UNIVERSITY

CHALLENGE CONVENTION.
CHANGE OUR WORLD.

Twitter Sentiment Analysis

Member name-

Rishav Mondal

Aaryan Pankaj Dabhade

Sangram Dedge

Anuja Bavkar

Prof name-

Yue Gao

Abstract

Sentiment analysis, a rapidly developing area, holds great promise for better understanding human-computer interactions, especially when it comes to interpreting textual data to determine emotional responses. This work investigates the effectiveness of a classification model in sentiment analysis using a dataset consisting of text from various sources, such as product reviews and social media posts. Using cutting-edge machine-learning techniques to categorize feelings as positive, negative, or neutral is the foundation of our methodology. According to the investigation, the classification model obtains a high degree of accuracy, which greatly improves our comprehension of user feelings on various digital platforms. As such, this study emphasizes how useful the classification model can be for gaining a practical understanding of customer behavior and interaction patterns.

Improving sentiment analysis tools' accuracy and efficacy is the main driving force behind this work, as these techniques are essential for deciphering the enormous and ever-increasing amounts of text data produced by diverse digital platforms. Businesses, governments, and researchers need to be able to reliably assess sentiments encoded in text as interactions and expressions on social media, consumer reviews, and online forums grow tremendously. This work aims to enhance the ability to identify and analyze sentiments by examining the effectiveness of a classification model in sentiment analysis. This project aims to improve public opinion and emotion analysis by developing sentiment analysis tools. This will help with better-informed decision-making in fields like marketing, CRM, and policy formation.

We used 1.6 million tweets using Twitter API. The tweets are annotated by 0,2 and 4 which represent negative, neutral, and positive respectively. Having 6 variables named as ids, dates, flags, users, text, and sentiments.
From the Decision tree, we got an accuracy of 66.31%, random forest gave us an accuracy of 66.32%, KNN and SVM gave us 62.71 & 66 percentage of accuracy respectively and Gradient boosting gave us an accuracy of 66.61%.

The research verified the efficacy of sophisticated classification models, specifically Gradient Boosting, which demonstrated the highest accuracy in sentiment analysis with a score of 66.61%. By improving the comprehension of public opinion, this advancement is essential for improving decision-making in marketing, customer relationship management, and policy formation.

Introduction

Motivation of the paper in detail

Recognizing the emotional undertone of textual information across platforms can provide important insights into public opinion and consumer behavior in an increasingly digital world. The need for sophisticated tools to efficiently evaluate, understand, and convert the enormous volumes of textual data produced every day into useful sentiment indicators is what spurred this work.

The specific problem under study

Accurately classifying textual expressions of sentiments into good, negative, and neutral emotions is the main issue this work attempts to solve. The subtleties and complexity of human language make it difficult to detect sentiment with high accuracy, even with the abundance of analytical techniques available.

Why studying the problem is important

Given its immediate uses in public relations, marketing, customer service, and social media monitoring, studying sentiment analysis is essential. Organizations can better satisfy client wants and react to market developments by customizing their strategy, offerings, and services based on an understanding of attitudes.

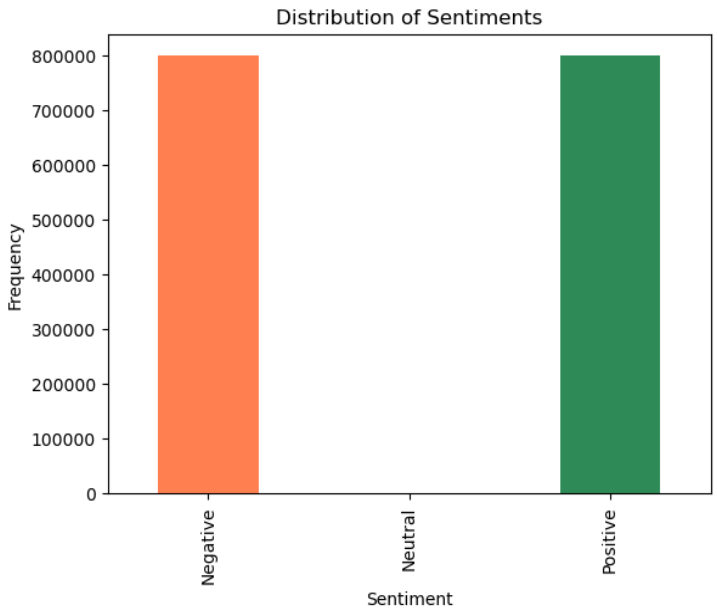
Research questions

1. To what extent is the sentiment of a given text predicted by the provided classification model?
2. What are the main determinants of the model's sentiment analysis performance?
3. In what ways does the model account for the intricacies and nuances of language used by people in various textual contexts?

Method and Analysis

Data description and source

This is the **sentiment140** dataset. It contains 1,600,000 tweets extracted using the twitter API. The tweets have been annotated (0 = negative, 2 = neutral, 4 = positive) and they can be used to detect sentiment. We got the dataset from the website Kaggle.



