Sentiment Analysis of Presidential Speeches using Natural Language Processing

Sreeram Sankarasubramanian
M.ENG. Electrical and
Computer Engineering
Carleton University
Ottawa, ON, Canada
Email: sreeramsankarasubram@cmail.carleton.ca

***Abstract*—Sentiment analysis has evolved over past few decades, most of the work in it revolved around textual sentiment analysis with text mining techniques. However, audio sentiment analysis is still in a coming to existence stage in the exploration network groups. In this proposed paper, we perform sentiment analysis on president speeches on State Union addresses from the year 1970-2019 using Natural Language Processing (NLP) to detect the impact of speech on public as positive or negative sentiment. We analyzed different speeches and tried different classification techniques to predict accuracy.**

***Index terms*—** Sentiment Analysis, State Union addresses, NLP, classification, Accuracy.

I. Introduction

Elections make a fundamental contribution to democratic governance. Because in a direct democracy—a form of government in which political decisions are made directly by the entire body of qualified citizens is impractical in most modern societies, democratic government must be conducted through representatives. Elections enable voters to select leaders and to hold them accountable for their performance in office. Nevertheless, the possibility of controlling leaders by requiring them to submit to regular and periodic elections helps to solve the problem of succession in leadership and thus contributes to the continuation of democracy.

Elections reinforce the stability and legitimacy of the political community. Like national occasions recognizing common experiences, elections connect residents to one another and along these lines affirm the viability of the polity. As a result, elections help to facilitate social and political integration. Elections provide political education for citizens and ensure the responsiveness of democratic governments to the will of the people.

Prediction is this extremely significant, and it's fundamental to science, to see whether emotional reality coordinates with the objective world. But it's not really a carnival show game. People have been asking sometimes to predict Kate Middleton stuff, and things like that. It's not really the idea, right? This stuff's hard."

US elections are predictable,for a few reasons. First, there is masses of polling data: there are regular nationwide polls and also, unlike in the UK, polls for each individual state. "If we would have only had national poll data, we'd still have had Obama ahead, but with much less confidence,". "We'd have had him winning by half a point instead of 2.5 points."

Natural language processing is amazing to the extent that it helps machines communicating with people by speaking with people in their own language and scales other tasks related to language. For instance, NLP causes it feasible for machines to understand the content and decipher it, measure estimation and figure out which parts are significant. The present machines surprises us by analyzing more information based on language than people, without exhaustion and in a predictable, impartial way. Considering the amazing measure of unstructured information that is created each day, from election manifests to social media tweets, computerization will be basic to completely break down content and speech information productively. NLP is significant on the grounds that it helps settle equivocalness in language and adds valuable numeric structure to the information for some downstream applications, for example, speech recognition or data analysis.

Sentiment Analysis is the way to identify the tone and emotions expressed through written or spoken online communication. Otherwise called sentiment mining or emotional AI. Sentiment analysis performs information mining, forms the outcomes, separates popular assessments out of the content and lets you reach the important resolutions. Sentiment Analysis is the investigation of individuals' feelings or mentalities towards an occasion, discussion on subjects or all in all. Sentiment analysis is utilized in different applications, here we use it to get a handle on the effect of President Speech dependent on their State Union addresses to the public each year. For a machine to comprehend the effect of the president's speech on public, it has to realize what is spoken, so we actualize a web scraping framework first to peruse the substance of addresses through URL and perform sentiment analysis on the information collected from earlier procedures.

Understanding the impact of the president speech on public can be very useful in many instances. For example, computers that possess the ability to perceive and predict before election which candidate will have better chance of winning will be based on impact of candidate speeches on public. In such a case, after detecting impact of speeches, the machine could customize the settings according his/her needs and preferences and can suggest user about usage of unique words, word count and sentence length etc.

The specialist's affiliation has tackled changing sound materials, for instance, tunes, news, political speeches, to text content. With the advancement in technology, audio speeches are converted to text and are available online. So, in this paper we propose a system which reads and interprets the speeches from a web source by making use of web scraping tools like Request and Beautiful Soup package from Python library. We use Natural Language Processing tools to perform Tokenization, Stemming, Lemmatization and Negation handling on the obtained data after web scrapping the speeches from URL source. Further, sentiment analysis is performed on the speech information which empowers the machine to comprehend what the Presidents were discussing and what sway people, in general, may feel.

The background and related work of sentiment analysis is discussed in segment-II. Segment-III contains a clarification about the proposed framework. Segment-IV deals with insights concerning the exploratory arrangement and Segment-V talks about the various results obtained. The conclusion drawn from this project is explained in Segment-V.

II. Related work and background

A numerous amount of literature survey dedicated to develop new tools and technologies for sentimental analysis.

Sentiment Analysis which recognizes the sentiment communicated in a text and then examines it to discover whether the speech communicates positive or negative feelings. In general work on the analysis of sentiment has concentrated on techniques, for example, Naive Bayes, Decision Tree, Support Vector Machine, Ridge classifiers [1][2][3]. In the work done by Mostafa et al [4] the sentences in each archive are named as abstract and target and afterward old-style AI strategies are applied for the emotional parts. With the goal that the extremity classifier overlooks the superfluous or misdirecting terms. Since gathering and marking the information is tedious at the sentence level, this methodology isn't anything but difficult to test. To perform an assumption analysis, they have utilized the accompanying strategies – Naive Bayes, Linear Support Vector Machines, VADER [5]. what's more, a correlation is made to locate a proficient calculation for their motivation.

In paper [6] they did a sentiment analysis on the data taken from twitter. They utilized cutting edge uni-gram model as a benchmark and detailed a general addition of over 4% for two grouping errands: a parallel, positive versus negative and a 3-way positive versus negative versus unbiased. They introduced an exhaustive arrangement of investigations for both these undertakings on physically clarified information that is an irregular example of a stream of tweets. Later researched two sorts of models: tree part and highlight based models and exhibited that both these models outflank the unigram benchmark. For their element based methodology, they featured analysis which uncovers that the most significant highlights are those that join the earlier extremity of words and their grammatical forms labels. Creator probably reasoned that conclusion analysis for Twitter information isn't that not quite the same as sentiment analysis for different sorts.

Sentiment analyzer (Jeonghee etal.,2003) [7] works by extracting sentiments about given topic. Sentiment Analysis comprises of a topic-specific feature term extraction, sentiment extraction, and relationship by relationship analysis. Sentiment Analysis uses two semantic assets for the examination: the sentiment lexicon and the sentiment design database. It analyses the documents for positive and negative words and attempts to give evaluations on the scale - 5 to +5.

Natural language processing (NLP) has as of late increased a lot of consideration for speaking to and dissecting human language computationally. It has spread its applications in different fields, for example, machine interpretation, email spam recognition, data extraction, synopsis, clinical, and question replying, and so on. [8] Right now recognizes four stages by talking about various stages of NLP and parts of Natural Language Generation (NLG) trailed by introducing the history and development of NLP, best in class introducing the different uses of NLP and current patterns and difficulties.

In this paper [9], they introduce a set of new approaches for text representation for automatic classification of Arabic textual documents. These approaches are based on combining the well-known Bag-of-Words (BOW) and the Bag-of-Concepts (BOC) text representation schemes and utilizing Wikipedia as a knowledge base. The proposed representations are used to generate a vector space model, which in turn is fed into a classifier to categorize a collection of Arabic textual documents. Three different machine learning based classifiers have been utilized in this work. The exhibition of proposed content portrayal models is assessed in contrast with utilizing a standard BOW conspire and an idea based plan, just as of late revealed comparative content portrayal plots that depend on increasing the standard BOW with the BOC.

