Comparison of various sentimental analysis for market segmentation

Minor Project Report

Prepared by

Specialization	Name	SAP ID
B. Tech CSE BAO	Aryan Thakur	500083130
B. Tech CSE BAO	Samriddh Goyal	500086928
B. Tech CSE BAO	Varnit Tomar	500083263

Department of Informatics
School Of Computer Science
UNIVERSITY OF PETROLEUM & ENERGY STUDIES,
DEHRADUN- 248007. Uttarakhand

Dr. Ashutosh Sharma
Project Guide

Dr. TP Singh
Cluster Head

Project Title:- Comparison of various sentimental analysis for market segmentation

Abstract

In today's world Social media platforms have become an integral part of people’s lives and are commonly used to convey their thoughts and feelings. This develops a wealth of information that can be used to study how people feel about any issue or any product. Companies and organizations use this data to better understand their target audience and improve their businesses. However, it becomes challenging to get some insights from this data to its sheer volume. Sentimental analysis is a technique that extracts subjective information from the text data, which can be used to analyze the social media data and get some insights into the opinions, and emotions expressed by the users towards any particular topic. This project aims to compare various sentimental analysis techniques including lexicon-based approaches, and machine learning-based approaches to classify tweets as positive, negative, or neutral. The accuracy of these techniques will be calculated on a dataset containing tweets related to any topic. Furthermore, machine learning will be used to create our own sentimental analysis model and evaluate its performance. Algorithms such as Naive Bayes, Random forest will be used to develop a model. We also aim to segment users based on their similar sentiments using the clustering algorithms such as k-means and perform market segmentation for a better understanding of the user needs and develop marketing strategies for specific target groups. Overall the project aims to demonstrate the effectiveness of various sentiment analysis techniques and their possible applications in market segmentation.

Keywords:- Sentimental analysis, machine learning, clustering, Twitter, VADER, TF - IDF, Bag of Words, random forest, accuracy

UPES

UNIVERSITY OF TOMORROW

Acknowledgment

We would like to convey our sincere gratitude to our mentor, Dr. Ashutosh Sharma, for all the guidance, inspiration, and unwavering support he provided us with throughout the course of working on our project. Without his encouragement and helpful recommendations, this effort would not have been feasible.

We really appreciate Dr. TP Singh, the cluster chair, for all of his help with our study at SOCS.

Additionally, we are grateful to our kind Dean of SOCS, UPES Dr. Ravi S. Iyer for providing us with the tools we needed to complete our project work effectively.

We appreciate the assistance and helpful feedback provided by all UPES faculty members during the course of our project work. Finally, we can only thank our parents in the deepest sense for showing us the world and for all of the support they have given us.

List of figures

Table of Contents

1. Introduction

This project focuses on sentiment analysis, which involves using natural language processing techniques to determine the sentiment of a piece of text. The project aims to compare the effectiveness of various sentiment analysis techniques and algorithms and to develop a machine learning model to segment users based on their similar sentiments. The project also aims to provide insights and analysis on the sentiment of Twitter users.

1.1 Purpose of the project

The goal of this project is to investigate and compare several strategies for sentimental analysis of social media data, such as pre-built libraries like TextBlob and VADER, as well as to construct a bespoke machine learning model. Furthermore, the project intends to use clustering algorithms to segment users based on their sentiments in order to gain insights into how different groups of users perceive a particular product, service, or brand. The ultimate goal is to give a platform that allows businesses and organizations to better understand their clients and enhance their products or services based on social media sentiment.

1.2 Target Beneficiary

This project's intended beneficiaries are businesses and organizations that rely on client input to improve their products or services. Businesses can acquire important insights into client preferences and make data-driven decisions to boost customer experience and loyalty by analyzing customer sentiment and segmenting them based on similar feelings. Furthermore, the project may be of interest to researchers and academics interested in sentiment analysis and clustering algorithms. This project's code and techniques can be utilized as a starting point for further research and development in the subject.

1.3 Project Scope

The scope of this project is to develop and test several sentimental analysis approaches on Twitter data to classify them as positive, negative, or neutral. The project will explore different machine learning algorithms such as Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine to build a sentiment analysis model. The model's accuracy will be assessed and compared to other techniques like TextBlob and VADER. It also intends to use market segmentation techniques to group Twitter users based on their sentiment toward a specific topic. K-Means clustering methods will be used to categorize users and detect patterns in their sentiments. Businesses and organizations will benefit from a better understanding of their clients and will be able to modify their marketing tactics accordingly.