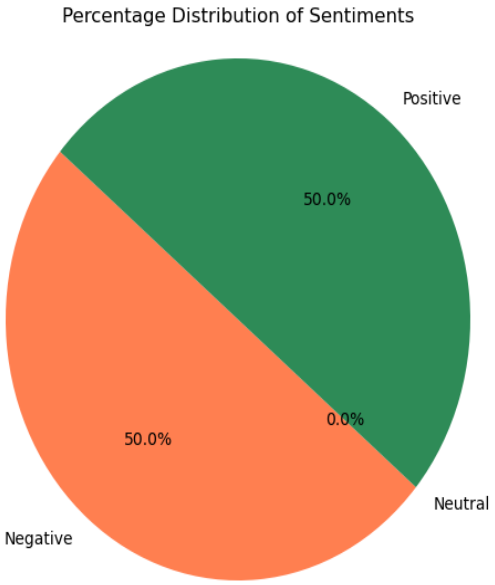


Fig: These figures

We can see from the plots, that this dataset does not have any neutral tweets, the reason for that is every tweet made always had either one negative or positive word in the tweets that made the sentiment of the tweet either negative or positive.

Variable description

The dataset contains 6 variables. They are:

1. **sentiment:** the polarity of the tweet
2. **ids:** the ID of the tweet
3. **date:** the date of the tweet
4. **flag:** the query of the tweet, if there is no query then it will have NO_QUERY
5. **user:** the user that tweeted
6. **text:** the text of the tweet

#	Column	Non-Null Count	Dtype
1	Sentiment	1600000	int
2	ids	1600000	int
3	date	1600000	string
4	flag	1600000	string
5	user	1600000	string
6	text	1600000	string

Data Cleaning

For this analysis we only used the **sentiment** and **text** variables hence we removed the other unnecessary variables from the dataset which included the **ids**, **date**, **flag** and **user** variables. These variables had no impact on the sentiment analysis.

	sentiment	text
0	0	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	0	is upset that he can't update his Facebook by ...
2	0	@Kenichan I dived many times for the ball. Man...
3	0	my whole body feels itchy and like its on fire
4	0	@nationwideclass no, it's not behaving at all....
...	...	...
1599995	4	Just woke up. Having no school is the best fee...
1599996	4	TheWDB.com - Very cool to hear old Walt interv...
1599997	4	Are you ready for your MoJo Makeover? Ask me f...
1599998	4	Happy 38th Birthday to my boo of alll time!!! ...
1599999	4	happy #charitytuesday @theNSPCC @SparksCharity...

1600000 rows × 2 columns

	sentiment	text
0	0	awww bummer shoulda got david carr third day
1	0	upset update facebook texting might cry result...
2	0	dived many times ball managed save 50 rest go ...
3	0	whole body feels itchy like fire
4	0	behaving mad see
...	...	...
1599995	4	woke school best feeling ever
1599996	4	thewdb com cool hear old walt interviews
1599997	4	ready mojo makeover ask details
1599998	4	happy 38th birthday boo alll time tupac amaru ...
1599999	4	happy charitytuesday thenspcc sparkscharity sp...

1600000 rows × 2 columns

Fig 2: Left- cleaned from unnecessary variables. Right-cleaned text from special characters

After that we cleaned the **text** variable further by removing unnecessary and special characters like '#', '@' and so on, and tokenized the text.

Data Preparation

This data was automatically created using the mechanism of distant supervision [1]. It was collected automatically using the twitter api. After that we cleaned the dataset and then performed some feature processing to extract various other features. The features we extracted are:

1. **word_count -** number of words in the text.
2. **char_count -** number of characters in text.
3. **emoticons_count -** number of emoticons in text
4. **exclam_question_marks_count -** number of exclamations in text
5. **capitalized_words_count -** number of capitalized word counts
6. **punctuation_count -** number of punctuations
7. **positive_words_count -** number of positive words in the text
8. **negative_words_count -** number of negative words in the text

To extract the number of positive and negative words we had to use the **SentimentIntensityAnalyzer** [2] from the **NLTK**(Natural Language Tool Kit [3]). This function takes a text as an input and returns all the negative and positive according to **polarityScore** of the words present in the text.

	sentiment	word_count	char_count	avg_word_length	emoticons_count	exclam_question_marks_count	capitalized_words_count	punctuation_count	positive_words_count	negative_words_count
0	0	8	44	4.625000	0	0	0	0	0	1
1	0	11	69	5.363636	0	0	0	0	0	3
2	0	10	52	4.300000	0	0	0	0	1	0
3	0	6	32	4.500000	0	0	0	0	1	2
4	0	3	16	4.666667	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...
1599995	4	5	29	5.000000	0	0	0	0	2	0
1599996	4	7	40	4.857143	0	0	0	0	1	0
1599997	4	5	31	5.400000	0	0	0	0	1	0
1599998	4	9	52	4.888889	0	0	0	0	1	0
1599999	4	5	57	10.600000	0	0	0	0	1	0

1600000 rows × 10 columns

Fig. 3: This figure shows all the variables in the dataset.

Handling Missing Values

Since the data was automatically generated using the twitter API, there are no missing values.

