SENTIMENT ANALYSIS OF TWEETS

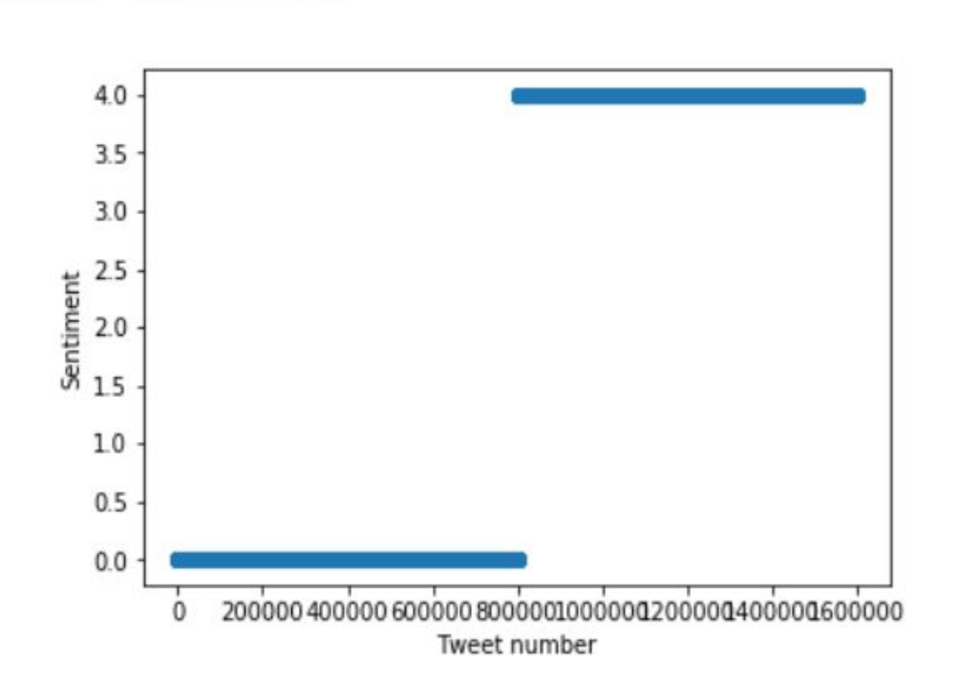
Sopnesh Gandhi (2018201064)
Rushitkumar Jasani (2018201034)
Priyendu Mori (2018201103)
Niharika Khare (2018201002)

APPROACH

1. Collection and analysis of dataset.
2. Preprocessing of data.
3. Extracting features from cleaned tweets.
4. Model Building.
5. Performance comparison.

Dataset Analysis

We have dataset of 1.6M tweets with data splitted equally among positive and negative class.



Pre-processing

1. Decoding HTML.
2. Removing username and tickers.
3. Removing hyperlinks.
4. Removing words of length less than two.
5. Removing punctuations.
6. Stemming. (ex:- changing playing to play)

Used regex to achieve above written things and Implemented porter stemming algorithm.