Late advances in machines and innovation brought about an ever-expanding set of records. The need is to characterize the arrangement of archives as indicated by the sort. Laying related records together is convenient for dynamic. Analysts who perform interdisciplinary research gain stores on various subjects. Grouping the vaults as per the subject is a genuine need to dissect the research papers. In this paper [10], Experiments are taken a stab at various genuine and counterfeit datasets, for example, NEWS 20, Reuters, messages, look into papers on changed points. Term Frequency-Inverse Document Frequency calculation is utilized alongside fluffy K-implies and various leveled calculations. At first, the trial is being completed on a little dataset and performed a bunch analysis. The best calculation is applied to the all-inclusive dataset. Alongside various bunches of the related records the came about silhoutte coefficient, entropy and F-measure pattern is introduced to show calculation conduct for every data set.

In this paper [11] they proposed a sentiment classification model to classify customer reviews into positive and negative reviews using ensemble machine learning method. The ensemble machine learning combined between five Classifiers they as follows Naive Bayes, Support Vector Machines (SVM's), Random forest, Bagging and Boosting.The proposed model is done using WEKA too, which have certain limitations. For word removal they are using uni-gram, bi-gram and tri-gram techniques. This approach provides greater accuracy and diversified sentimental polarity.

In this paper [12] the author represented noun as sentimental words and which have a good impact on sentiment detection. In addition to that some words has duel sentiment base on its application, mostly those words are a noun. Yet the accuracy is less compared to other machine learning approach. There needs to be some improvement done on the duel sentimental analysis to increase the performance of the model.

In this paper [13] they have built a system which analyses Sentiments based on historical data to forecast prices over a future selling period and then to use the dynamic pricing model to increase the revenue generated. Here the raw data from tweets are converted into JSON format where the parsing of data can be done at an ease, which makes the prepossessing much efficient. The scoring mechanism developed in the model and the time series which is generated are not efficient to predict sentiments of the matched tweets.

In this paper [14], they have used the techniques and methods involved to perform sentiment analysis are RNN algorithm and NLP, to improve the competence power and accuracy of the model they have introduced Stanford library. Google translator is being used here to remove the linguistic issues. There is a major drawback with the google translator, the accuracy of which the translator algorithm works is seemingly low and it needs some improvement. Moreover there would be delay in the process which might affect the overall performance of the model.

In order to evaluate the reliability of e-commerce products from subjective aspect more properly and easily, in this paper [15] they have developed an evaluating procedure based on weighted sentiment analysis of products' comment content. In this procedure, they have calculated Sentimental Value and Usefulness for each comment to obtain the Subjective Reliability value of a certain group of e-commerce product. The massive user data lessen the performance of neural networks, its index parameters. The calculation methods which are used here still needs improvement. Here while calculating the reliability, picture comments are not analyzed and identified. The reliability evaluation of the product with fewer comments is less accurate.

In this study [16], they have done the research of the sentiment analysis of review text for online micro video by using big data analysis, to predict the type of the user's favorite video based upon reference value from the algorithm. This algorithm will be useful to analyze the users favorite genres and suggest videos based on the data analyzed. But the downside is there are few more parameters which needs to be considered like Impressions, Traffic sources for impressions, Views from impressions, Watch time from impressions to efficiently predict the user desired videos. This approach is currently implemented in the YouTube algorithm.

This paper [17] tackles a fundamental problem of sentiment analysis, sentiment polarity categorization. The data for this research is collected from Amazon.com product reviews. Sentiment polarity classification and POS tagger have been proposed to improve the performance. There are more innovative and efficient machine learning techniques which can be used instead of this model to over come the performance issues faced with the opinion mining based on the POS taggers.

In this paper [18], they have developed the concept and extraction method of sentimental context, and have proposed two models for short text sentiment classification (i.e. Sentimental Context Term Model, C. Sentimental Context Topic Model) by integrating the sentimental context. The results shows that the sentimental context helps to improve the performance of sentiment classification. This model has proved it performance in lesser number of data, but the topic based models will get affected by the number of selected topics.

In this paper [19] the sentimental analysis is calculated by collecting data from five companies i.e. Oracle, Microsoft, Google, Apple and Facebook in the form of tweets and the user comments from the largest community of investors and traders website, Stocktwits. Here the author fed the sentimental score with market values to an artificial neural network, Levenberg-Marquardt algorithm and used mean square error to calculate errors and to predict the future market values. The use of Artificial neural network is the major advantage in this paper, based upon the number of neurons used the efficiency varies. But I highly doubt the precision of this approach for a wide variety of data, which would be its major disadvantage.

To improve the overall performance of the analysis, in this paper [20] the author takes a probabilistic approach in machine learning techniques by using Naive Bayes Classifier for the amazon product review dataset. The main advantage of this classifier is that it is mainly based on the prior and posterior probability. The word occurrence is mainly considered for this method. Still the accuracy of 89% can be improved by doing proper pre-processing and adding various datasets.

To anticipate the future extremity of the organization in this paper [21] they have utilized the KNN calculation to store all the accessible cases and classifies the new cases dependent on the likeness measures. The main advantage of this approach is the categorization for that they have used non-probabilistic binary linear classifier – SVM algorithm and for sentimental analysis they have used tweets to classify the positive and negative words. All three of these results can be used by the investor to get the accurate result about the next day's trends before buying or selling of company's stocks. The disadvantage is the use of three more parameters which has to be separately analyzed for the market prediction.

To calculate the sentimental analysis in this paper [22] they have used the tweets from twitter as an input data and used the Stanford core NLP which provides a set of natural language analysis tools to predict the sentimental value. The main disadvantage is the lack of pre-processing and the limitation of the Stanford library.

In this paper [23], they have proposed a new word sentimental similarity calculation method to compute words sentimental value with the modified HowNet knowledge, on the basis of existing primitives which is combined with transductive learning for judgment words sentimental orientation, which is the main advantage. The performance of this model is far superior to that of SVM and traditional semantic comprehension.

In order to identify eight kinds of sentiment like Joy, love, sorrow, disgust, surprise, anxiety, anger, hate for the Chinese language reviews, in this paper [24], they have used a sentimental agent identification based on the Chinese sentimental sentence dictionary. The sentence dictionary is consists of sentence patterns which can be used to calculate the consistency of the conversations and can easily get rid of the sentences without sentiment. The model totally depends upon the consistency of the sentimental sentence pattern which needs improvement.

In this paper [25], they are using the traditional classifier Naive Bayes Classifier to get a set of positive, negative and neutral sentences which are used for feedback. Then the clustering operation is done within the positive and negative feedback, which is used to determine the broad topics like quality of foods, ambiance, Service, on which the feedback has been obtained. For that reason K-means Clustering is utilized here. K-means clustering algorithm has a certain drawbacks, it needs the number of clusters to be specified initially and it has a great dependency on initial cluster center, which is unstable and gives inaccurate results and it is also sensitive to noisy data. In order to address these limitations they have modified K- means clustering with dynamic threshold, which helps in creating clusters dynamically depending on the dataset.