2. Literature review

Sentiment analysis is a novel type of text analysis that seeks to determine reviewers' opinions and subjectivity. The reason for this is that among the vast number of evaluations, there are some that do not contain any overtly subjective terms yet represent a powerful viewpoint. [1]One of the previous research used the lexicon methods in which first, the reviews are preprocessed: they are split into sentences which are corrected and the method "Part Of Speech" (POS) is used to tag and store each word of the sentence. The performance has been assessed with an accuracy of 91% at the review level and 86% at the sentence level. The problem countered was the link between generated clusters and latent variables in probabilistic models has yet to be rigorously explained.

Another study on market segmentation was based on using the[3] K - means and the agglomerative clustering algorithms for high-level customer segmentation retailers have begun to incorporate aspects of machine learning into the analysis of their customers. Retailers are approaching analysis in ways that cannot be matched without the use of unsupervised machine learning technologies such as clustering and dimensionality reduction.

Sentimental analysis is also in high demand in the clinical and health informatics sector. To utilize the huge data available on social media about sentiment and emotional expressions NLP(Natural Language Processing) is taken under consideration. The main problem encountered here was data collection, it was very difficult for the researchers and practitioners in psychiatry to access data that truly reflects the mental state of the patient. For NLP concepts of text preprocessing, Lexical analysis, Syntactical analysis, and Semantic analysis are important to understand.

3. Problem statement

The problem statement for this project is the requirement for appropriate sentiment analysis algorithms to handle the large volume of text data created by social media platforms, customer reviews, and other sources. The sudden boom of social media and online review platforms has made it difficult for businesses to maintain track of public opinions of their products and services. We need an effective system for sentimental analysis that can classify opinions as positive, negative, or neutral. The existing traditional sentimental analysis approaches sometimes fall short due to the nuances of natural language processing and the constant evolution of language trends. The project will also determine the effectiveness of sentimental analysis techniques such as TextBlob, and VADER. Additionally, there is a need to segment users based on similar sentiments using clustering algorithms which will be useful for targeted marketing and personalized customer experience. Overall, the project aims to give a comprehensive solution to the problem of analyzing massive amounts of social media data and extracting useful insights for businesses and organizations.

4. Existing system issue

Based on the literature review done for this project the major issues in the existing projects are:-

- Accuracy - Accuracy of the sentimental analysis techniques varied depending on the complexity of the tweets, sarcasm in text and some contextual nuances. Improving the accuracy of the sentimental analysis techniques is a challenging task.
- Data quality - Data extracted from Twitter includes a lot of noise in the form of irrelevant content present unfortunately, this degrades the accuracy of the sentimental analysis model.
- Multilingual analysis - Performing sentimental analysis in languages other than English is a challenge.
- Data imbalance - The dataset for sentimental analysis suffers a class imbalance maximum times, classifying maximum tweets as neutral.

5. Project Description

5.1 Reference Algorithm

The algorithms which are taken into account for this project are:-

- TextBlob - It is a Python library used for processing textual data. It is based on the Natural Language Toolkit (NLTK) library and offers a simple interface for doing natural language processing tasks such as sentiment analysis. TextBlob assigns polarity scores to text inputs using a pre-trained machine learning model.

- VADER - It is the "Valence Aware Dictionary and Sentiment Reasoner", a lexicon and rule-based sentiment analysis tool specifically designed to analyze sentiments in social media texts. It determines the sentiment of a given text using a combination of linguistic rules and a sentiment lexicon, and it is noted for its accuracy and efficiency in analyzing social media data.

- TF IDF - It stands for term frequency-inverse document frequency and is a common strategy in natural language processing for determining the relevance of words in a document. It provides a weight to each word depending on its frequency in a document and rarity in the corpus.

- Bag of words - It is a popular text analysis method in Natural Language Processing. It counts the frequency of each word in a document and displays it as a bag of words ignoring the order of occurrence of the words. This method is effective in classification and clustering tasks.

- Naive Bayes - It is a classification technique that employs the Bayes theorem to predict the likelihood of a specific class based on a set of features. In sentiment analysis, the features are the words in the text input, and the classes are the various sentiment categories (positive, negative, neutral). Naive Bayes assumes that all features are independent of one another, which can be a convenient but sometimes incorrect assumption.

