SVKM'S NMIMS Deemed to be UNIVERSITY

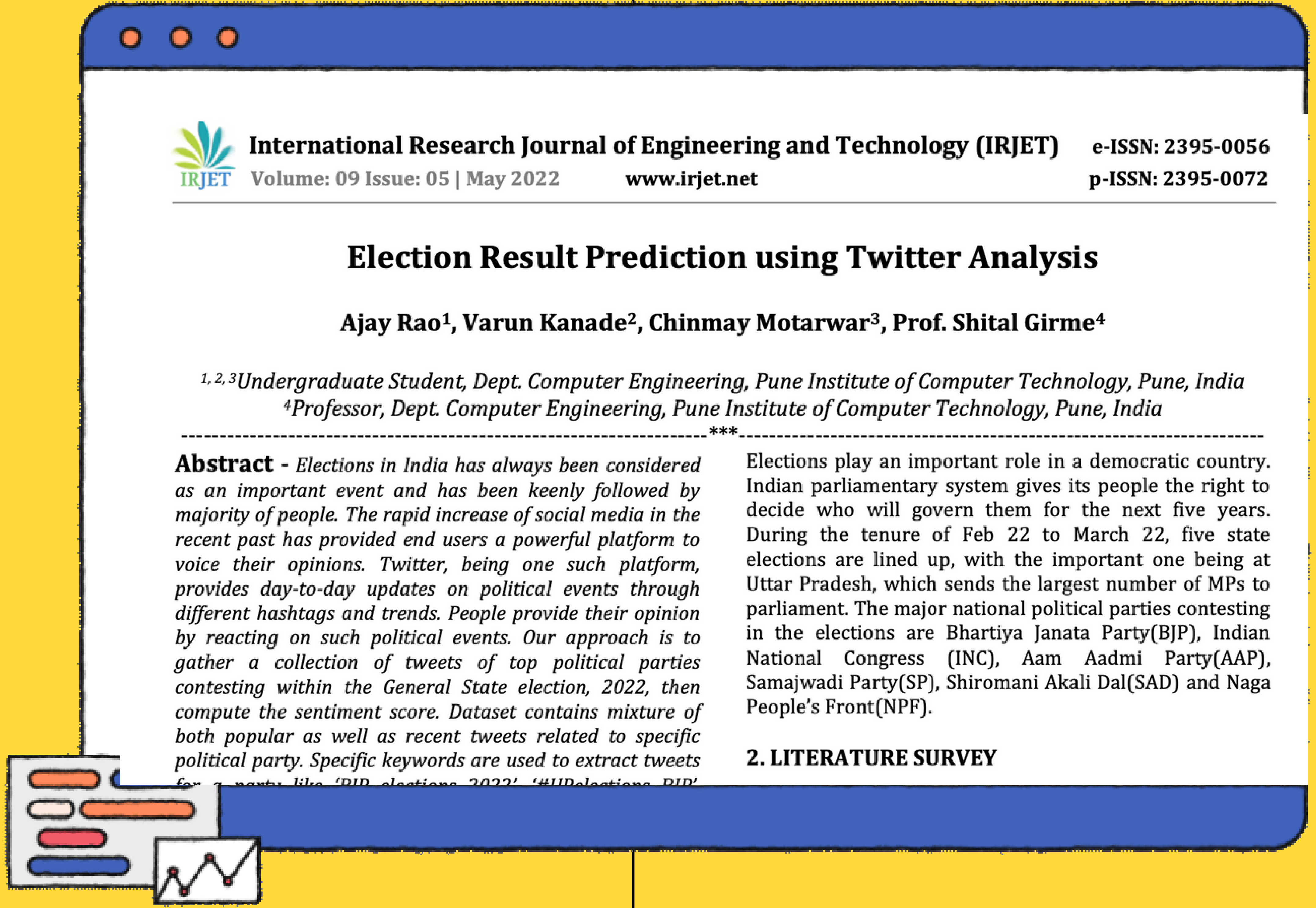
SUNANDAN DIVATIA SCHOOL OF SCIENCE

Election (2023) Prediction Using Twitter Data

Mentor: **Mr. Rajesh Maurya**

Presented by:

Shagun Kulshreshtha – 38
Vinayak Mokashi – 40
Triveni Bisen – 43
Mahesh Ahire – 48
Rohit Saigaonkar – 58



International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 | May 2022 www.irjet.net p-ISSN: 2395-0072

Election Result Prediction using Twitter Analysis

Ajay Rao[1], Varun Kanade[2], Chinmay Motarwar[3], Prof. Shital Girme[4]

[1, 2, 3]*Undergraduate Student, Dept. Computer Engineering, Pune Institute of Computer Technology, Pune, India*
[4]*Professor, Dept. Computer Engineering, Pune Institute of Computer Technology, Pune, India*

---***---

Abstract - *Elections in India has always been considered as an important event and has been keenly followed by majority of people. The rapid increase of social media in the recent past has provided end users a powerful platform to voice their opinions. Twitter, being one such platform, provides day-to-day updates on political events through different hashtags and trends. People provide their opinion by reacting on such political events. Our approach is to gather a collection of tweets of top political parties contesting within the General State election, 2022, then compute the sentiment score. Dataset contains mixture of both popular as well as recent tweets related to specific political party. Specific keywords are used to extract tweets for a party like 'BJP elections 2022', '#UPelections BJP'*

Elections play an important role in a democratic country. Indian parliamentary system gives its people the right to decide who will govern them for the next five years. During the tenure of Feb 22 to March 22, five state elections are lined up, with the important one being at Uttar Pradesh, which sends the largest number of MPs to parliament. The major national political parties contesting in the elections are Bhartiya Janata Party(BJP), Indian National Congress (INC), Aam Aadmi Party(AAP), Samajwadi Party(SP), Shiromani Akali Dal(SAD) and Naga People's Front(NPF).