Results

Exploratory results

With 1.6M data at hand we plotted the distribution of the positive and negative words across all the texts in the dataset.

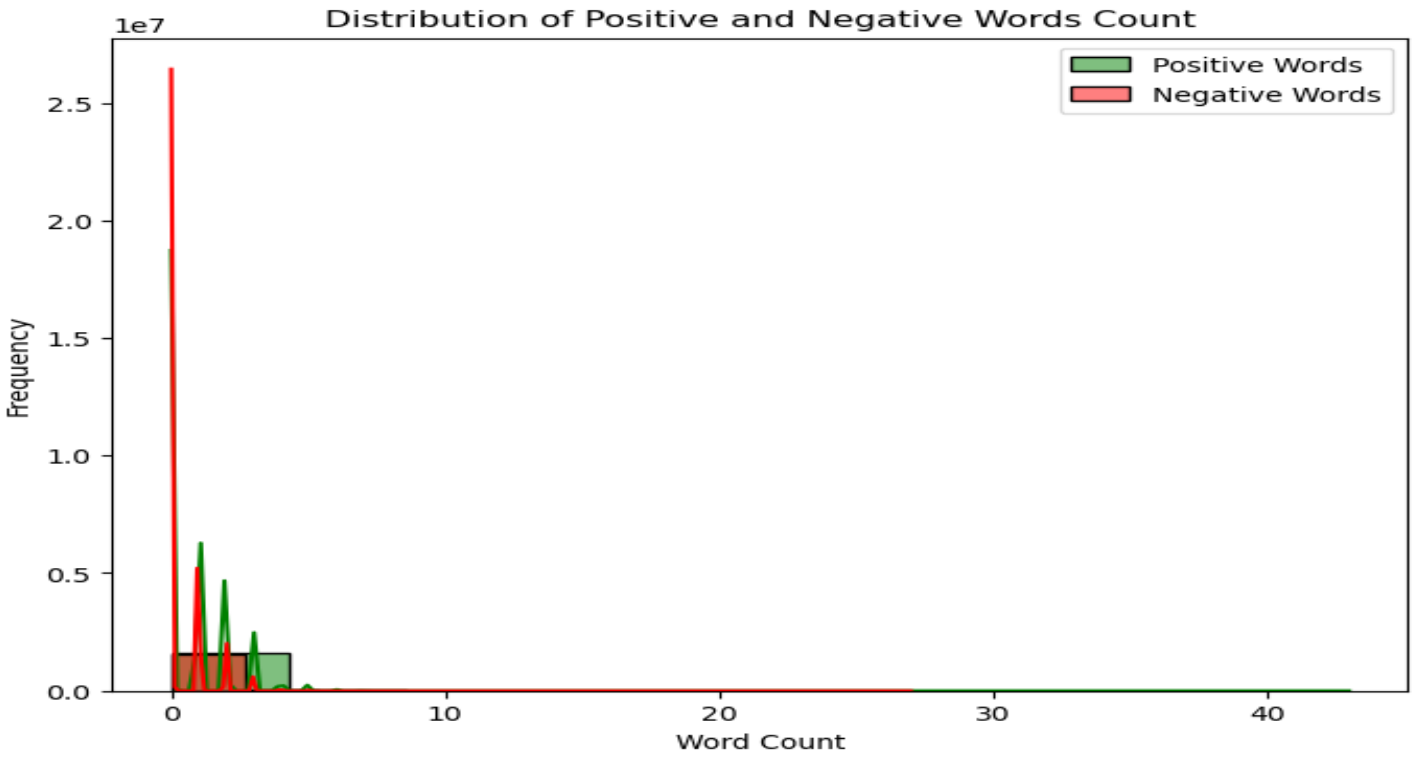


Fig.4: Distribution of positive words and negative words

We found out that about 250K tweets had one negative word and about 200k had positive words in the dataset.

This plot also shows that people are generally making more negative tweets than positive ones. After that we plotted the distribution of the average length of the words present in the text of the tweets.