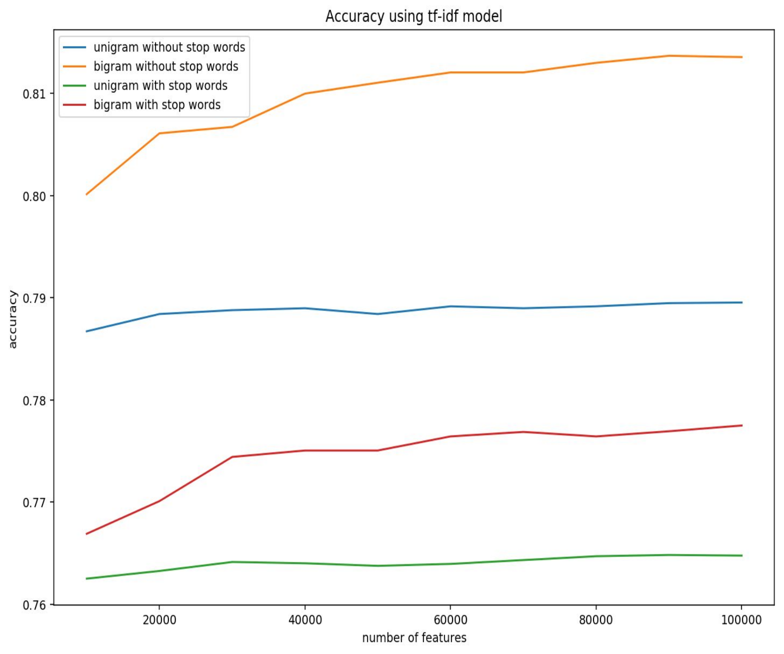
Model Building

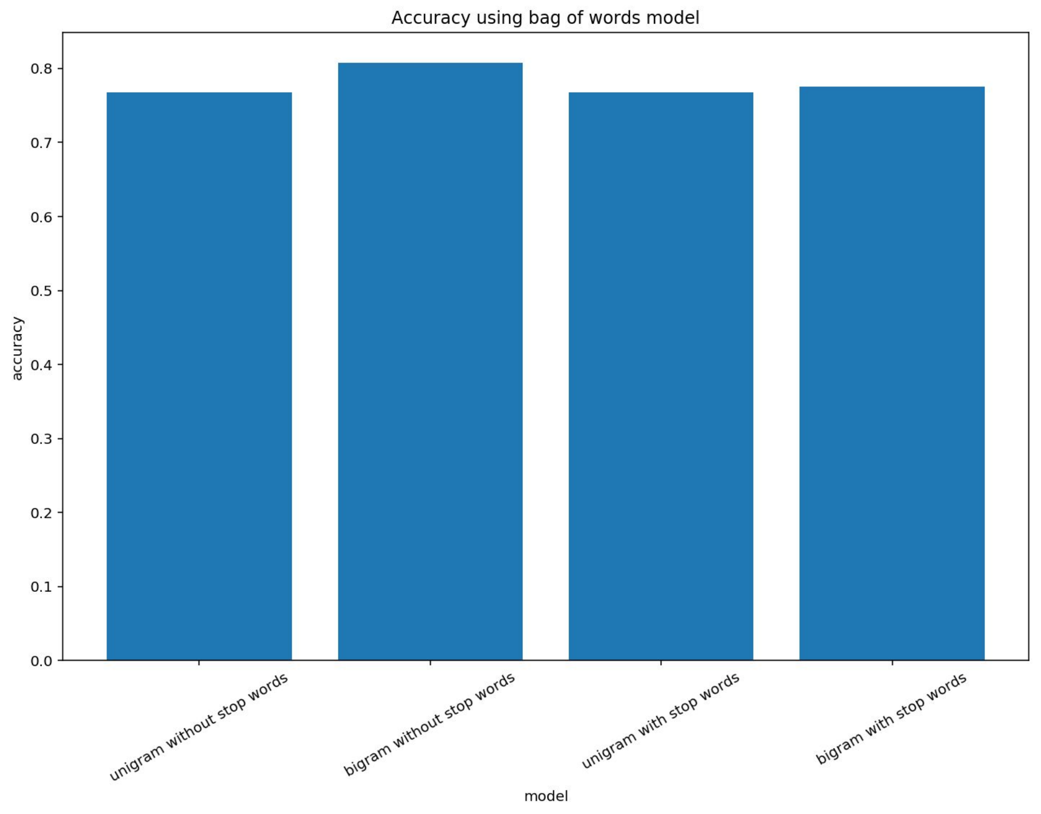
1. Naive Bayes
2. Logistic Regression
3. SVM

Trained each models considering:

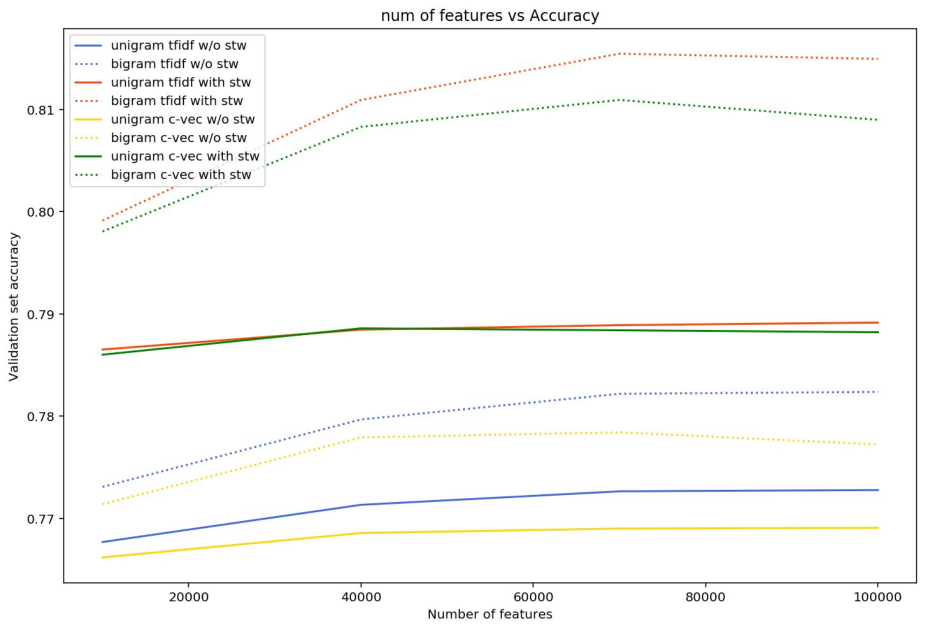
1. Bag of words and TF-IDF
2. Unigram and Bigram
3. Keeping and removing stop words