2. LITERATURE SURVEY

RESEARCH PAPER:

Election Result Prediction using Twitter Analysis

Published in May 2022

Motivation

- Social media: Shrinking the world, connecting millions of people across the globe
- Twitter, Facebook, Instagram, Google+, and more
- Sharing of opinions, experiences, reviews, ratings
- A democracy: of the people, by the people and for the people
- Election is the most important aspect of a democracy
- Social media enables the people to voice their strong and various opinions about leaders

Objective

- To analyze tweets collected from Twitter
- To build a robust model to predict future election outcomes

BMC Elections 2023

Contender A

Contender B

Overview

Approach

Data Collection

Using TweePy and SnsScrape to extract tweets

Data Preprocessing

Turning unstructured tweets to structured and clean data

Sentiment Analysis

Using VADER for sentiment analysis

Model Training

SVM, Naïve Bayes, Decision Trees, Multinomial Logistic Regression, KNN, Bagging (SVM) was used

Predictions

Calculation of popularity score and making prediction of winner

Data Collection

DATA EXTRACTION FROM TWITTER

Using TweePy:

A Python library for accessing the Twitter API

Drawback: Only allows 3,200 and 7 days old Tweets to be scraped

Using SnScrape:

A scraper for social networking services (SNS)

Data extracted using Hashtags

Attributes of Tweet could be Extracted

50,000 for each dataset

30th June 2022 - 15 November 2022

Datasets Collected:

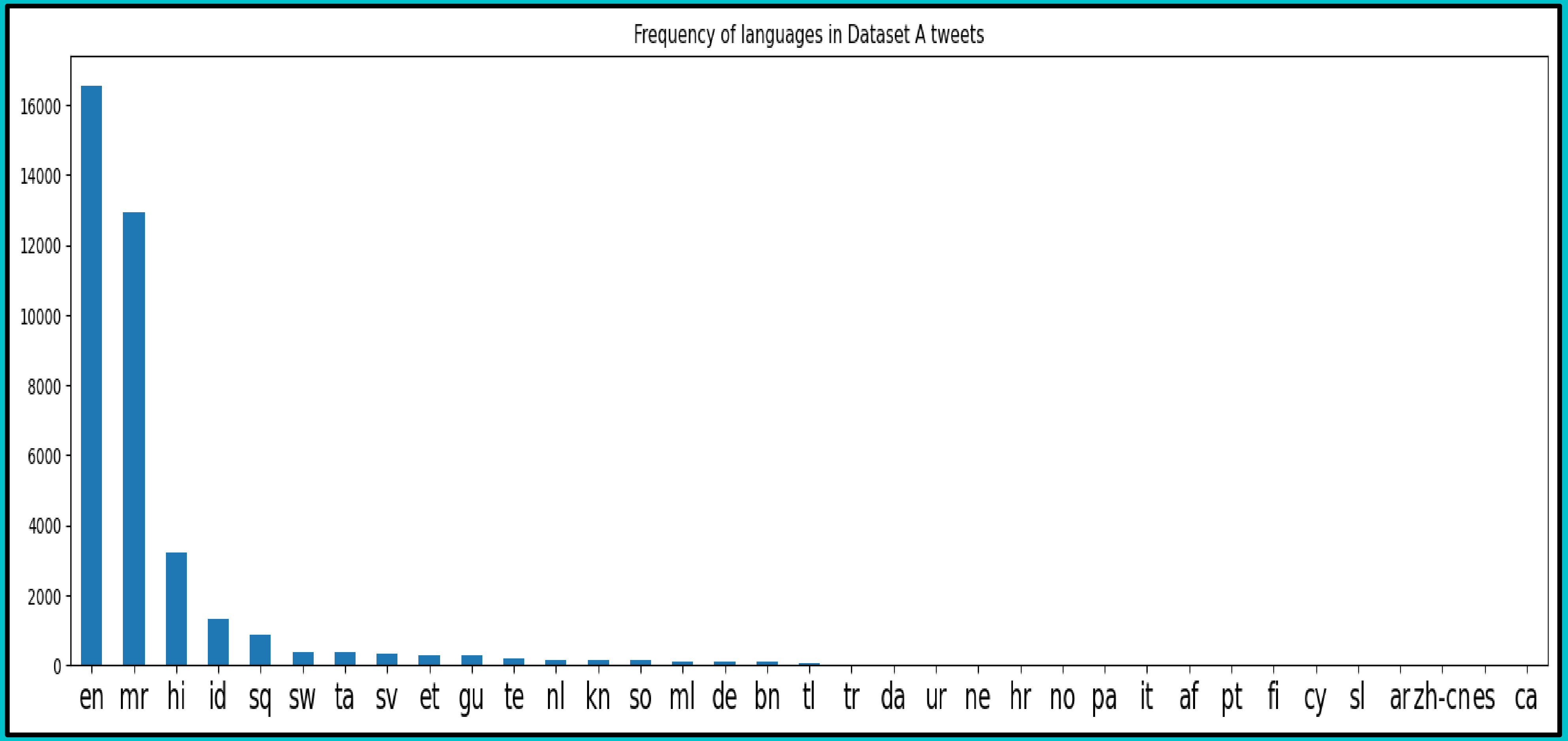
Tweets from Contender A

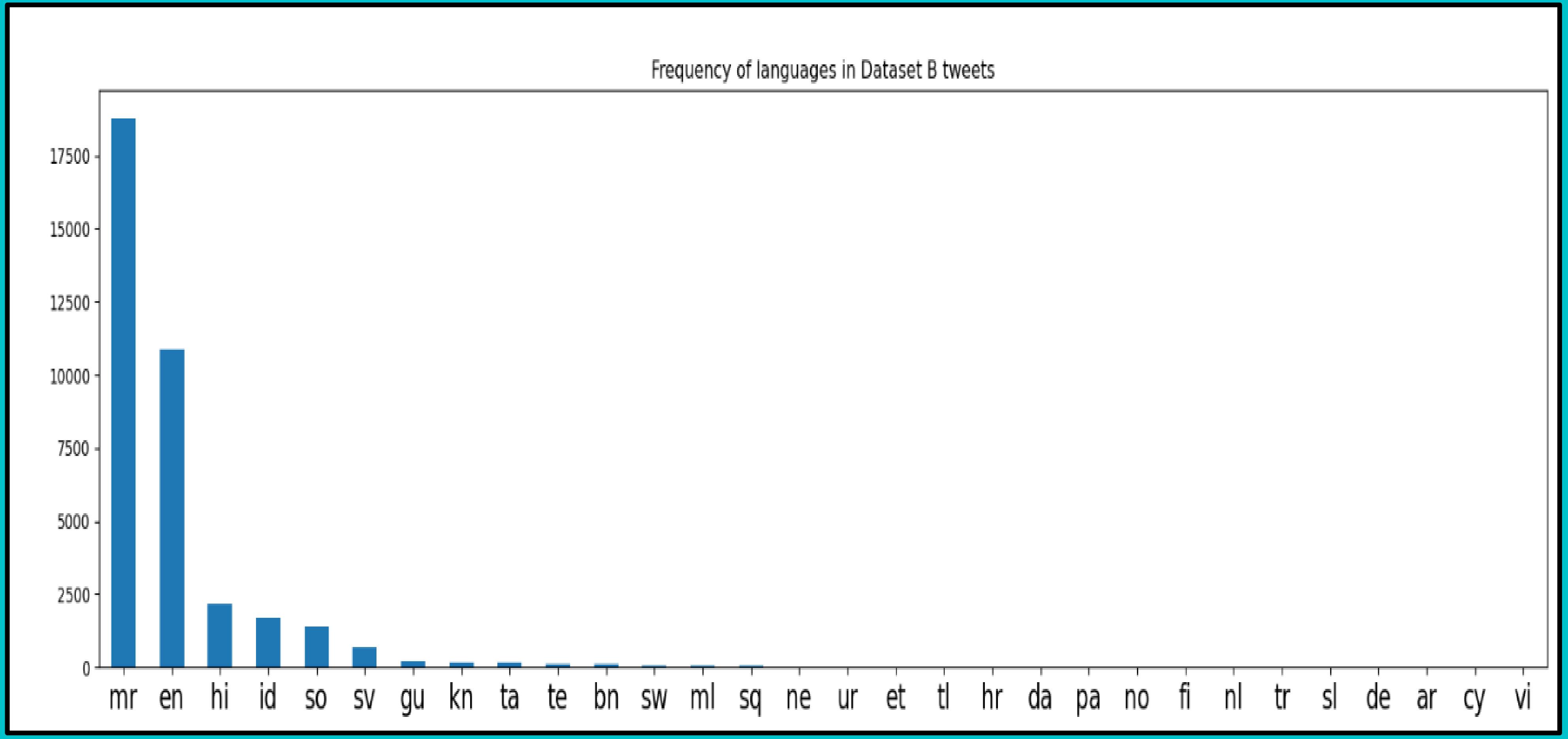
Tweets from Contender B

Attributes:

Likes_Count
Retweets_Count
UserName
Date
Tweet

Data Preprocessing


Frequency of languages in Dataset A tweets


Frequency of languages in Dataset B tweets

- **Duplicates removed**: 37717 Tweets left
- **Language barriers**: Only English Tweets kept A: 16569 and B: 10878 left
- **Text cleaning**: Removal of Punctuations Special characters, URLs and Hashtags, extra white spaces

Data Preprocessing

- **Stop-word removal**:

  Removing the words that occur commonly across all the documents in the corpus using NLTK

- **Stemming**:

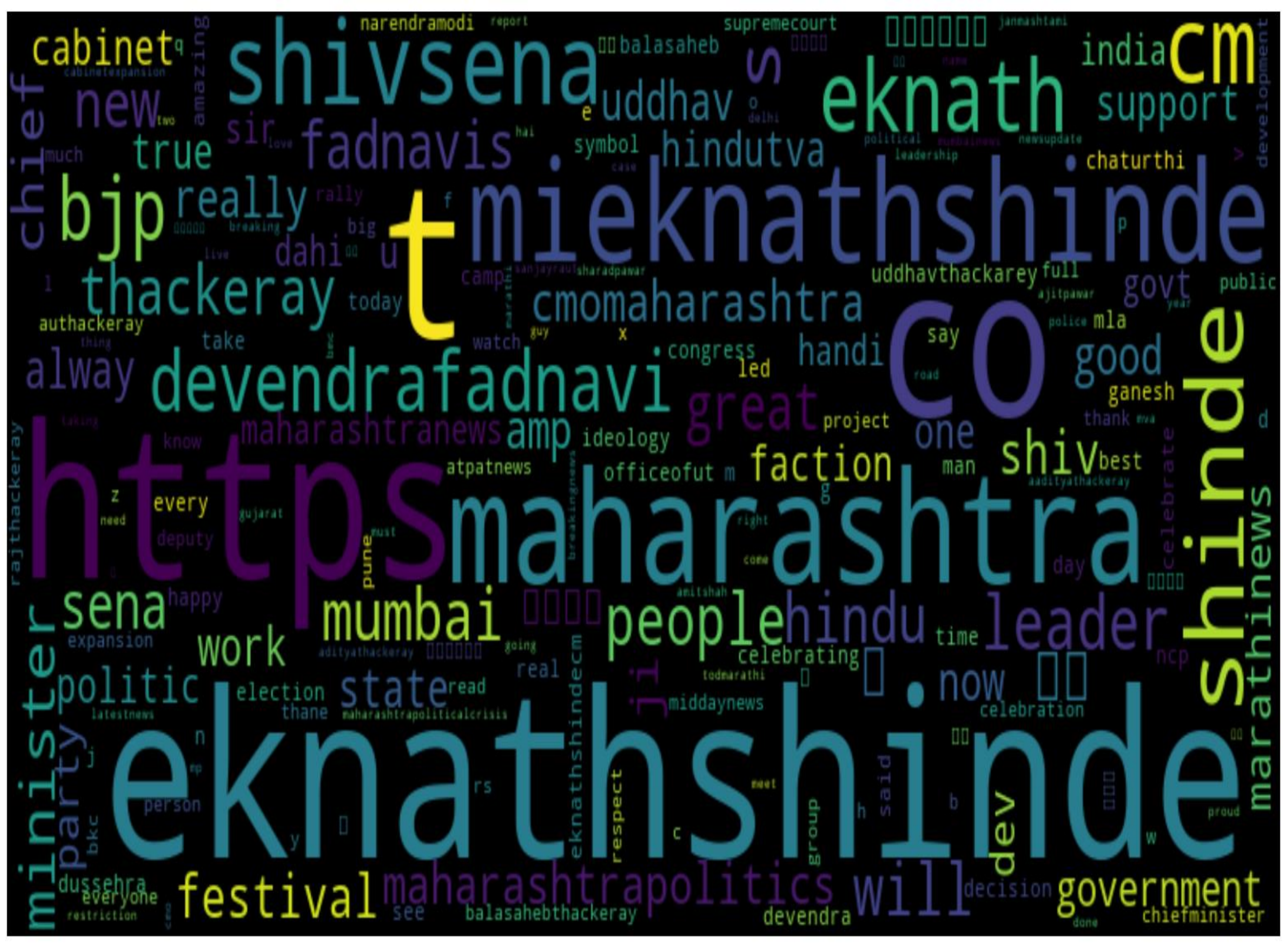
  Process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as stem