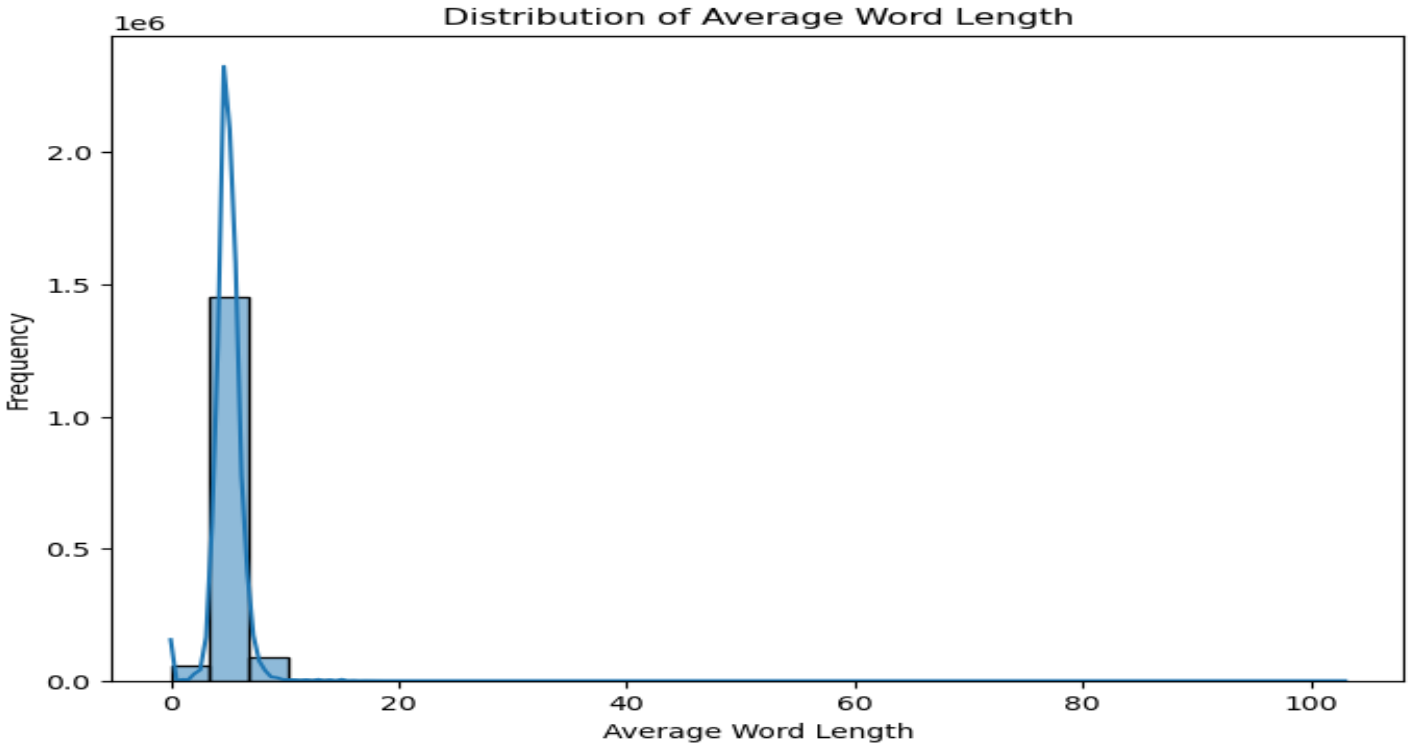


Fig. 5: Distribution of average word length across tweets

We can see that most of the tweets have an average word length of 6, across more than 250K tweets. Then we plotted the distribution of the word count and the character counts across the tweets.