- K - Means - It is an unsupervised machine learning algorithm that divides the data points into k clusters based on their similarity. K - means can be used in sentimental analysis to cluster people based on the similarity of their sentiment scores for different inputs.

5.2 Characteristics of data

The data used for this project is Twitter data extracted through API have the following characteristic:-

- Textual data - The data contains textual information in the form of tweets that are short messages of 280 characters or less.

- Unstructured data - The extracted tweets are unstructured which means they do not have any predefined format

- Noisy data - The tweets contain noise in the form of misspellings, emojis, abbreviations, and URLs that can affect the accuracy of the sentimental analysis.

- Volume - Twitter generates a large volume of data.

- Realtime data - Data is generated in real-time.

- User-generated content - Reflects opinions, and emotions of people on any particular product or issue.

5.3 SWOT Analysis

Strengths

- The project can aid in the comprehension of consumer sentiment and preferences, hence improving the customer experience.
- The project can be altered to meet particular company needs, and the insights gained can be added to other systems to improve decision-making capacity.
- Uses NLP techniques to analyze the sentiment of textual data.
- Results can potentially be useful for the business in developing new marketing strategies for the target market.

Weaknesses

- Accuracy depends on the data extracted.
- May face legal and ethical concerns due to the privacy issue of data.
- NLP techniques may not accurately capture the intended meaning of the text.

Opportunities

- The information gathered can be utilized to improve marketing efforts and client interaction.
- Can be integrated with other analytical tools to provide a more precise picture of the market and customer behavior.
- The use of Twitter data for market segmentation is a relatively new area of study, with much room for additional investigation and development.

Threats

- Privacy concerns for the personal data of the users extracted through API.
- Continuously changing the nature of social media can make it difficult to maintain the analysis's relevance and accuracy.

5.4 Project features

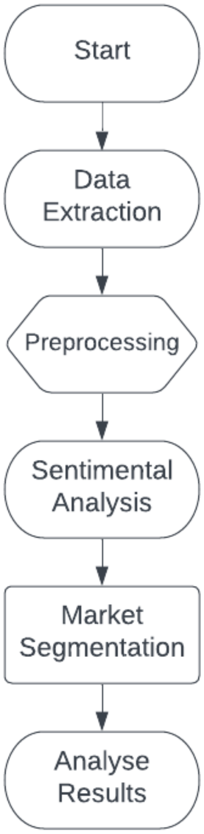


Fig 1. FlowChart

The basic flow of the project goes starts from data extraction through API, creating a CSV dataset which goes on for the preprocessing part. On the preprocessed database created sentimental analysis techniques are applied and the accuracy of the same is recorded. Based on the best accuracy recorded segmentation of tweets is done using clustering methods for getting insights into the market about brand reputation.