- **Lemmetization**

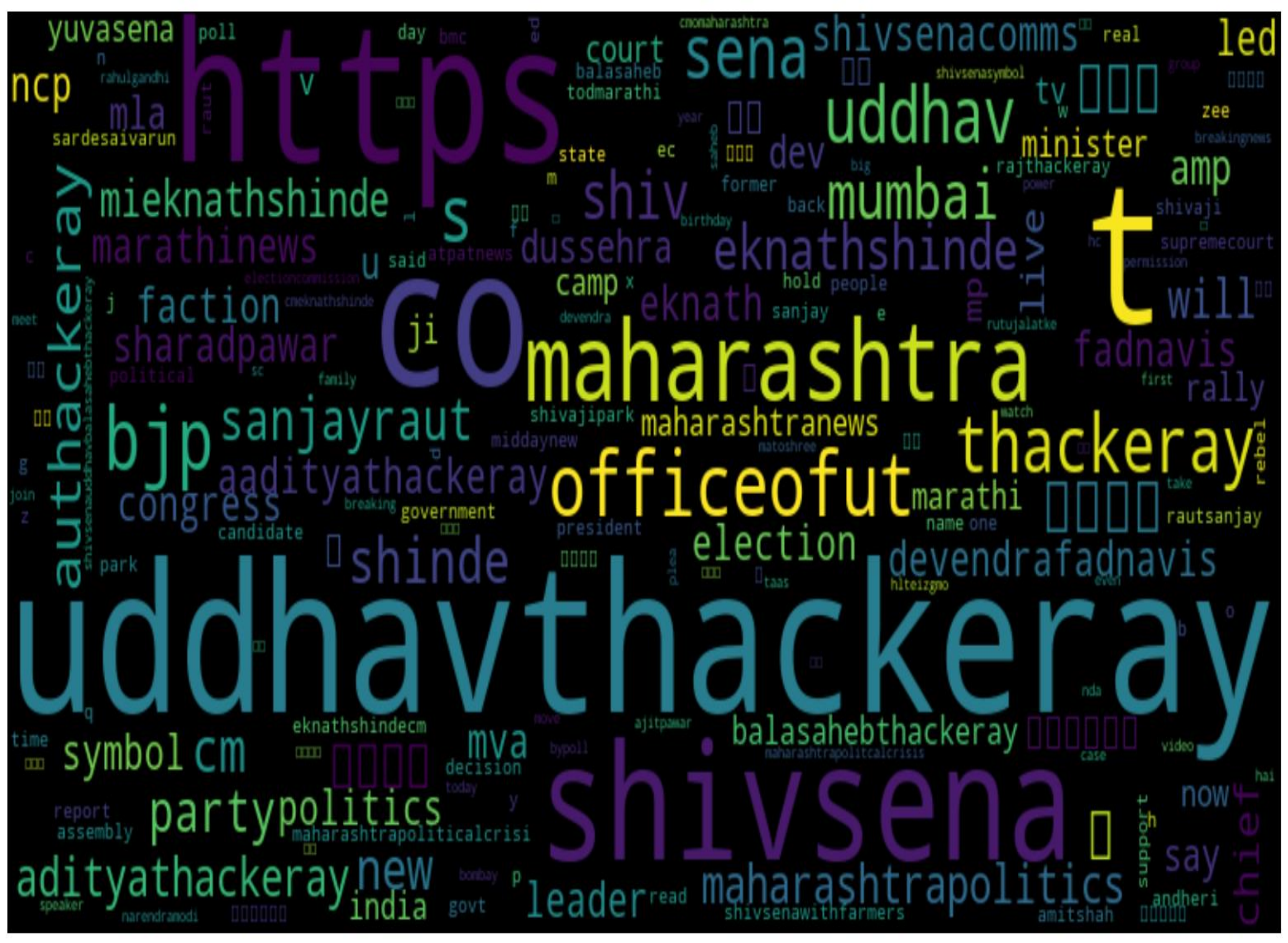
  Method that switches any kind of a word to its base root mode

Word	Stemming	Lemmetization
celebrating	celebr	celebrate
ideology	ideolog	ideology