In this approach [26], the tweets from twitter are continuously downloaded via the streaming API and converted to JSON and sent to Flume sink. The tweets are fetched from Twitter using Apache flume on the basis of the keywords provided by the user. They have collected tweets related to IPL in order to discover the public opinion and rating about IPL players and its matches. In continuation of sentimental analysis, the hashtag analysis is done here which helps in categorizing the tweet's topics. Furthermore it optimize the tweets and these words get maximum exposure. Followed by count analysis which helps the analyst in knowing about the impact the person could create using Twitter.

In this paper [27], the author collected number of tweets and Facebook posts and stored them into JSON which can be used as a data model. Here NLTK library is used for pre-processing of the data. The ontology model can classify tweets with the negative sentiments to do sentiment analysis. SentiStrength tool is used to identify the tweet with the negative sentiments. The tool can identify the polarity of words in the tweet sentences. The Semantic Orientation (SO) of a word as the distinction between its relationship with negative words. Followed by applying Naive Bayer algorithm and the fuzzy functions is used to calculate overall sentimental score which is achieved by using linguistic variables like more negative, less negative, negative. The combination of these approaches gives a good accuracy. The problem arises with the reliability of the fuzzy function.

In this paper [28], experiments uses POECS(Platform for Opinion Extraction, Classification and Summarization) and CSV file to link opinion words to their orientation. Here they have used WEKA tool to collect and categorize opinions reviews about a product. The limitation of WEKA tools and the lack of proper pre-processing affects the overall performance of this model.

In this paper [29], they have used a simple approach and used twitter tweets and comparing with files containing a dictionary of positive and Negative words. The sentiment score is calculated by considering the positive and negative words used in the tweets, these calculations can be used to do the sentimental analysis. But there is a major drawback where the sarcastic conversations cannot be identified.

Electoral results can be predicted by analyzing the social media feeds, in this paper [30], they are using the concept of decision tree to show the output with the help of tree and nodes, it is a simple classifier which helps in text mining. Naive Bayers also used here and the final results are compared between decision tree and the Naive Bayes classifier. Quite often for visual image recognition CNN is used. It wont be suitable for analyzing the words because most of the online comments are short texts and always have character limitation. The word embedding produced by unsupervised pre-training using Word-Level Embedding representation is high-dimensional and sparse. Unlike image RGB information, adjacent point in word vector has not strong correlation that is the major disadvantage of this model.

III. Proposed Solution

The main scope of this project is to develop an efficient sentimental analysis using NLP algorithm. The type of data set chosen here is from online resources like wikisource and Kaggle. The proposed solution uses python NLTK as the main library, in which we are going to analyze different speeches on the dataset which we haven chosen, we are going to use NLTK python library to find the sentimental score (positive and negative emotions) on each speeches which we have done in the first step. Then we are going to use the classifier like Naive Bayes, AdaBoost and Random Forest. Once the model has been developed from the dataset we will be able to efficiently strategize and plan speeches based on sentiment it assigns for a speech.

IV. Problem Identification

From the literature survey it is evident that all the papers solely focus on the opinion mining, but not on the sentimental analysis of multiple feature classification, which targets the real time issue faced when classifying president speeches. In real time scenario, a speech might have thousands of paragraphs and public mood might have different opinions about each sentences.The overview of the project has been explained in the below diagram.

Moreover from all those papers it is evident that the datasets used are from the social media sites like Twitter, Facebook and e-commerce sites like amazon, where the data is collected and

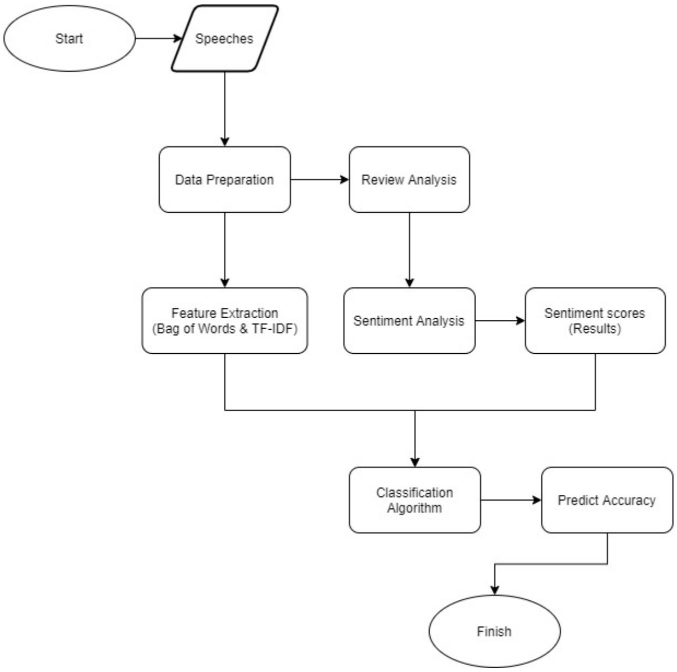


Fig. 1: Block diagram for sentiment analysis

NLP models are implemented on that to find the statistical opinion analysis of about a certain feature. But here we are trying to implement our sentimental model for a real time President speech data from the state union addresses and analyze based on the same. Initially for our project we are going to use the data from the online resources, once the model is stable we can use it to make the sentimental analysis.

V. Scientific Reasons

To improve the performance and to create a highly efficient model to do sentimental analysis for the real time President speech data by using machine learning approaches, which helps in improving our machine learning results by combining several classifiers. We are utilizing this technique in our predictive model so as to diminish differences. To efficiently mitigate the feature classification we are using different classifiers such as Random Forest tree, AdaBoost and Naive Bayes algorithm. Here for two feature classification one can go for Naive Bayes, for multiple feature classification Random forest tree is more suitable. Here we are going to experiment with both of these classifiers and compare the efficiency to choose our ideal model.

These classifiers helps in determining the speech sentiments given by Presidents, (positive and negative opinion). To implement this feature we are first doing the text classification (positive and negative words) by using NLTK library, which is widely used python library. The diagram in [Fig 2] shows the steps involved in sentimental analysis.

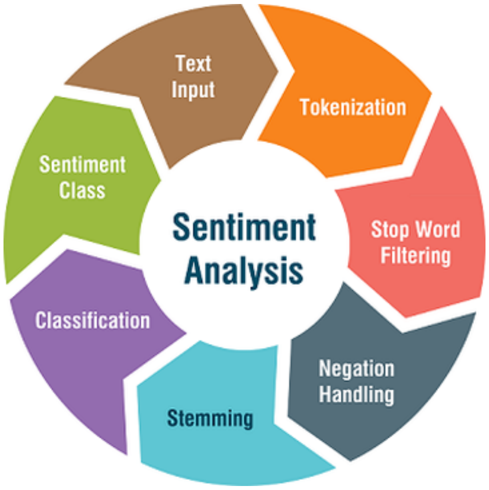


Fig. 2: NLP methods in sentiment analysis

VI. Experiments and Methodologies

A. Datasets

The dataset chosen for this project is from online resource through kaggle, wikisource and Miller Center. It is President speeches on State Union Addresses and Annual messages dataset which describes the President speech given every year from 1970-2019. When I went through different speeches in the dataset, I found that tone in speech was different and sentence length of each speech varied too. The opinions on different speeches will be interesting to calculate on such datasets. Once Data is cleaned and processed using Natural Language Processing techniques, First we apply sentiment analysis to predict impact of psoitive or negative sentiment on our speeches and then evaluated using classifiers and accuracy is predicted with help of bag of words and TF-IDF.