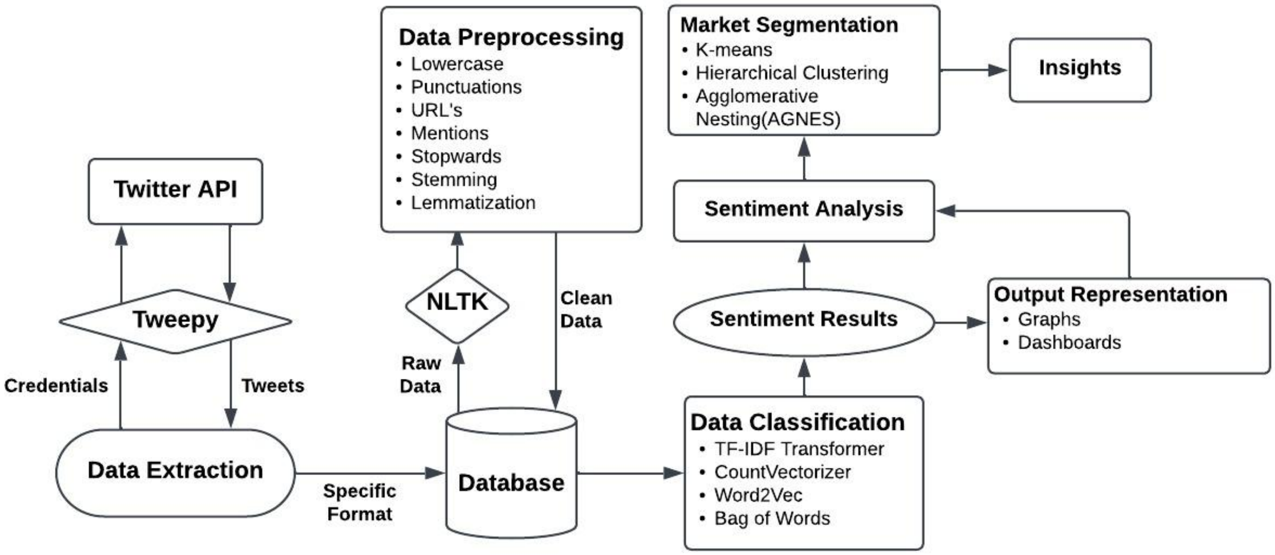


Fig 2. UML

- Data collection through API
- Preprocessing of data
- Implementation of various sentiment analysis
- Evaluation of accuracy
- Implementation of clustering algorithms for market segmentation
- Visualization of results through graphs

4.5 <u>User Classes and Characteristics</u>

Potential user classes that would benefit from this project will be -

1. Customers: These will be the primary target audience for this project and will be segmented based on their sentiments towards the product or the company.

2. Competitors: These are also an important user class for this project, analysing their data can help in improving competitive strategies and product offerings.