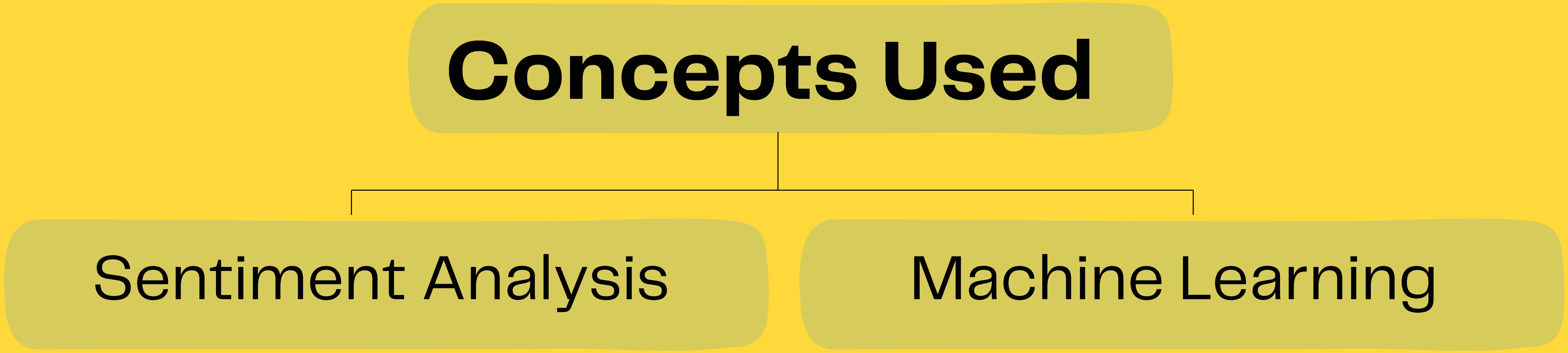
Word Clouds



Contender A



Contender B



Concepts Used

Sentiment Analysis

- VADER

Machine Learning

- SVM
- Naive Bayes
- Decision Trees
- Multinomial Logistic Regression
- K Nearest Neighbours
- Bagging SVM

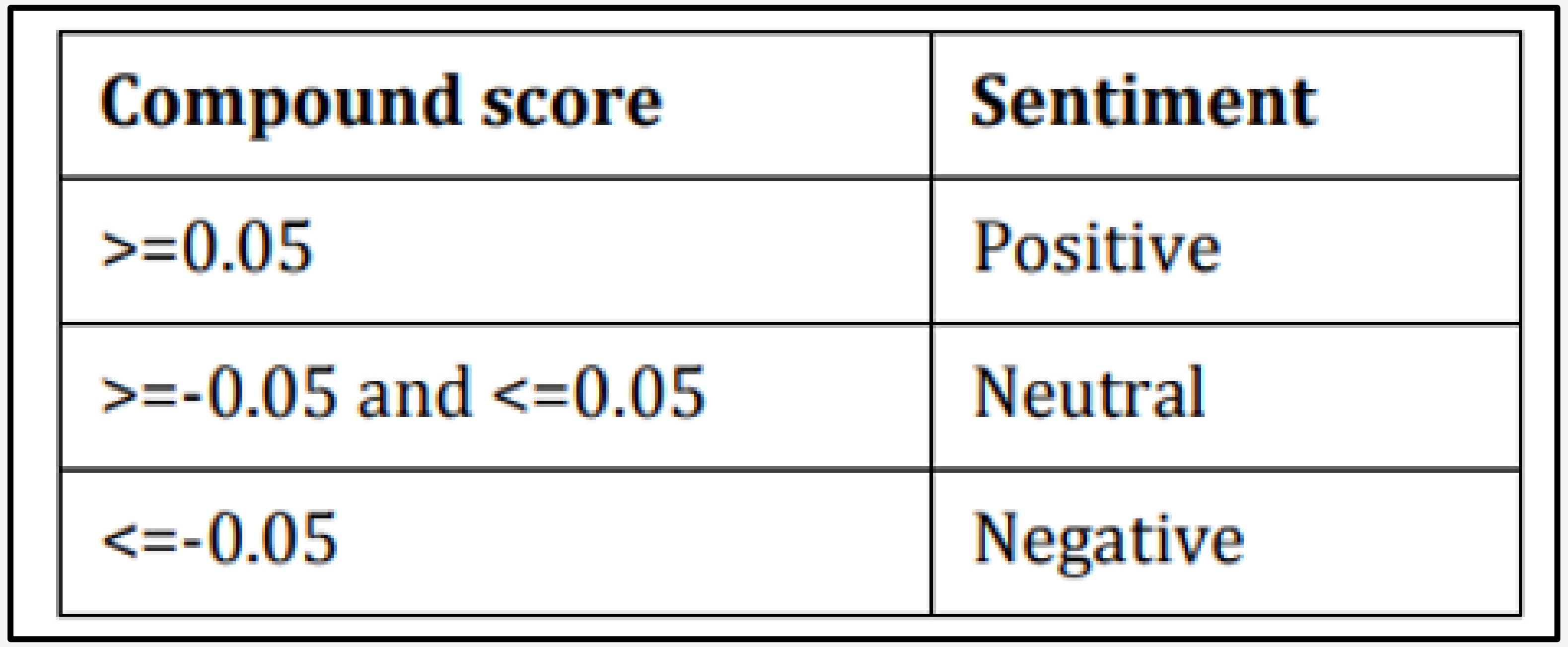
Model Training: Labelling dataset A

Sentiment Analysis

- A natural language processing technique used to determine whether data is positive, negative or neutral
- Becoming an essential tool to monitor and understand sentiment in all types of data