B. Tools and Setup

Platform	Windows, Linux
Tools	Anaconda distribution, PyCharm, Jupyter Notebook and google colab
Language	Python
Libraries	Numpy, Matplotlib, Seaborn, NLTK, urllib, bs4, heapq, wordcloud, gensim, sklearn, string

Table 1: Tools

C. Algorithms

In this section, we will discuss about the supervised ML algorithms which we are used in this paper. This paper will use two ML algorithms, namely, Naive Bayes, AdaBoost and Random forest.

1) Naive Bayes: It is a simple classification technique which is mainly based upon the Assumption that all input attributes are independent to each other. Naive Bayes is based on Bayes' Theorem and known by different terms such as independence Bayes or simple Bayes. Moreover Naive Bayes classifier based on the assumption that all the features are independent of each other, for example a fruit can be classified as an apple, just because it is round, red color and 8 cm in diameter. The classifier considers the features like roundness, color and size independently before predicting it as apple. Naive Bayes considered as simple, easy and used in the very large data set.

The least difficult solutions are generally the most remarkable ones, and Naive Bayes is a genuine case of that. Notwithstanding the extraordinary advances of Machine Learning in the most recent years, it has demonstrated to not exclusively be basic yet in addition quick, exact, and dependable. It has been used widely for many reasons, but it works particularly well with problems with the Natural Language Processing (NLP)

Naive Bayes is a group of probabilistic algorithms that utilizes the likelihood hypothesis and Bayes Theorem to anticipate a text tag (like a bit of news or a client survey). These are probabilistic, meaning these are measuring the likelihood of every tag for a given text, and afterward yield the tag with the most noteworthy one. The manner in which they get these probabilities is by utilizing Bayes Theorem, which depicts the likelihood of a feature, in view of earlier information on conditions that may be identified with that highlight.

2) Random Forest: Random forest is ensemble supervised machine learning classifiers, which creates multiple or different decision trees and at the end it integrates them to get more stable and accurate prediction. For training data N random forest select N randomly generated data with the allowable replacement of training data. After creating several trees, it makes prediction to find the best possible solution by the majority voting.

D. AdaBoost

AdaBoost is an iterative algorithm that at every cycle extricates a powerless classifier from the arrangement of frail classifiers and weight is being assigned to the classifier as indicated by its importance. The boosting procedure has pulled in a ton of consideration among specialists in the field so as to legitimize its great performance in practice and its relative immunity to over-fitting.

VII. Architecture

The high level architecture diagram of the project is given in Fig 1.

A. Design and Implementation

In this project, the dataset is taken from online resource such as Miller and wikisource, which contains the URL links for President State of Union Addresses and Annual messages.

B. Data Preparation

There are 4 feature variables available in the dataset, if the data is imbalanced, it can be balanced by adding additional set of data. This is also the stage of data cleaning. We use library called "urllib request" to open the url link. Beautiful soup which is supported by python 3.x is used for web scraping. it is a html/xml parser that is used to read and navigate through contents of url. Once we have the content, we tokenize the paragraph and store it in a corpus. For each words in corpus we do cleaning , stemming and lemmitization/lemma which are NLP steps. We make use of library called "regular expression" for substituting and replacing the data to remove unwanted words when web scraped. We then perform stop word removal.

C. Text processing

Text processing is main step in towards the sentimental analysis. It converts the text into more readable form of data, so that the machine learning algorithm could perform better. The NLP techniques can be achieved by using NLTK python library. There are certain NLP techniques we are using in this project. They are explained below:

1) Tokenization: Given a character succession and a characterized archive unit, Tokenization might be characterized as the way toward breaking it up into pieces, called tokens, may be simultaneously discarding certain characters, for example, punctuation. These tokens are regularly approximately alluded to as terms or words, yet it is some of the time-critical to make a type-token distinction. We use the Punkt sentence tokenizer from the NLTK library. This tokenizer separates a corpus into a rundown of sentences, by utilizing a solo calculation to construct a model for shortened form words, collocations, and words that start sentences. It must be prepared on a huge assortment of plain content in the objective language before it very well may be utilized. The NLTK information bundle incorporates a pre-processed Punkt tokenizer for English.

2) Stop Words Removal: To expand the processing time and improve the presentation of the model it is important to expel the stop words like "the", "an", "a", wherein it takes additional memory space during the procedure. Removing stopwords are definitely not a firm standard or fast rule in NLP. It relies on the undertaking that we are dealing with. For errands like content grouping, where the content is to be arranged into various classes, stopwords are expelled or barred from the given content with the goal that more center can be given to those words which characterize the significance of the content. On evacuating stopwords, dataset size reductions and an opportunity to prepare the model additionally diminishes. Removing stopwords can conceivably help improve the performance as there are less and just significant tokens left. Accordingly, it could expand order exactness. Indeed, even web search tools like Google remove stopwords for quick and important recovery of information from the database.

3) Stemming: For syntactic reasons, archives are going to utilize various types of a word, for example, compose, sorts out, and arranging. Also, there are groups of derivationally related words with comparative implications, for example,

majority rule government, just, and democratization. By and large, it appears as though it would be valuable for a quest for one of these words to return reports that contain another word in the set. The objective of both stemming and lemmatization is to diminish inflectional structures and now and again derivationally related types of a word to a typical base structure. For instance:
am, are, is can be reduced to base form ”be”.
car, cars, car’s, cars’ can be reduced to base form ”car”.
The result of this mapping of text will be something like: the boy’s cars are different colors results in:
the boy car be differ color
In any case, the two words contrast in their flavor. Stemming typically alludes to a rough heuristic procedure that hacks off the parts of the bargains the expectation of accomplishing this objective accurately more often than not, and frequently incorporates the expulsion of derivational affixes.

The most common algorithm for stemming English, and one that has repeatedly been shown to be empirically very effective, is Porter’s algorithm (Porter, 1980). The whole algorithm is excessively long and complicated to introduce here, however, we will demonstrate its general nature. Doorman’s calculation comprises of 5 periods of word decreases, applied successively. Inside each stage there are different shows to choose rules, for example, choosing the standard from each standard gathering that applies to the longest suffix.

4) Lemmatization: Lemmatization, as a rule, alludes to doing things appropriately with the utilization of a jargon and morphological analysis of words, typically intending to expel inflectional endings just and to restore the base or word reference type of a word, which is known as the lemma. At whatever point went looking with the token saw, stemming may return just s, while lemmatization would endeavor to return either watch or saw subordinate upon whether the utilization of the token was as an action word or a thing. The two may in like manner differentiate in that stemming most commonly folds derivationally related words, while lemmatization normally just falls the assorted inflectional kinds of a lemma. Semantic planning for stemming or lemmatization is routinely done by an extra module part to the mentioning procedure, and diverse such sections exist, both business and open-source.

D. Review Analysis

We are collecting a list of positive and negative words in a text file and comparing our tokenize words with the list of words and count the number of positive and negative words. Each speech will have numerous amount of positive and negative words. It is essential to identify the count of positive and negative words to correctly classify a speech will have positive or negative impact. Based on overall negative words usage in all speeches we will try to set a threshold for usage of negative words to classify them as negative impact speech.

E. Word Net

WordNet is the lexical database for example word reference for the English language, explicitly intended for natural language processing. Synset is an exceptional sort of a straightforward interface that is available in NLTK to look into words in WordNet. Synset examples are the groupings of synonymous words that express a similar idea. A portion of the words have just a single Synset and some have a few.