3. Marketers: This class contains the individuals who are responsible for marketing the product. They should have expertise in marketing and knowledge of market segmentation. They should have the ability to apply the insights gained from sentimental analysis to marketing strategies.

4. Analysts: They are the individuals who are responsible for analyzing the data generated by sentimental analysis. They should have expertise in data analysis and statistical methods.

5. Researchers: They may be interested in analyzing sentiment in large datasets to identify patterns or trends in public opinion towards a particular topic or issue.

6. PERT Chart

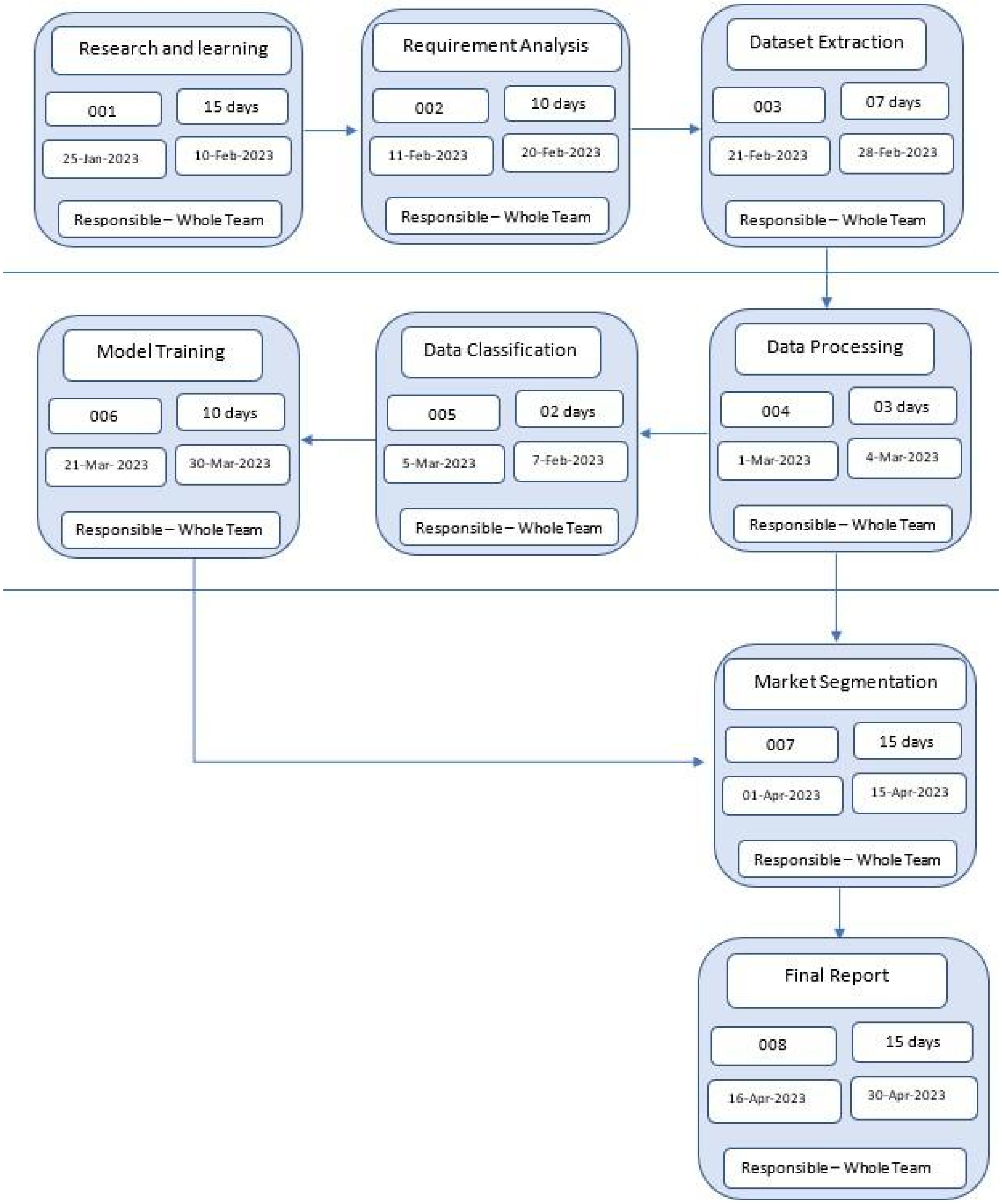
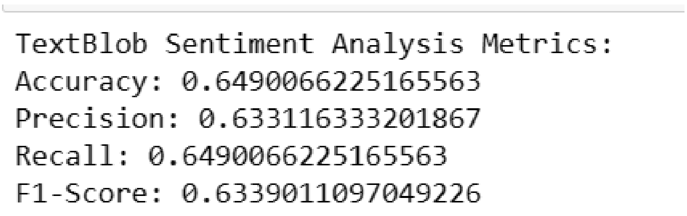