VADER:
Valence Aware Dictionary and sEntiment Reasoner

A lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media

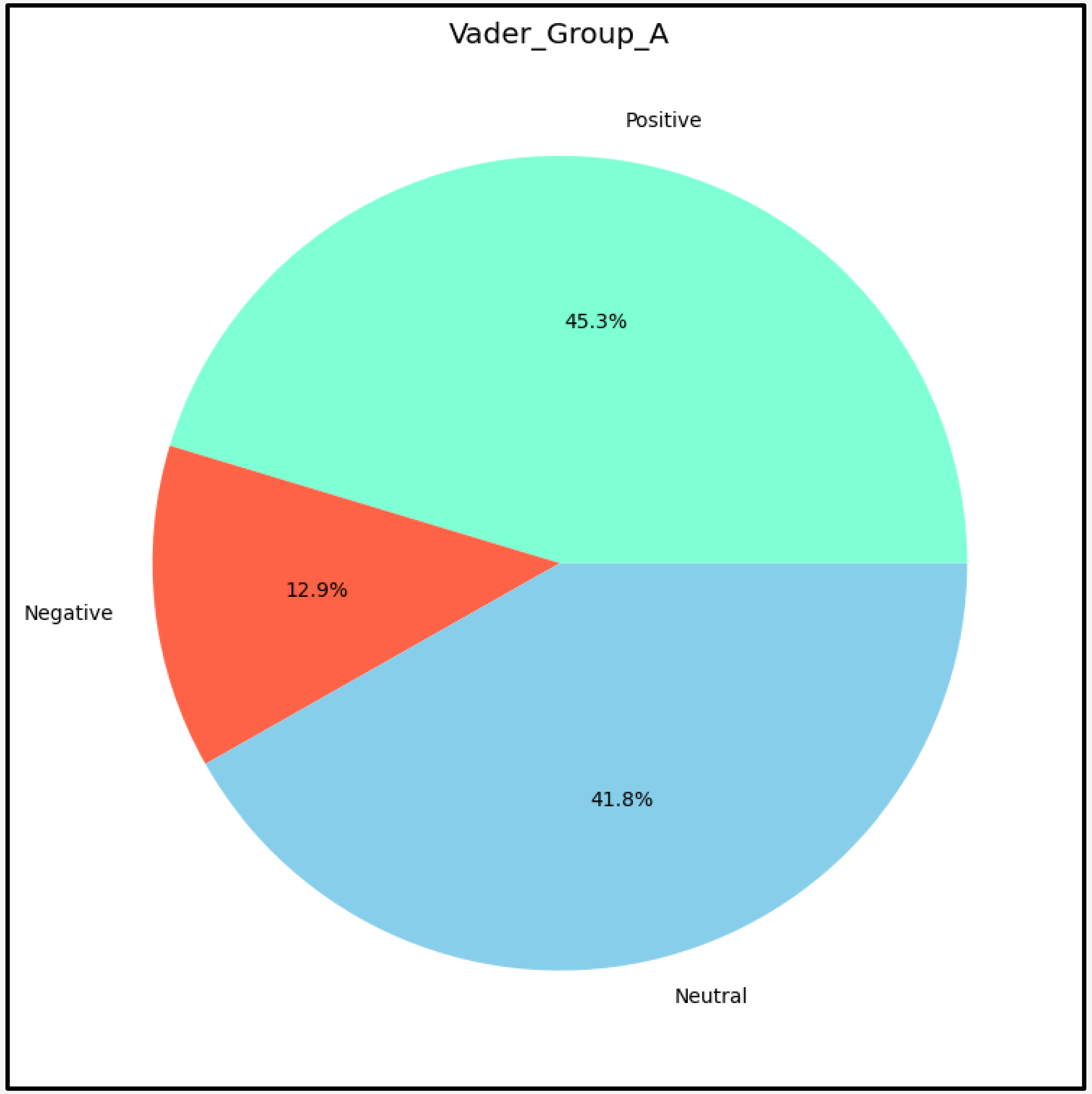


Compound score	Sentiment
>=0.05	Positive
>=-0.05 and <=0.05	Neutral
<=-0.05	Negative

Examples

Sentence	Compound Score
A good leader	0.4404
A really good leader	0.4927
A great leader	0.6249
A terrible leader	-0.4767
went vote today :)	0.4588
went vote today :(	-0.4404
A fine cm	0.2023
A fine cm!	0.2714

Distribution of Sentiments in Dataset A



Train – Test Split

SPLITTING THE DATASETS INTO TRAIN SET AND TEST SET

Train Set	Test Set
70%	30%

tf – idf

FEATURE EXTRACTION

Term Frequency–Inverse Document Frequencies

A technique to quantify words in a set of documents

Computes a score for each word to signify its importance in the document and corpus

The rare words have higher tfidf value and considered important to model training

Word2Vec

FEATURE EXTRACTION

Word2Vec

Employs the use of a dense neural network with a single hidden layer that has no activation function, that predicts a one-hot encoded token given another one-hot encoded token

Capable of capturing context of a word in a document, semantic and syntactic similarity, relation with other words, etc.

MODEL TRAINING

Naïve Bayes Classifier

Classification algorithms based on Bayes' Theorem

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

Predicts on the basis of the probability of an object

Decision Trees

A tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome

Multinomial Logistic Regression

Models the relationship between a set of predictors and a nominal response variable. A nominal response has at least three groups which do not have a natural order