F. Sentiment Polarity

After all the text processing one need to find the sentimental polarity of the review data. So that we can determine the amount of positive and negative reviews in the dataset. For this purpose we are using python Vader library. If needed further POS tagging can be done which improves the accuracy. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a vocabulary and rule-based supposition examination apparatus that is explicitly receptive to slants communicated in web-based social networking. VADER utilizes a blend of A supposition dictionary is a rundown of lexical highlights (e.g., words) which are commonly marked by their semantic direction as either positive or negative. VADER does not just tell about the positive and negative score yet in addition enlightens us regarding how positive or negative a supposition is.

G. Bag of Words

Bag of Words (BoW) is a calculation that checks how often a word shows up in a record. It’s a count. Those word checks permit us to think about archives and measure their similitudes for applications like inquiry, document classification and topic modeling. On the off chance that the new sentences contain new words, at that point our jargon size would increment and in this manner, the length of the vectors would increment as well. Moreover, the vectors would likewise contain a huge number, accordingly bringing about a meager lattice (which is the thing that we might want to keep away from). We are holding no data on the language structure of the sentences nor on the requesting of the words in the content.

H. TF-IDF

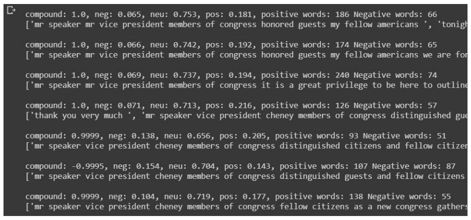
Term-Frequency Inverse Document Frequency (TF-IDF) is another approach to pass judgment on the subject of an article by the words it contains. With TF-IDF, words are given weight – TF-IDF estimates pertinence, not recurrence. That is, word checks are supplanted with TF-IDF scores over the entire dataset. To start with, TF-IDF gauges the occasions that words show up in a given record (that is ”term recurrence”). But since words, for example, ”and” or ”the” show up oftentimes in all records, those must be efficiently limited. That is the reverse report recurrence part. The more records a word shows up in, the less significant that word is as a sign to separate any given archive. That is expected to leave just the regular AND unmistakable words as markers. Each word’s TF-IDF pertinence is a standardized information position that likewise signifies one.

Bag of Words just makes a lot of vectors containing the include of word events in the record (audits), while the TF-IDF model contains data on the more significant words and the less significant ones also. BOW vectors are anything but difficult to decipher. Nonetheless, TF-IDF for the most part performs better in AI models.

I. *Classification*

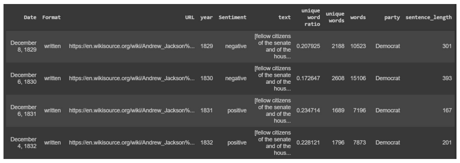
Once all the pre-processing have been done the vectorized words can be splitted into training data and testing data. This can be achieved by the python library Scikit learn. Then Split data is used for training the model using Naive Bayes and Random Forest classifiers. Once the classification have been done with the training dataset, now we can implement it in the testing data and compare the accuracy between each classifier and finally choose the best performance model and we plot ROC and Precision-Recall curve.

VIII. Results



```
compound: 1.0, neg: 0.065, neu: 0.753, pos: 0.181, positive words: 186 Negative words: 66
['mr speaker mr vice president members of congress honored guests my fellow americans ', 'tonigh

compound: 1.0, neg: 0.066, neu: 0.742, pos: 0.192, positive words: 174 Negative words: 65
['mr speaker mr vice president members of congress honored guests my fellow americans we are for

compound: 1.0, neg: 0.069, neu: 0.737, pos: 0.194, positive words: 240 Negative words: 74
['mr speaker mr vice president members of congress it is a great privilege to be here to outline

compound: 1.0, neg: 0.071, neu: 0.713, pos: 0.216, positive words: 126 Negative words: 57
['thank you very much ', 'mr speaker vice president cheney members of congress distinguished gue

compound: 0.9999, neg: 0.138, neu: 0.656, pos: 0.205, positive words: 93 Negative words: 51
['mr speaker vice president cheney members of congress distinguished citizens and fellow citizen

compound: -0.9995, neg: 0.154, neu: 0.704, pos: 0.143, positive words: 107 Negative words: 87
['mr speaker vice president cheney members of congress distinguished guests and fellow citizens

compound: 0.9999, neg: 0.104, neu: 0.719, pos: 0.177, positive words: 138 Negative words: 55
['mr speaker vice president cheney members of congress fellow citizens as a new congress gathers
```

Fig. 3: Sentiment Scores



Date	Format	URL	year	Sentiment	text	unique word ratio	unique words	words	party	sentence_length
December 8, 1829	written	https://en.wikisource.org/wiki/Andrew_Jackson%...	1829	negative	[fellow citizens of the senate and of the hous...	0.207925	2188	10523	Democrat	301
December 6, 1830	written	https://en.wikisource.org/wiki/Andrew_Jackson%...	1830	negative	[fellow citizens of the senate and of the hous...	0.172647	2608	15106	Democrat	393
December 6, 1831	written	https://en.wikisource.org/wiki/Andrew_Jackson%...	1831	positive	[fellow citizens of the senate and of the hous...	0.234714	1689	7196	Democrat	167
December 4, 1832	written	https://en.wikisource.org/wiki/Andrew_Jackson%...	1832	positive	[fellow citizens of the senate and of the hous...	0.228121	1796	7873	Democrat	201

Fig. 4: Analysis Model

As you can see in this Analysis model we have constructed useful features from unstructured raw data. This is very useful in predicting election results and for planning speeches for elections.

```
array([[0, 0, 0, ..., 0, 0, 0],
       [1, 1, 1, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [1, 1, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]])
```

Fig. 5: Bag of words model

```
array([[0.        , 0.0662974 , 0.        , ..., 0.03660169, 0.        ,
        0.        ],
       [0.        , 0.03508429, 0.        , ..., 0.07747782, 0.        ,
        0.        ],
       [0.        , 0.02349758, 0.        , ..., 0.        , 0.        ,
        0.        ],
       ...,
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ]])
```

Fig. 6: TF-IDF

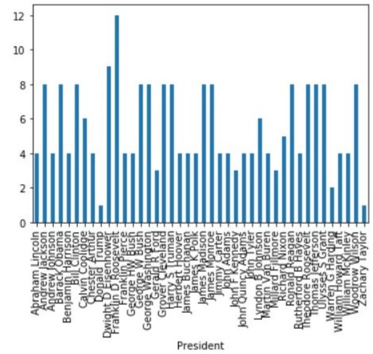


Fig. 7: Total number of speeches given by each presidents

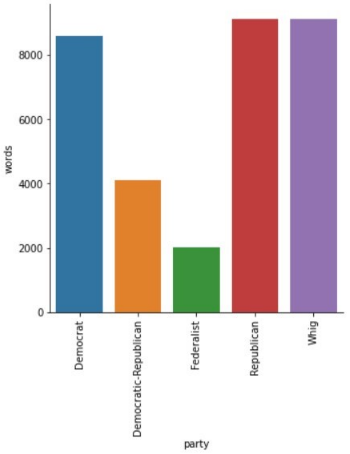
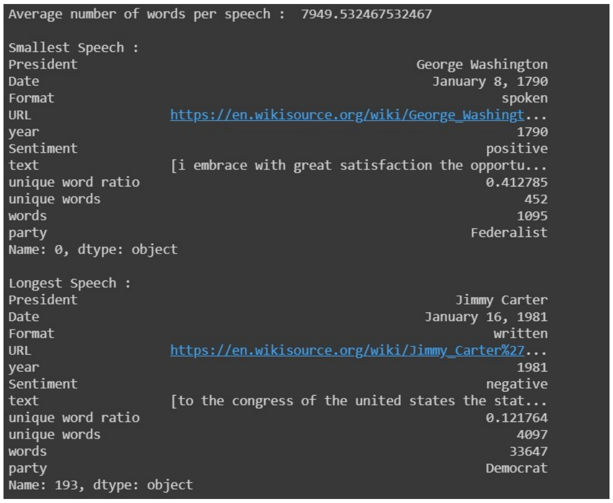