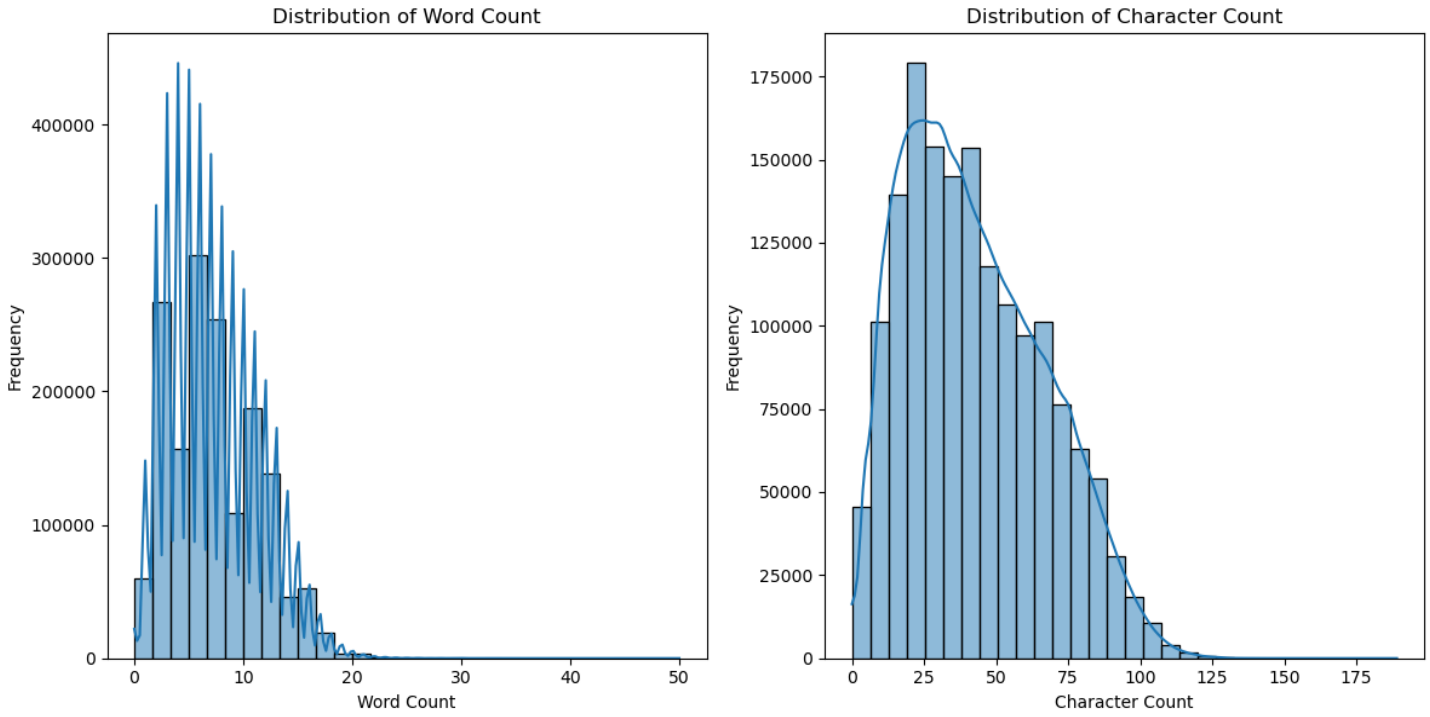


Fig.6: Distribution of word counts and character counts

We can see that most tweets had about 6 words and most tweets have 25 characters. The maximum character count for twitter is 280 characters which would be 47 words in average if we consider the average word length of 6. This experiment shows that people generally like to keep their tweets short and precise.

Modeling

The data used for the analysis are all categorical data, so we had to use classification models. The models we used are **Decision Tree, Random Forest, Gradient Boosting, KNN** and **SVM.** The plots below explain clearly the most important features for this analysis, where it is quite clear that the **negative word count** variable is the most important followed by closely with **positive word count.** For gradient boosting **character count** is more important than **average word length** compared to the other two classification models where the opposite is true.

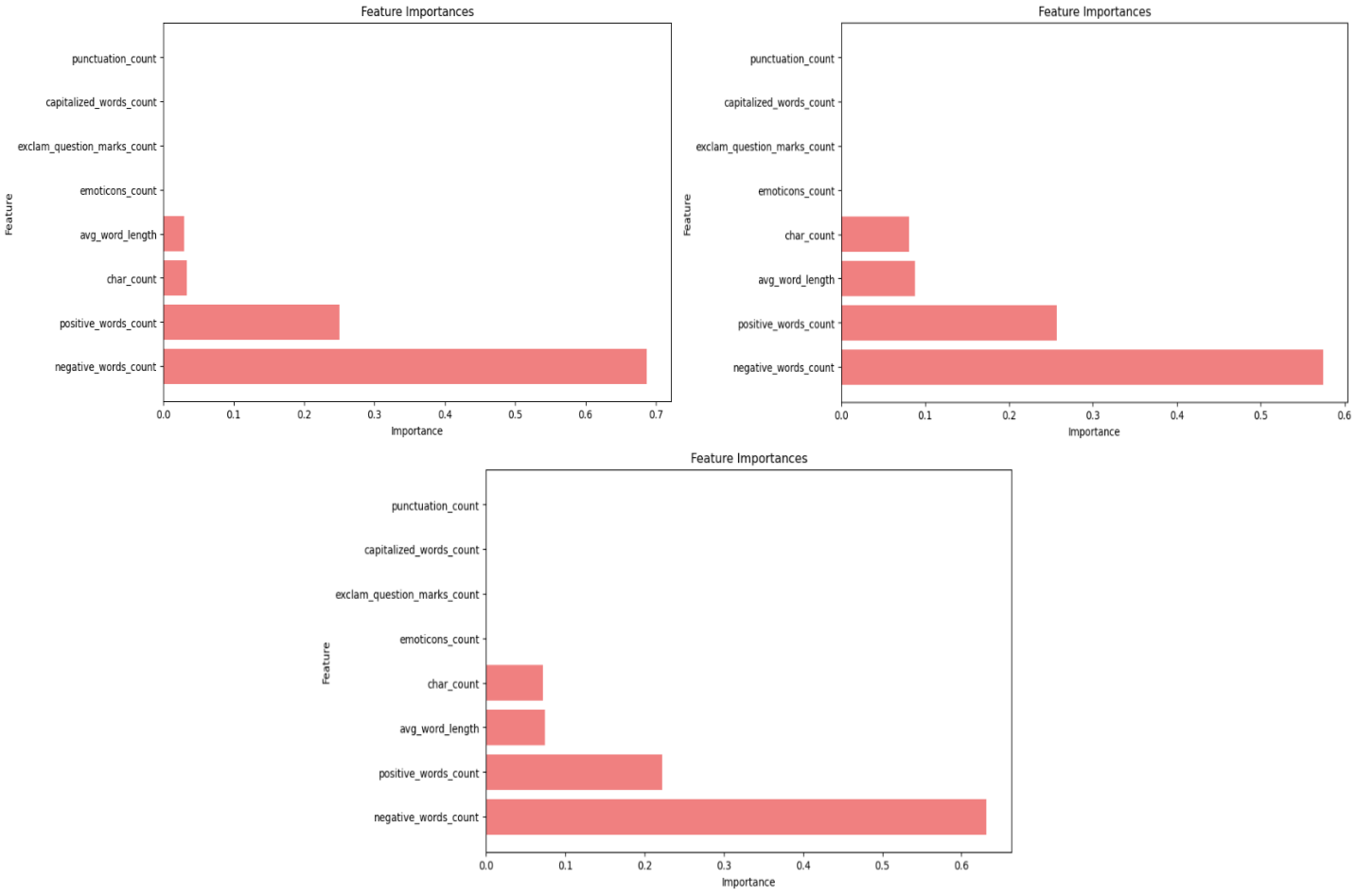


Fig. 7: Top-Left Gradient boosting, Top-Right Random Forest, and Bottom Decision tree feature importance plots

It was quite surprising to find that the **negative word count** had such a vast contribution to the performance of the models compared to the **positive word count variable.** This proves that people generally make tweets only when they find something negative, as the negative sentiment appears to be the more prevalent of the two sentiments.