MODEL TRAINING

k Nearest Neighbours

Assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories

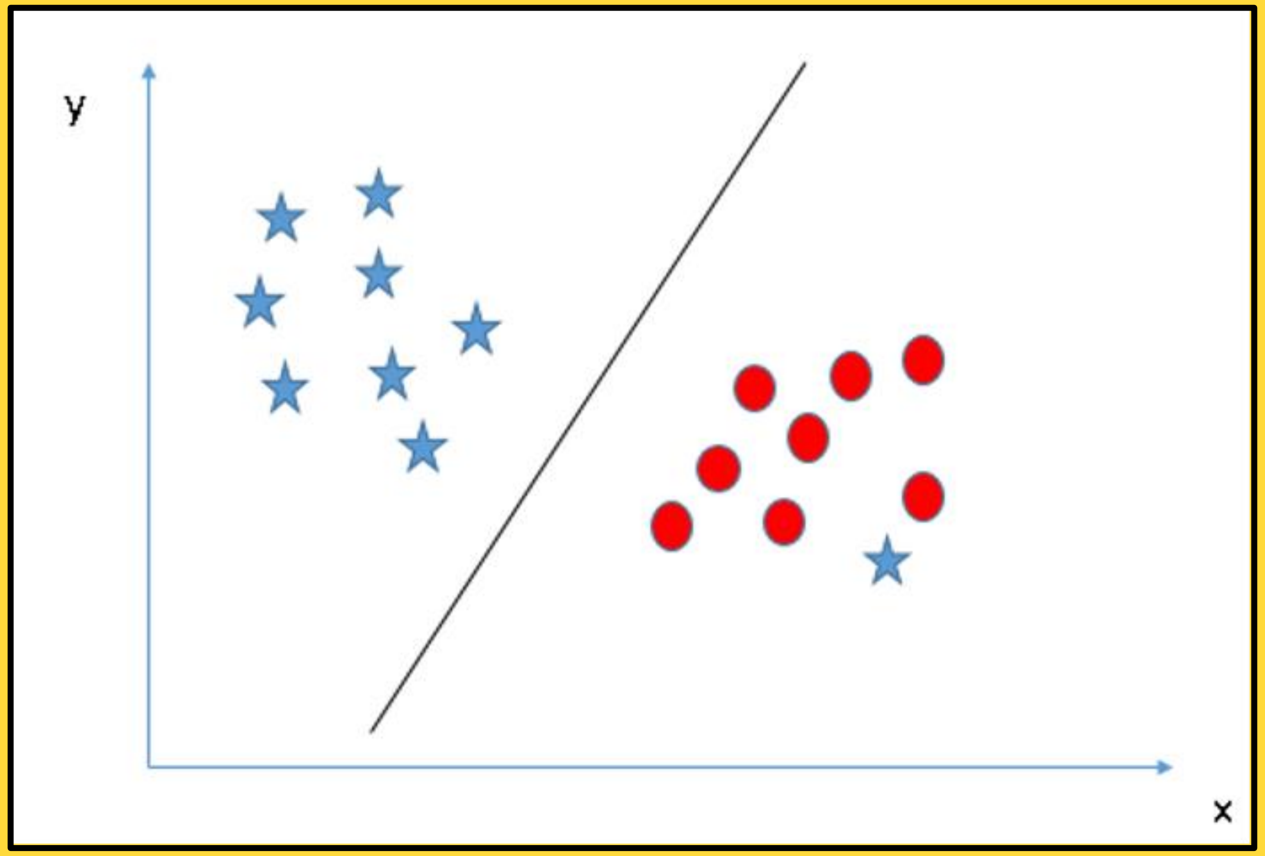
Bagging

An ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction

Model Training: Support Vector Machine

Goal

To create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future



Predictions

Model Performance

tf – idf:

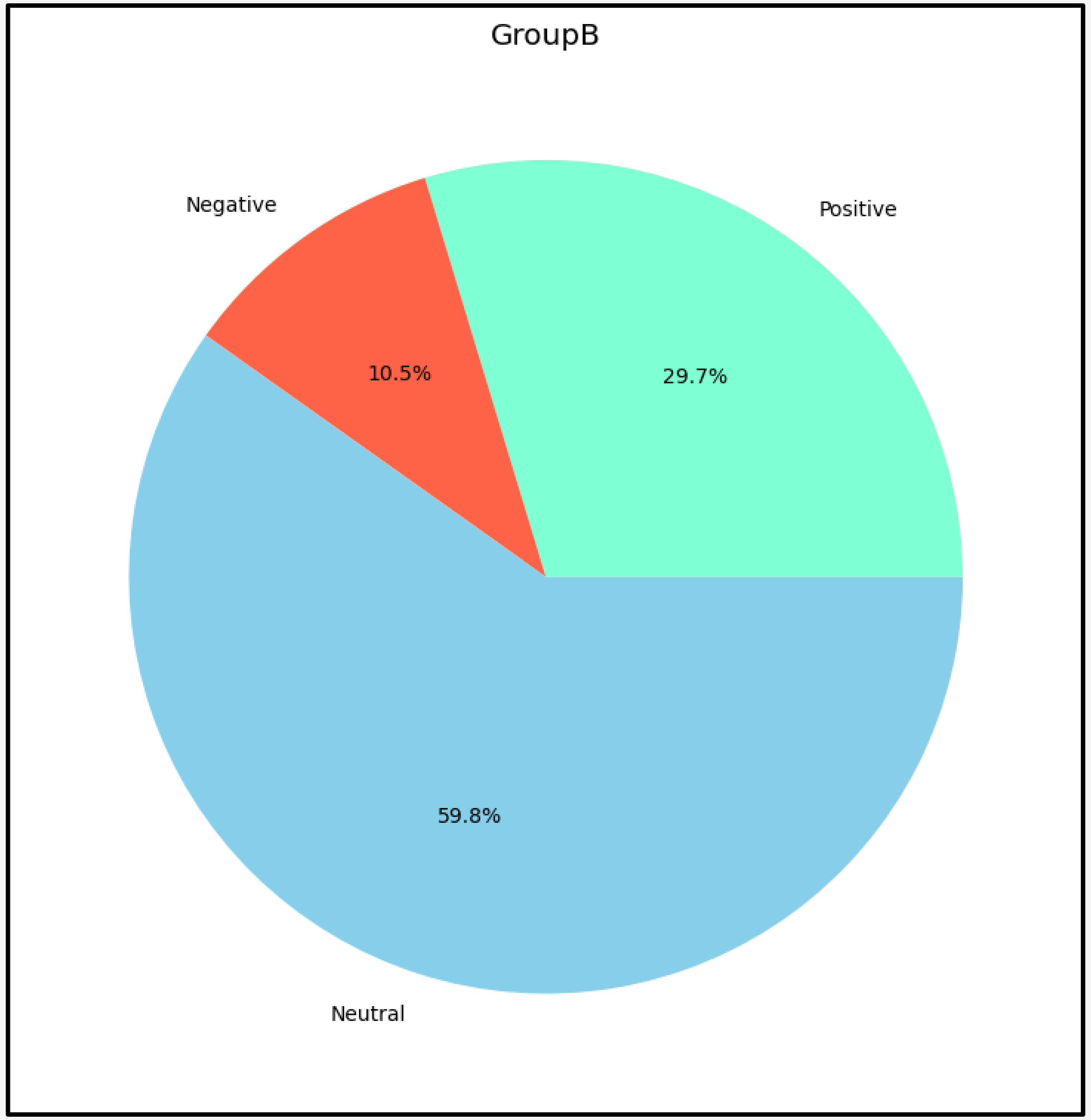
Algorithm	**Accuracy (In %)**
SVM	**86.94**
KNN	76.66
Naïve Bayes	73.22
Multinomial Logistic Regression	84.14
Decision Trees	77.65
Bagging (SVM)	**86.86**