Fig. 8: Words usage by each party



```
Average number of words per speech :  7949.532467532467

Smallest Speech :
President                                      George Washington
Date                                             January 8, 1790
Format                                                    spoken
URL                  https://en.wikisource.org/wiki/George_Washingt...
year                                                        1790
Sentiment                                               positive
text                 [i embrace with great satisfaction the opportu...
unique word ratio                                       0.412785
unique words                                                 452
words                                                       1095
party                                                 Federalist
Name: 0, dtype: object

Longest Speech :
President                                           Jimmy Carter
Date                                            January 16, 1981
Format                                                   written
URL                  https://en.wikisource.org/wiki/Jimmy_Carter%27...
year                                                        1981
Sentiment                                               negative
text                 [to the congress of the united states the stat...
unique word ratio                                       0.121764
unique words                                                4097
words                                                      33647
party                                                   Democrat
Name: 193, dtype: object
```

Fig. 9: Longest and shortest speech

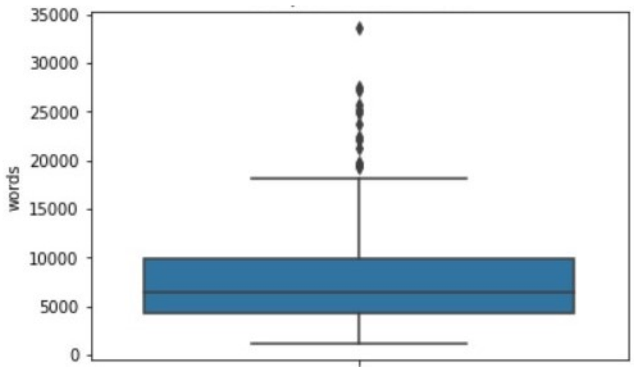


Fig. 10: Box plot of words

The usage of words in all speeches indicate that 50% of the president speeches are in the range of 4800 to 10000 words. We can also see some outliers and extreme data. This plot gives an idea to upcoming president on what is the average number of words used in speech by previous presidents.

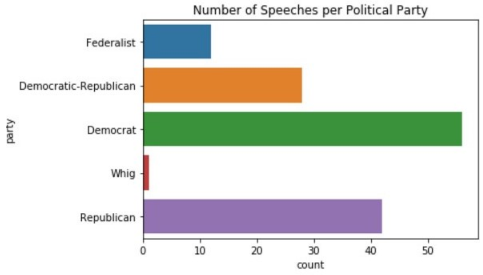


Fig. 11: Number of speeches by each party

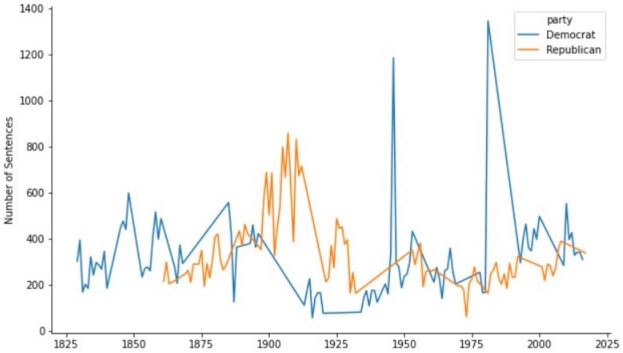
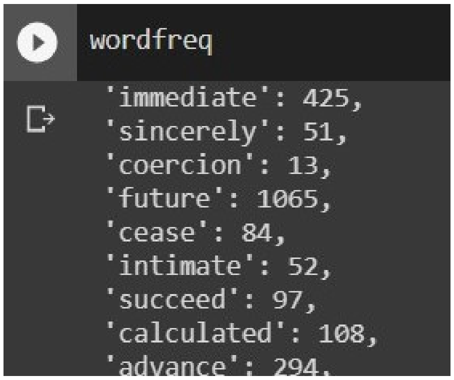
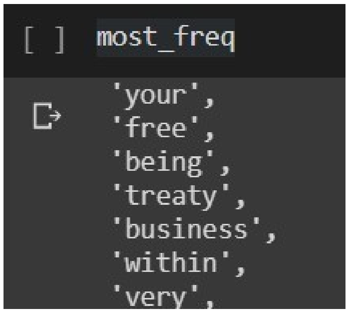


Fig. 12: Speech sentence length by Democrat and Republic party



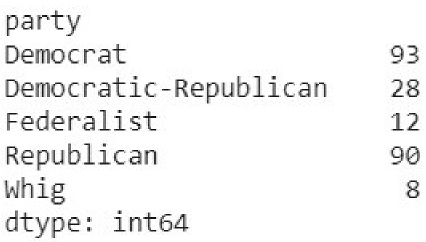
```
wordfreq

'immediate': 425,
'sincerely': 51,
'coercion': 13,
'future': 1065,
'cease': 84,
'intimate': 52,
'succeed': 97,
'calculated': 108,
'advance': 294,
```

Fig. 13: Word frequencies



```
[ ]  most_freq

'your',
'free',
'being',
'treaty',
'business',
'within',
'very',
```

Fig. 14: Most frequent words



```
party
Democrat                 93
Democratic-Republican    28
Federalist               12
Republican               90
Whig                      8
dtype: int64
```

Fig. 15: speeches given by each party

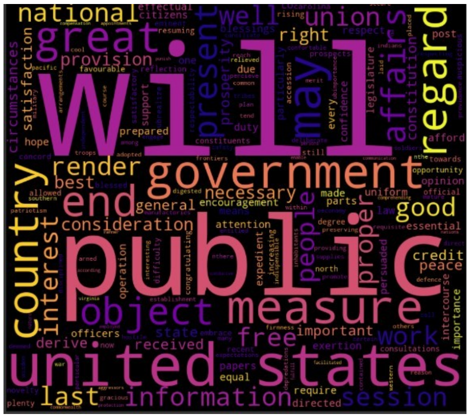
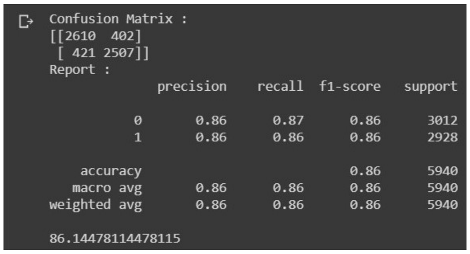


Fig. 16: Word Cloud

Word clouds are fun to use as a visual aid with blog posts to underscore the keywords on which you're focusing. The public will see the bigger, intense words and comprehend their significance to speech. Furthermore, for speakers, word mists are extraordinary to ensure you're concentrating on the correct words in their addresses.



```
Confusion Matrix :
[[2610  402]
 [ 421 2507]]
Report :
              precision    recall  f1-score   support

           0       0.86      0.87      0.86      3012
           1       0.86      0.86      0.86      2928

    accuracy                           0.86      5940
   macro avg       0.86      0.86      0.86      5940
weighted avg       0.86      0.86      0.86      5940

86.14478114478115
```