Model Performance

(in %)	Decision Tree	Random Forest	Gradient Boosting	KNN	SVM
Accuracy	66.31	66.32	66.61	62.72	66.00
F1-Score	66.25	66.25	66.56	62.71	66.07

Table 1: Performance scores of the classification models

We can see from the table that **Gradient Boosting** has the best performance while **KNN** has the worst performance. We performed **data reduction [4]** and used only one-fifth of the data for the **SVM** classification, since **SVM** is a complex classifier, it takes a lot of resources to train it on a dataset of size 1.6M. Hence this was a necessary step. For the Decision tree we can see the structure of the classification task performed by the model.

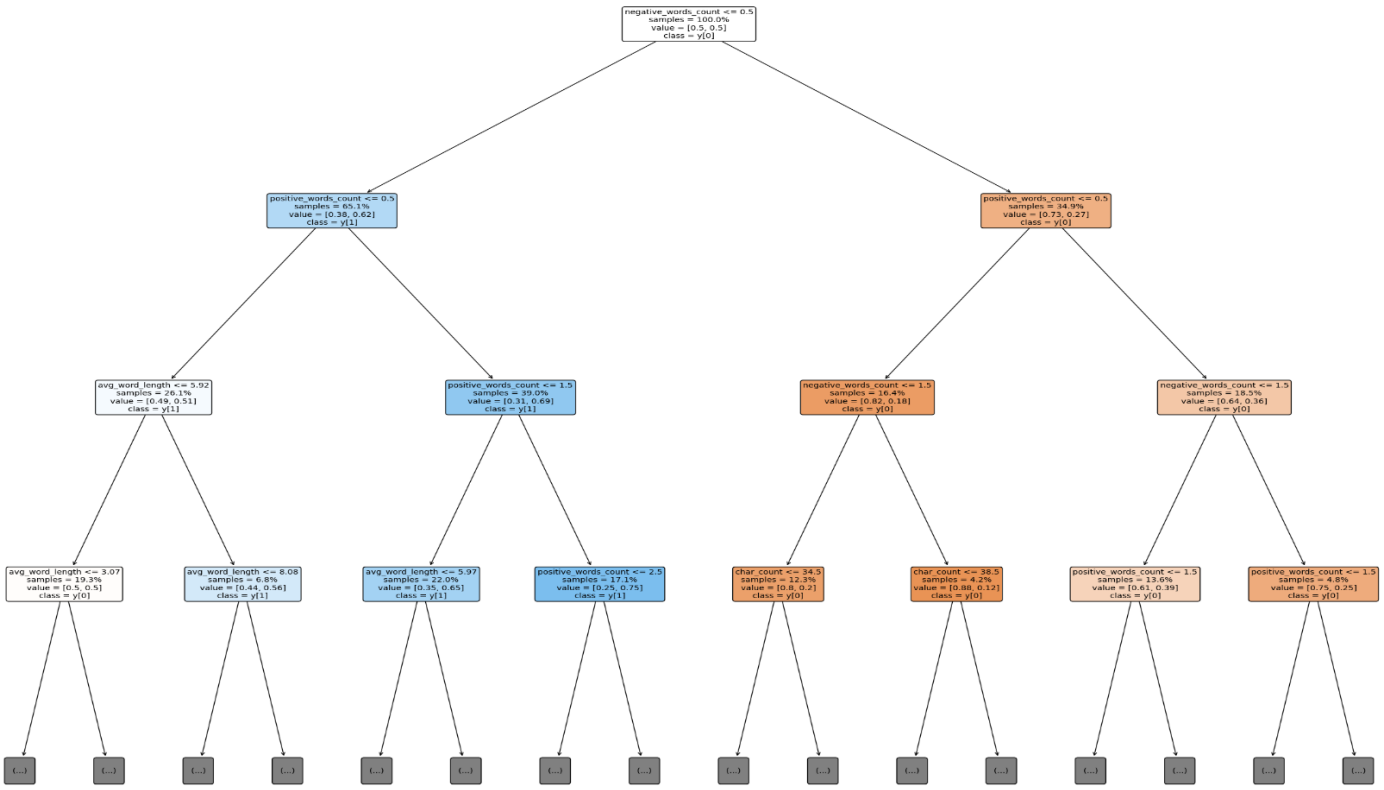


Fig.8: Decision Tree Structure

The above image shows the structure of the **Decision Tree** classifier and describes the structure that the decision tree produces for the analysis. This structure explains the inner workings of the decision tree. We also plotted confusion matrices for the classifiers **SVM** and **Gradient boosting.** The confusion matrices show how the prediction were done by the two classifiers and performance of the two models.