Word2Vec:

Algorithm	**Accuracy (In %)**
SVM	71.13
KNN	69.92
Naïve Bayes	61.65
Multinomial Logistic Regression	71.25
Decision Trees	65.49
Bagging (SVM)	71.07

Predictions

Dataset B using SVM (tf – idf)



Prediction of Winner

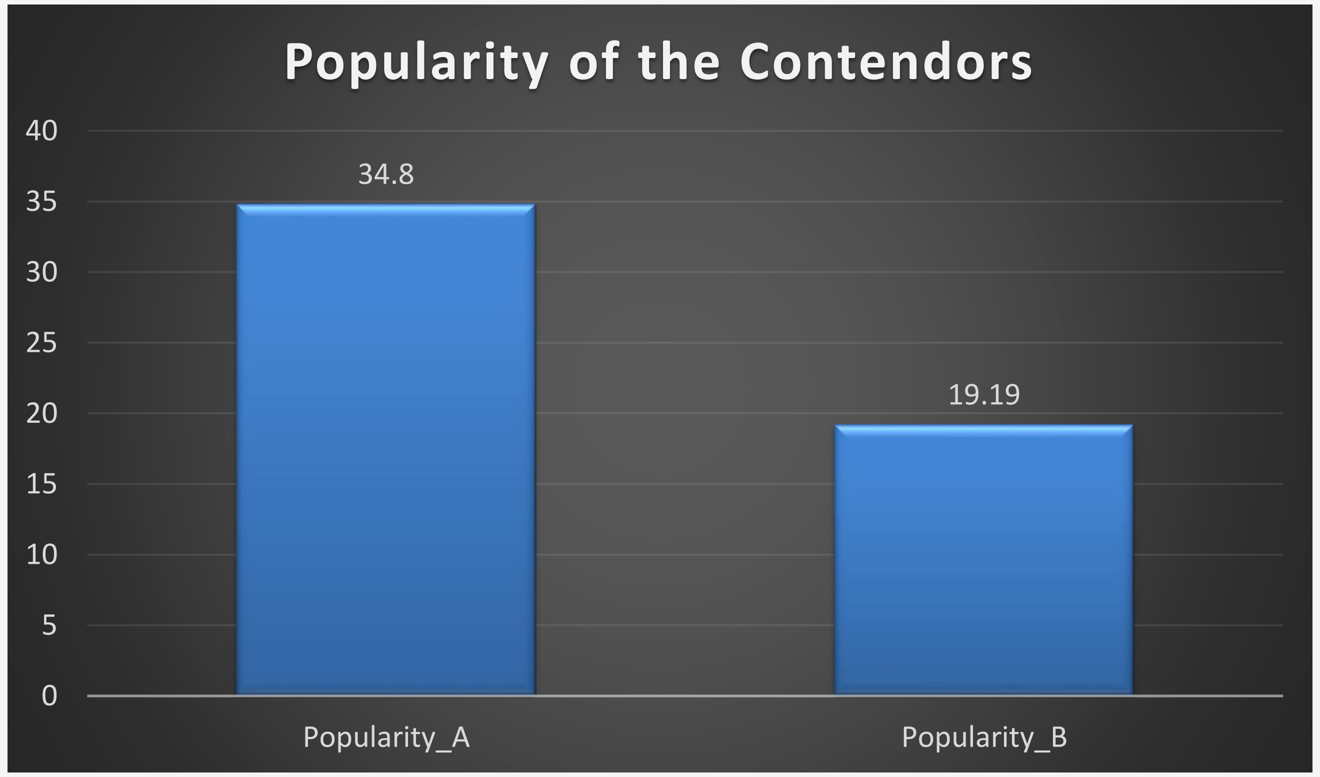
Calculating Popularity Scores

Popularity Score:

$$\frac{(\sum \text{Tweets with positive sentiment} - \sum \text{Tweets with negative sentiment})}{\text{Total Tweets over the time period}} \times 100$$

Group A – 34.80
(SVM tf-idf)

Group A – 32.46
(Using VADER)



Group B – 19.19
(SVM tf-idf)

Group B – 9.39
(Using VADER)

Conclusion

- **Best Model:**

  SVM tf-idf with an accuracy of **86.94**

- **Group A is a clear winner Based on the popularity score**

Recommendations

Use case for political parties

The proposed system can be used by political parties to improve their campaigning strategies during the election period

Political analyst and strategist can use this methodology, as application, as a long term plan for a political party to study the sentiments of people over a long time period

Use case for the people

Can be used by users to make informed decisions in voting by seeing the current trends of political parties

Future Scope

Geographic Locations and work profiles of users can be taken into consideration

Other Hashtags can be used to improve analysis

Native languages were not translated

Sarcasm was not detected

Events taking place closer to the Election date would influence the results more, hence doing the analysis of tweets closer to the election date is required

Thank you!

Vote Responsibly!