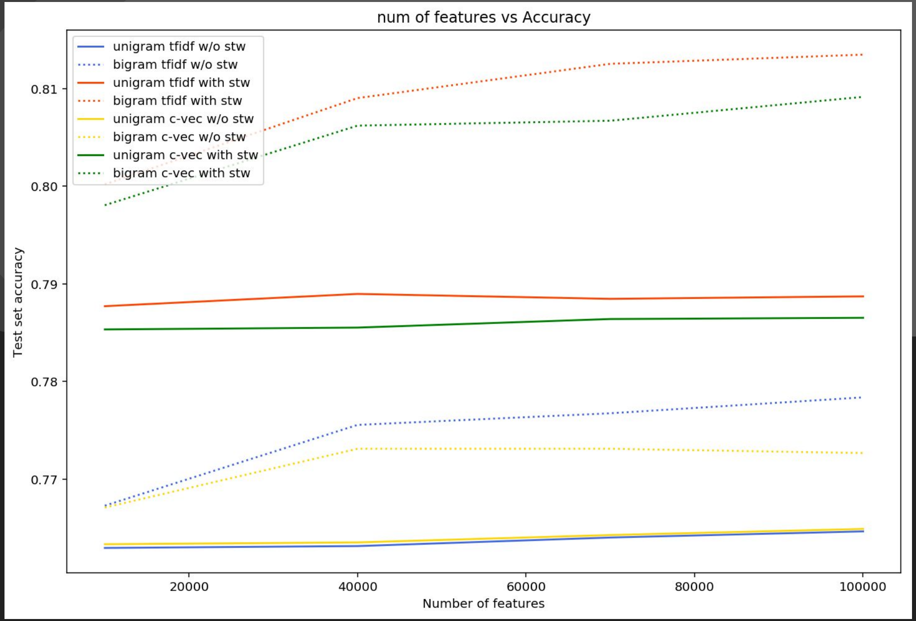
Performance of SVM



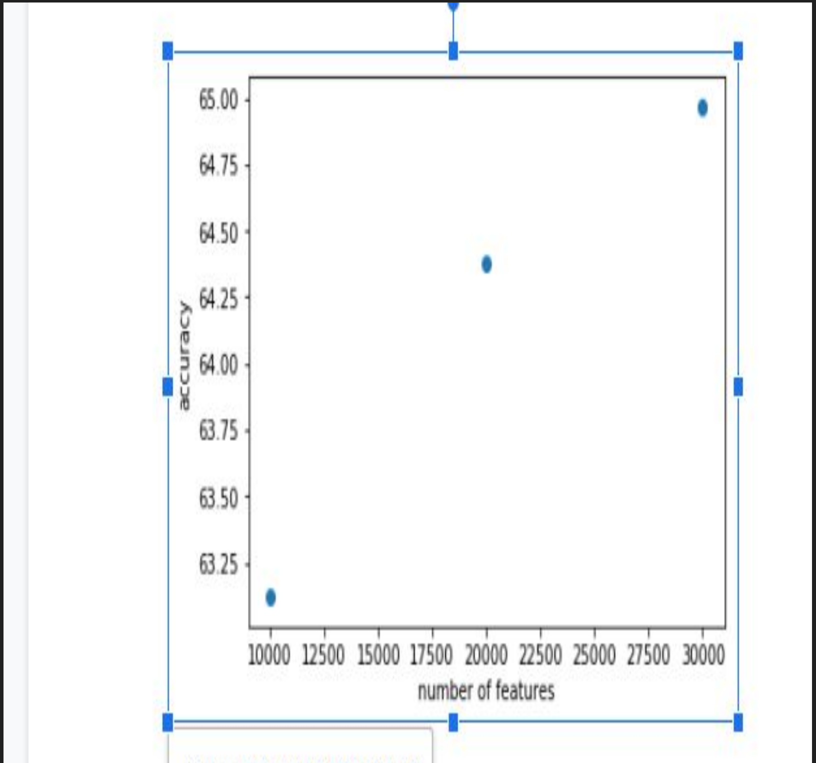


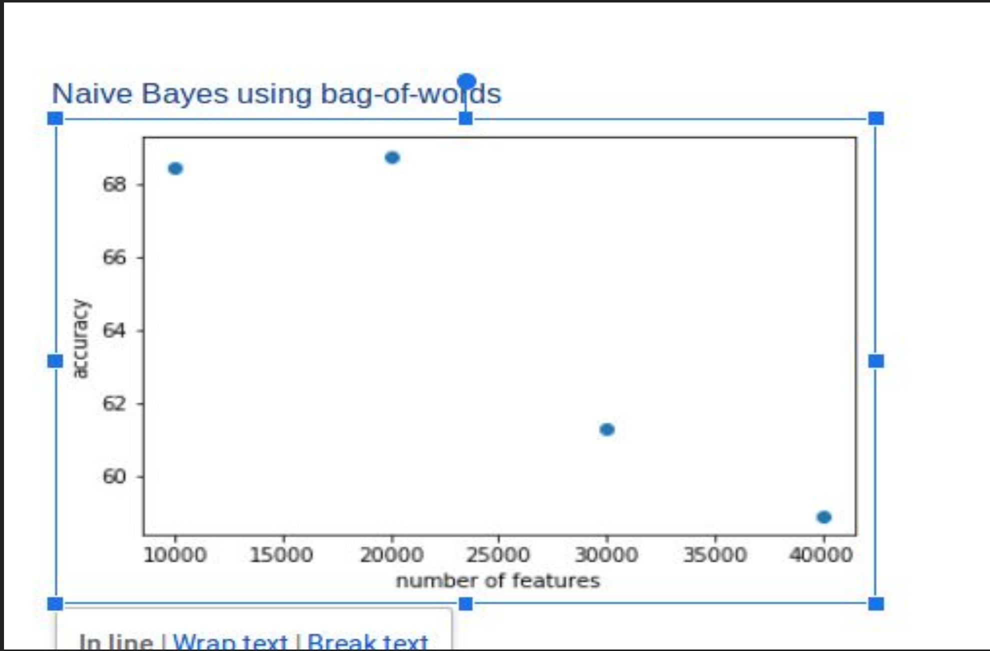
Logistic Regression Performance





Naive Bayes Performance




Naive Bayes using bag-of-words

Testing

- Tested on twitter data.
- Tested on facebook comments.
- Tested on amazon reviews.

Last two are add-ons and were not asked to implement.

Performance on FB comments and Amazon reviews

	SVM	Logistic Regression
Facebook comments	84.00	89.15
Amazon reviews	77.52	72.10

Challenges

1. Understanding how to work with text and pre-process data.
2. Implementing stemming and other functions that are inbuilt in NLTK etc.
3. Deciding the models for training.
4. Working with huge data leads to huge training time especially on a low computing device like a personal computer.

Thank You