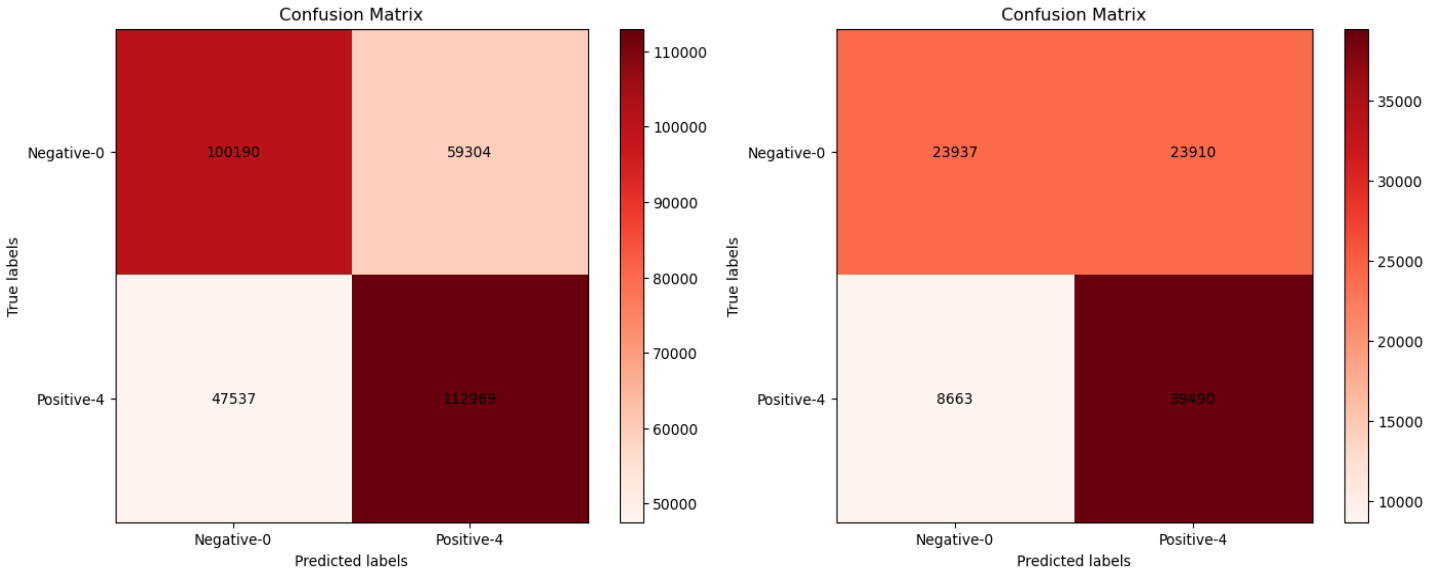


Fig.9: Left- CM for Gradient Boosting and Right- CM for SVM

From the two confusion matrices it becomes quite clear that the **SVM** was confused with the negative sentiments as it predicted nearly equal number of negative sentiments wrong, while its performance for the positive sentiments were nearly all correct. As for **Gradient Boosting,** the models have an overall good performance over both the classes.

Discussion

Initially, we attempted to solve the problem of accurately classifying text sentences as neutral emotions, negative emotions, or positive emotions. We can say that our models have achieved this. Confusion matrix dual analyzes show that the SVM performed equally well in detecting predominantly positive emotions but poorly in detecting negative emotions, which were approximately misclassified. However, the gradient boosting model consistently showed stronger performance in both sensory groups. Sensory classification in textual data is the main theme of the work, which also provides insights from the analytical results. It shows how well the Support Vector Machine (SVM) and Gradient Boosting models distinguish between positive and negative emotions. Our work demonstrates the effectiveness of the Gradient Boosting model in emotion classification in general and highlights the need for more research on the effectiveness of SVM for negative emotions.

Limitations of the paper

Informal language challenges: The first disadvantage of the project is the difficulty in understanding the spoken and non-verbal language of the meeting. Because social networking sites often have multiple languages, jargon and dialects, it is difficult for natural language processing algorithms to understand the content.

Linguistic complexity: Slang, irony, and complex contextual cues can make even the most elaborate natural language processing systems difficult to understand This complexity can lead to misinterpretation or it is inaccurate in sentiment analysis because algorithms could not fully grasp the subtleties of some information.

Subjectivity in textual expression: Textual expression can be subjective and dependent on circumstances. A text can evoke many different emotions in different people depending on their point of view. The subjective nature of the story creates uncertainty and difficulties in accurately categorizing its emotional tone.

Emotional labeling: transcripts can have a profound effect on the bias and interpretation of narratives. The inconsistencies in ratings may be due to the different narratives in which emotions are observed and interpreted differently. This unpredictability can threaten the accuracy and reliability of sensitivity analysis findings.