Fig 3. PERT chart

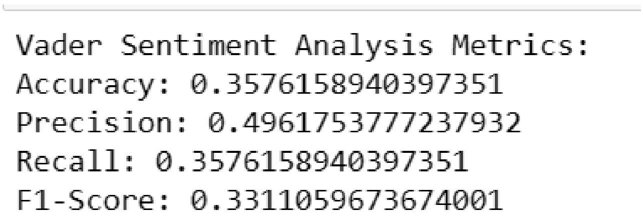
7. Results

We analyzed the Twitter data of MAC Cosmetics and tested various sentimental analysis techniques on it including TextBlob, and VADER.

These are the accuracies that were recorded for these techniques using the manual modeling method.

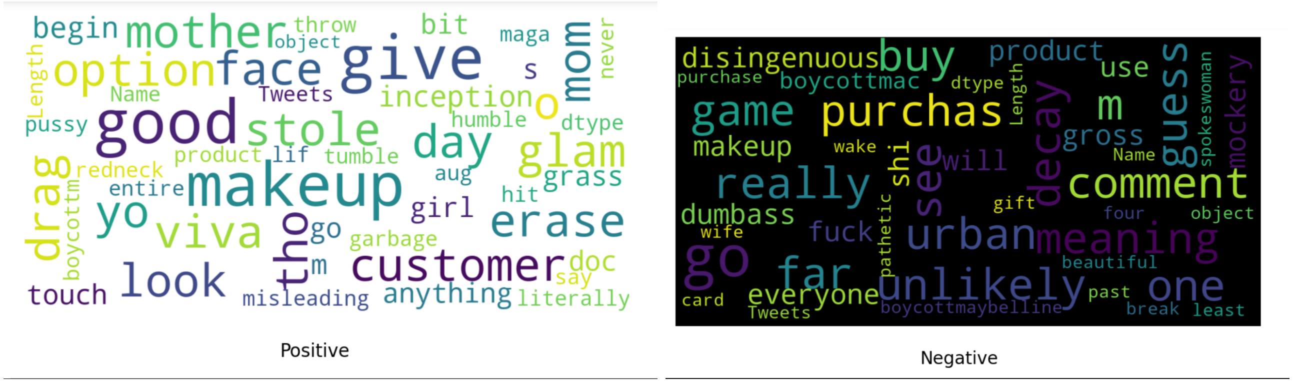


```
TextBlob Sentiment Analysis Metrics:
Accuracy: 0.6490066225165563
Precision: 0.633116333201867
Recall: 0.6490066225165563
F1-Score: 0.6339011097049226
```



```
Vader Sentiment Analysis Metrics:
Accuracy: 0.3576158940397351
Precision: 0.4961753777237932
Recall: 0.3576158940397351
F1-Score: 0.3311059673674001
```

For further classification of these tweets, we created word clouds for negative and positive tweets.