Fig. 17: Naive Bayes Model score

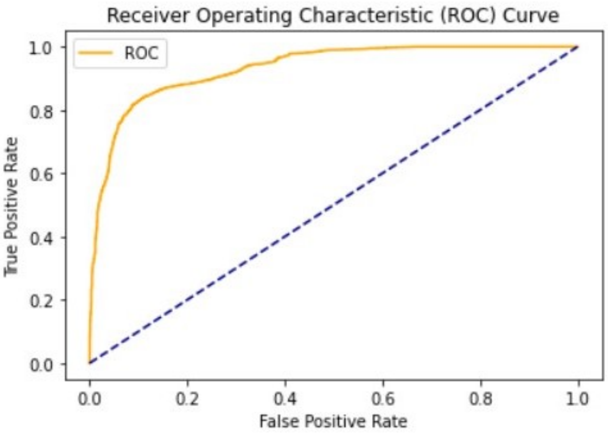


Fig. 18: ROC curve - Naive Bayes

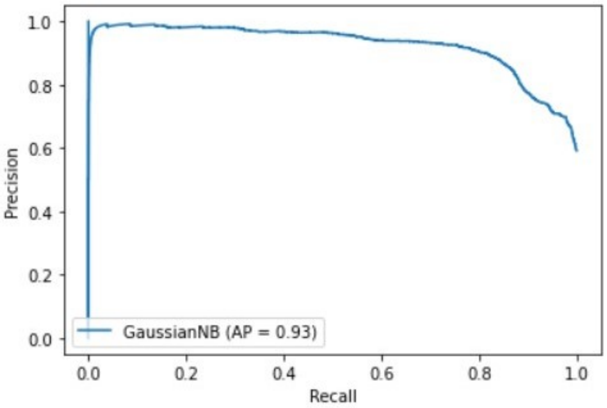
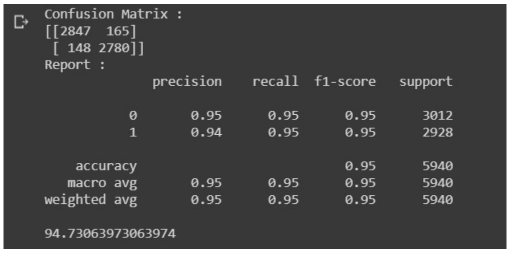


Fig. 19: Precision-Recall curve - Naive Bayes



```
Confusion Matrix :
[[2847  165]
 [ 148 2780]]
Report :
              precision    recall  f1-score   support

           0       0.95      0.95      0.95      3012
           1       0.94      0.95      0.95      2928

    accuracy                           0.95      5940
   macro avg       0.95      0.95      0.95      5940
weighted avg       0.95      0.95      0.95      5940

94.73063973063974
```

Fig. 20: Random Forest model score

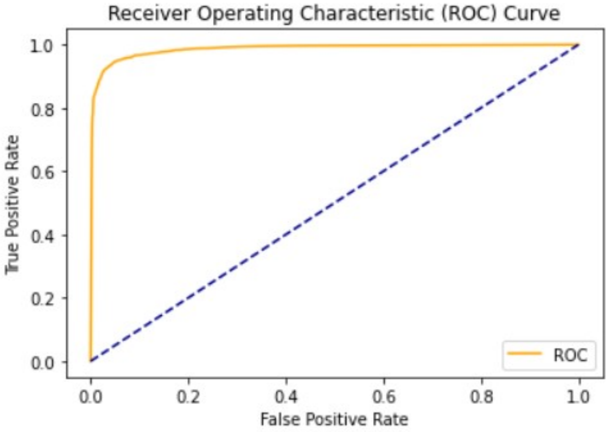


Fig. 21: ROC curve - Random Forest

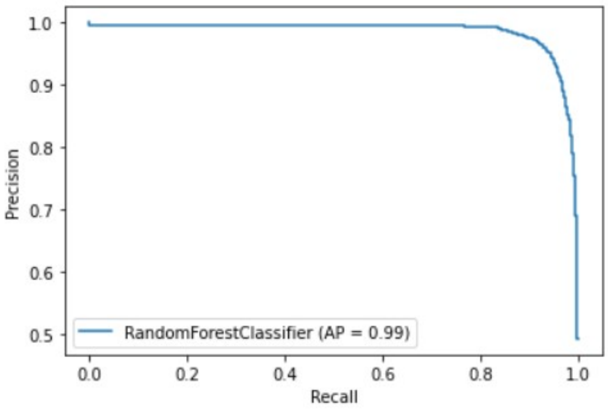
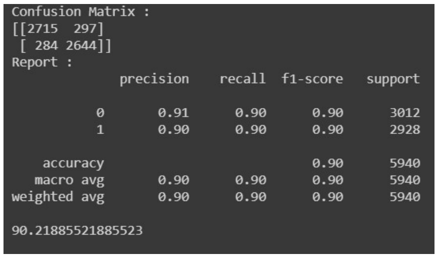


Fig. 22: Precision-Recall curve - Random Forest



```
Confusion Matrix :
[[2715  297]
 [ 284 2644]]
Report :
              precision    recall  f1-score   support

           0       0.91      0.90      0.90      3012
           1       0.90      0.90      0.90      2928

    accuracy                           0.90      5940
   macro avg       0.90      0.90      0.90      5940
weighted avg       0.90      0.90      0.90      5940

90.21885521885523
```