Inability to Detect Sarcasm and Non-Linearity: The work's inability to distinguish between satire and conventional language is another shortcoming. For algorithms that interpret natural speech, sarcasm, irony, and subtle nuances in speech can make it difficult to understand the intended meaning behind such words.

These limitations point to the difficulties and challenges of sentiment analysis, especially when considering Twitter data with different languages and micro expressions To overcome these shortcomings, natural language must be further studied of design methods and update them to improve the accuracy and simplicity of sensitivity analysis procedures

Future work and conclusions

Find more powerful models: Advances in natural language processing (NLP) and the use of deep learning to find more powerful models can drive the results in this study effectiveness. Modern architectures and techniques can be explored to pursue higher efficiency and accuracy in tasks such as sensitivity analysis. Deep learning improvements: Deep learning models have shown incredible ability to understand and process natural language. Future research should focus on establishing new architectures or improving current models specifically for existing tasks to achieve more accurate sensitivity analysis and improved insight extraction.

Using NLP for marketing strategies: Brands have an exciting opportunity when NLP and deep learning models are integrated into marketing strategy. Brands can use this technology to learn more about consumer sentiment, preferences, and behavior. This will allow them to streamline their marketing activities, increase outreach and build stronger relationships with their target markets.

Enhancing support research during elections: Understanding the dynamics and intentions of voter support during elections is essential. To better understand the contribution process, future research should focus on improving NLP models to provide insightful information for politicians, journalists and decision makers. This can include sentiment analysis of social media conversations, psychoanalysis and public sentiment development to support informed decision-making and policy making.

Conclusion

We have shown in our work how NLP and deep learning models work well for sentiment analysis and support flow analysis. We used this technology to gain insightful insights into user sentiment and engagement, which can have a significant impact on marketing strategy and decision-making, especially during critical times such as election season in. Finally, our work highlights the importance of embracing NLP and deep learning as essential tools for textual content analysis and interpretation in multiple domains. By leveraging this technology, companies can maintain their innovation leadership, encourage informed decision-making, and gain a competitive edge in a data-driven world

References

[1] Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford, 1(2009), p.12*. TwitterDistantSupervision09.pdf (stanford.edu)

[2] Yao, J., 2019, April. Automated sentiment analysis of text data with NLTK. In *Journal of Physics: Conference Series* (Vol. 1187, No. 5, p. 052020). IOP Publishing. Microsoft Word - FB6065 (iop.org)

[3] Bird, S., 2006, July. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions* (pp. 69-72).Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions (aclanthology.org)

[4] Bevington, P.R. and Robinson, D.K., 2003. Data reduction and error analysis. *McGrawâ€“Hill, New York*. textalk.pdf (spy-hill.com)

[5] Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J. and Scholkopf, B., 1998. Support vector machines. IEEE Intelligent Systems and their applications, 13(4), pp.18-28. https://www.zemris.fer.hr/predmeti/su/nastava/ag20022003/SVM-uvodni-clanak.pdf

[6] Friedman, J.H., 2002. Stochastic gradient boosting. Computational statistics & data analysis, 38(4), pp.367-378. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=48caac2f65bce47f6d27400ae4f60d8395cec2f3

[7] Song, Y.Y. and Ying, L.U., 2015. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), p.130. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/

[8] Peterson, L.E., 2009. K-nearest neighbor. Scholarpedia, 4(2), p.1883. http://scholarpedia.org/article/K-Nearest_Neighbor

[9] Biau, G. and Scornet, E., 2016. A random forest guided tour. Test, 25, pp.197-227. https://arxiv.org/pdf/1511.05741

[10] Zien, A., Krämer, N., Sonnenburg, S. and Rätsch, G., 2009. The feature importance ranking measure. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II 20 (pp. 694-709). Springer Berlin Heidelberg. https://arxiv.org/pdf/0906.4258

[11] Townsend, J.T., 1971. Theoretical analysis of an alphabetic confusion matrix. Perception & Psychophysics, 9, pp.40-50. https://link.springer.com/content/pdf/10.3758/BF03213026.pdf

[12] Heydarian, M., Doyle, T.E. and Samavi, R., 2022. MLCM: Multi-label confusion matrix. IEEE Access, 10, pp.19083-19095. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9711932

[13] Giachanou, A. and Crestani, F., 2016. Like it or not: A survey of twitter sentiment analysis methods. ACM Computing Surveys (CSUR), 49(2), pp.1-41.Like-It-Or-Not

[14] Sarlan, A., Nadam, C. and Basri, S., 2014, November. Twitter sentiment analysis. In Proceedings of the 6th International conference on Information Technology and Multimedia (pp. 212-216). IEEE. https://www.researchgate.net/profile/Aliza-Sarlan/publication/301408174_Twitter_sentiment_analysis/links/581a897508ae30a2c01caf20/Twitter-sentiment-analysis.pdf

[15] Zimbra, D., Abbasi, A., Zeng, D. and Chen, H., 2018. The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation. ACM Transactions on Management Information Systems (TMIS), 9(2), pp.1-29. https://dl.acm.org/doi/pdf/10.1145/3185045