Positive

Negative

Fig 3. Wordcloud

To get insights about the brand and the classification we used the bag of words method.

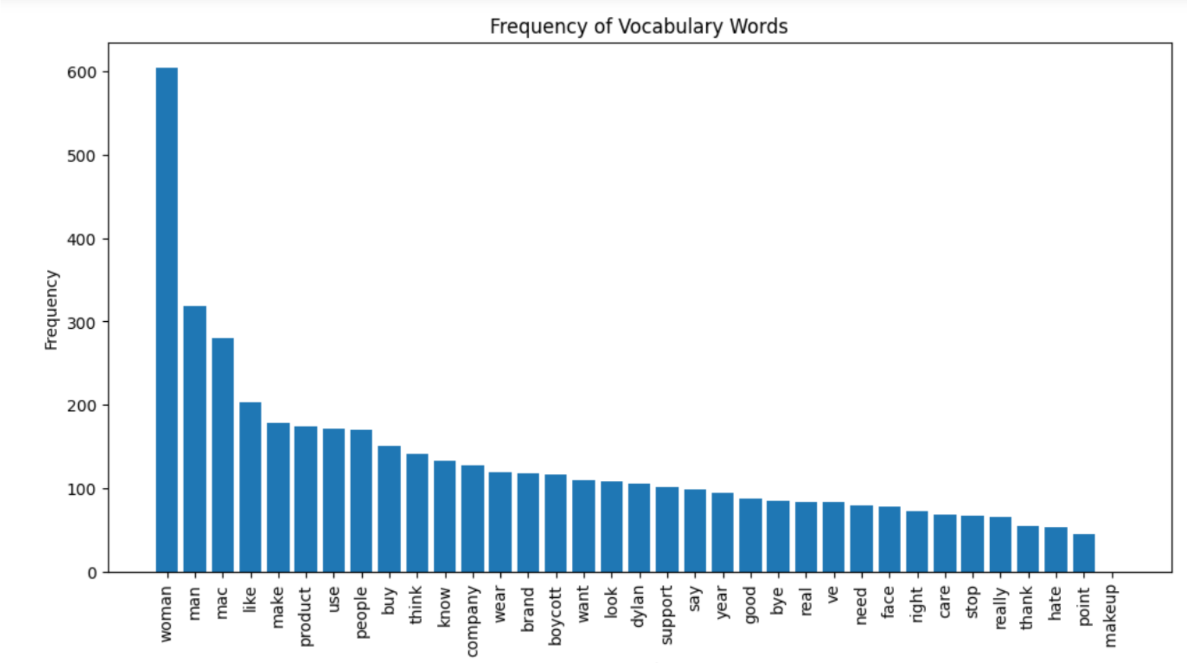


Fig 4. Feature frequency graph

Graph for Word2Vec method

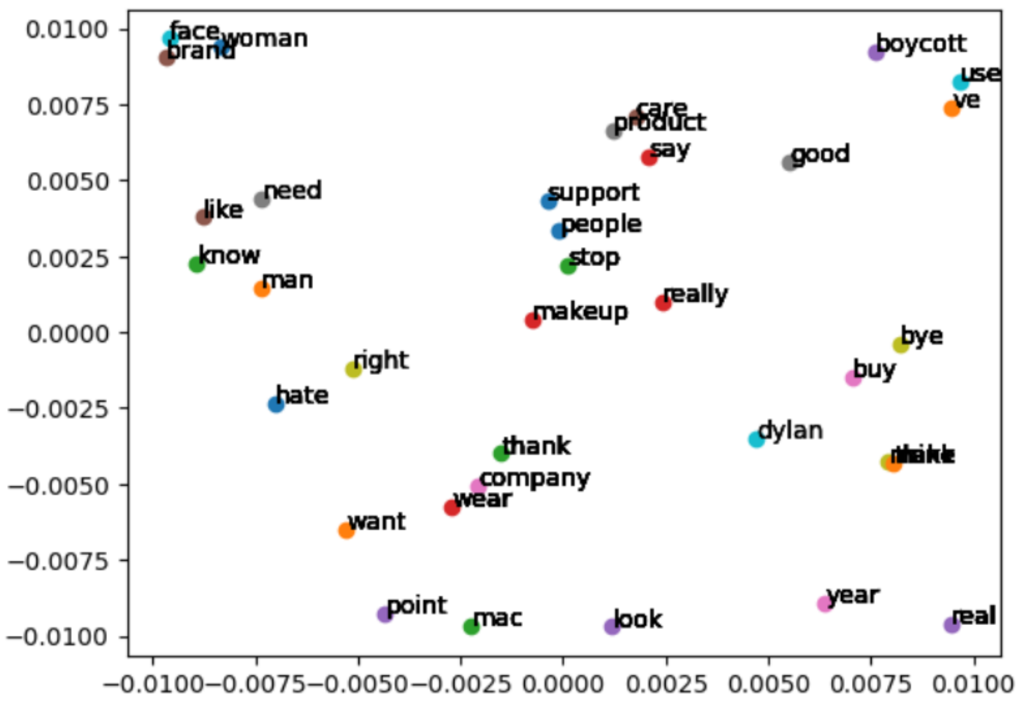


Fig 5. Scatter word plot

K means clustering gave a result that most of the users fall in the positive region and the brand positivity is more in market.