Fig. 23: AdaBoost Model score

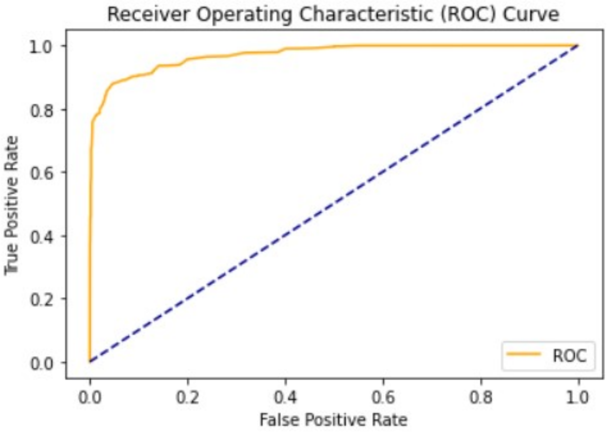


Fig. 24: ROC curve - AdaBoost

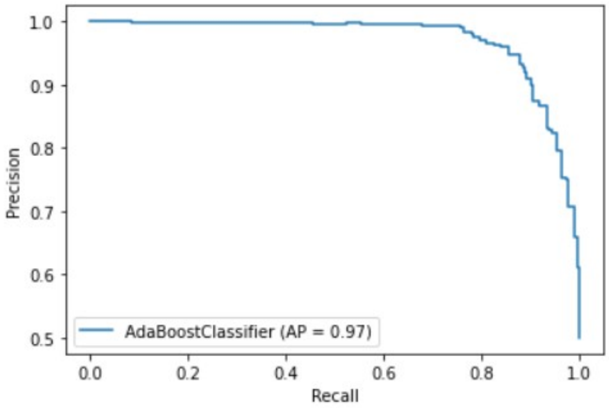


Fig. 25: Precision-Recall curve - AdaBoost

IX. Conclusions

This work presents a generalized model that takes an website link which contains a speech of particular president and studies the content by automatically extracting the speech content in website and stores the speech in corpus. Right now, we have proposed a basic framework to do the previously mentioned task. The framework functions admirably with the misleadingly created dataset, we are chipping away at gathering a bigger dataset and expanding the versatility of the framework. Naive Bayes ends up being increasingly successful and beats Random Forest and AdaBoost. Both Random Forest and AdaBoost get overfit. It is common in these algorithms to get overfit. Tree size must be adjusted to reduce overfitting. Hence it is advisable to use Naive Bayes for Natural Language Processing. Despite the fact that the framework is precise in fathoming the sentiment of the speakers in speech data, it endures a few defects, at this moment the framework can deal with a discussion between two speakers and in the discussion just a single speaker should talk at a given time, it can't comprehend if two individuals talk all the while. Our future work would address these issues and improve the exactness and versatility of the framework. When we can predict outcomes of sentiment for election speeches of elected presidents, if the same algorithm is applied for speeches of presidential debate or speeches of incumbents, we can predict the outcome of elections.

References

[1] Pang, B., and Lee, L. (2004, July). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (p. 271). Association for Computational Linguistics.

[2] Pang, B., and Lee, L. (2005, June). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd annual meeting on association for computational linguistics (pp. 115-124). Association for Computational Linguistics.

[3] Pang, B., Lee, L., and Vaithyanathan, S. (2002, July) : sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

[4] Shaikh, M., Prendinger, H., and Mitsuru, I. (2007). Assessing sentiment of text by semantic dependency and contextual valence analysis. Affective Computing and Intelligent Interaction, 191-202.

[5] Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea (2004). Sphinx-4: A flexible open source framework for speech recognition.

[6] Hutto, C. J., and Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth International AAAI Conference on Weblogs and Social Media.

[7] Yi, J., Nasukawa, T., Bunescu, R., and Niblack, W. (2003, November). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In Data Mining, 2003. ICDM 2003. Third IEEE International Conference on (pp. 427-434). IEEE.

[8] Diksha Khurana1, Aditya Koli, Kiran Khatter and Sukhdev Singh, Department of Computer Science and Engineering,(August 2017) Natural Language Processing: State of The Art, Current Trends and Challenges,publication-319164243.

[9] Alaa Alahmadi, Arash Joorabchi, Abdulhussain E. Mahdi, Combining Bag-of-Words and Bag-of-Concepts representations for Arabic text classification, published in 25th IET Irish Signals and Systems Conference 2014 and 2014 China-Ireland International Conference on Information and Communications Technologies (ISSC 2014/CIICT 2014).

[10] Prafulla Bafna, Dhanya Pramod, Anagha Vaidya, Document clustering: TF-IDF approach, published in 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT).

[11] Ahlam Alrehili ; Kholood Albalawi, ”Sentiment Analysis of Customer Reviews Using Ensemble Method”, International Conference on Computer and Information Sciences (ICCIS), 2019.

[12] Prosanta Kumar Chaki ; Ikrum Hossain ; Probhas Ranjan Chanda ; Shikha Anirban, “An Aspect of Sentiment Analysis: Sentimental Noun with Dual Sentimental Words Analysis” INSPEC Accession Number: 18075945,DOI: 10.1109/CTCEEC.2017.8455159

[13] Lin Zhao, ”A Dynamic Pricing Mechanism Model Based on Sentiments Analysis”, International Conference on Intelligent Transportation, Big Data and Smart City (ICITBS), 2019

[14] Dipti Mahajan, Dev Kumar Chaudhary,”SENTIMENT ANALYSIS USING RNN AND GOOGLE TRANSLATOR” , 8th International Conference on Cloud Computing, Data Science and Engineering (Confluence), 2018

[15] Xinyu Zhang, Guanxu Xie, Daqing Li*, Rui Kang, Reliability Evaluation Based on Sentiment Analysis of Online Comment”, 12th International Conference on Reliability, Maintainability, and Safety (ICRMS), 2018

[16] Zhengzheng Liu , Nan Yang, Sanxing Cao, ”Sentiment-Analysis of Review Text for Micro-video”, 2nd IEEE International Conference on Computer and Communications, 2016

[17] Pankaj, Prashant Pandey, Muskan, Nitasha Soni, ”Sentiment Analysis on Customer Feedback Data: Amazon Product Reviews ”, International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16th Feb 2019

[18] Wenjie Zheng1, Zenan Xu1, Yanghui Rao1, Haoran Xie2, Fu Lee Wang3, Reggie Kwan4, ”Sentiment Classification of Short Text Using Sentimental Context”, International Conference on Behavioral, Economic, Socio-cultural Computing (BESC), 2017

[19] Sunil Kumar Khatri ; Ayush Srivastava, “Using Sentimental Analysis in Prediction of Stock Market Investment” INSPEC Accession Number:16544223,DOI: 10.1109/ICRITO.2016.7785019

[20] Surya Prabha PM; Subbulakshmi.B, ”Sentimental Analysis using Naïve Bayes Classifier”, International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), March 2019

[21] Sunil Kumar Khatril, Ayush Srivatsava, “Capital Market Forecasting By Using Sentimental Analysis”, 2nd International Conference on Next Generation Computing Technologies (NGCT), 2016

[22] Hase Sudeep Kisan ; Hase Anand Kisan ; Aher Priyanka Suresh, “Collective Intelligence and Sentimental Analysis of Twitter Data By Using StandfordNLP Libraries with Software as a Service (SaaS)”, 2016

[23] Bin Wen1 ,Shanrong Duan1 ,Bin Rao1, Wenhua Dai1, “Research on Word Sentimental Classification based on Transductive Learning“, 2015

[24] DongLIU , Changqin QUAN , FujiREN , PengCHEN , “Sentiment and Sentimental Agent Identification Based on Sentimental Sentence Dictionary”, International Conference on Natural Language Processing and Knowledge Engineering, 2008

[25] Atharva Patil, Nishita S. Upadhyay, Karan Bheda, Rupali Sawant, “Restaurant's Feedback Analysis System using Sentimental Analysis and Data Mining Techniques”, International Conference on Current Trends towards Converging Technologies (ICCTCT), 2018

[26] G. Kavitha ; B. Saveen ; Nomaan Imtiaz, ”Discovering Public Opinions by Performing Sentimental Analysis on Real Time Twitter Data”, International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), 2018

[27] Ruchi Mehra1 , Mandeep Kaur Bedi2 , Gagandeep Singh3 , Raman Arora4 , Tannu Bala5 , Sunny Saxena6, ”Sentimental Analysis Using Fuzzy and Naive Bayes”, International Conference on Computing Methodologies and Communication (ICCMC), 2017

[28] Sanjay K.S Dr.Ajit Danti , ”Sentimental Analysis on Web Mining using Statistical Measures”, International Conference on Power, Control, Signals and Instrumentation Engineering, 2017

[29] M J Adarsh ; Pushpa Ravikumar, ”An Effective Method of Predicting the Polarity of Airline Tweets using sentimental Analysis”, 4th International Conference on Electrical Energy Systems (ICEES), 2018

[30] Neha Gigi, Amanpreet Kaur,”Sentimental Analysis On Social Feeds to Predict the Elections” , First International Conference on Secure Cyber Computing and Communication(ICSCCC), 2018