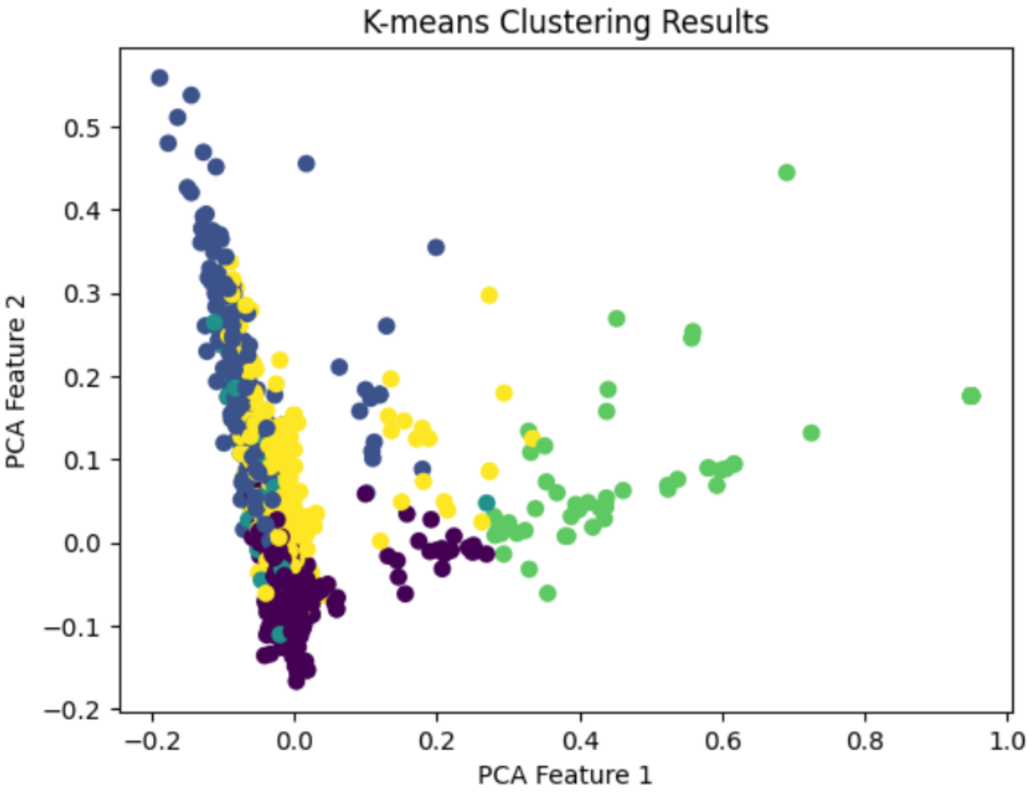


Fig 6. K Means

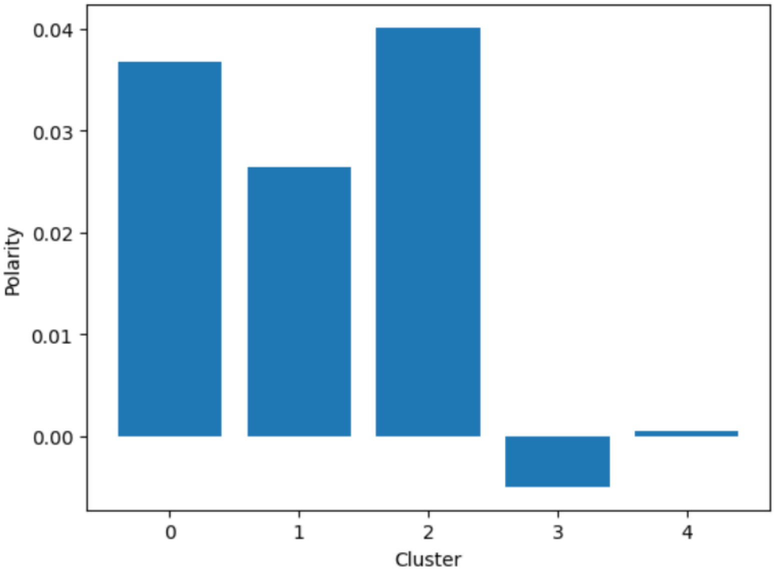


Fig 7. K means bar

8. Conclusion

Finally, this project demonstrated the utility of various sentiment analysis techniques on Twitter data. The accuracy of each method was tested and compared, and a new machine-learning model was constructed to accurately classify sentiment. Furthermore, using clustering algorithms to segment Twitter users based on similar sentiments provided valuable insights for targeted marketing campaigns. Overall, this study demonstrated the utility of sentiment analysis across a wide range of businesses and contexts, from customer service to political analysis. As the relevance of social media grows, sentiment analysis will become even more important in understanding and responding to public opinion.

9. Future work

This project can become a good marketing tool for businesses. Using some of the advanced sentimental analysis models and integrating multiple data sources to it like social media platforms and online surveys a sentiment-based recommendation system can be built. The system can provide personalized product and service recommendations

Real-time analysis is also possible to analyze customer sentiments as they evolve over time, this will also allow businesses to promptly respond to the emerging trends in the market

References

[1] A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation
URL - https://liris.cnrs.fr/Documents/Liris-6508.pdf
Author Ana¨ıs Collomb, Crina Costea, Damien Joyeux, Omar Hasan, Lionel Brunie

[2] Natural Language Processing, Sentiment Analysis, and Clinical Analytics
URL
https://www.researchgate.net/publication/330871275_Natural_Language_Processing_Sentiment_Analysis_and_Clinical_Analytics/link/5c6023d792851c48a9c73793/download
Author - Adil Rajput

[3] CUSTOMER SEGMENTATION ANALYSIS OF CANNABIS RETAIL DATA: A MACHINE LEARNING APPROACH
URL - https://core.ac.uk/download/pdf/288175101.pdf
Author - Ryan Henry Papetti

[4] Yener, Y. (2021, December 16). Step by Step: Twitter Sentiment Analysis in Python - Towards Data Science. Medium.
https://towardsdatascience.com/step-by-step-twitter-sentiment-analysis-in-python-d6f650ade58

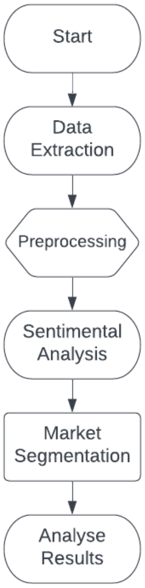
[5]A Sentiment Analysis Based Approach for Customer Segmentation
Author - Anisha Bhatnagar and Madhulika Bhatia

Appendix A: Glossary

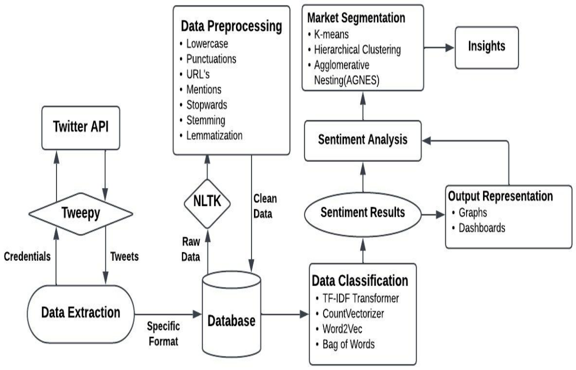
- NLP - Natural Language Processing
- API - Application programming interface
- VADER - Valence Aware Dictionary and Sentiment Reasoner
- TF IDF - Term frequency-inverse document frequency

Appendix B: Analysis Model

- Flow chart



- UML



Appendix C: Issue List

- Data privacy
- API Limitations