Predicting Twitter Sentiment of the US public about 2016 Presidential Candidates

First Name	**Last Name**
Hemanth	Inukonda Kamalakannan
SreeKumar	Selvam
Srilekha	Napa Ugandhar

Table of Contents

1.Introduction

As all of us know that the interesting and most highlighted topic currently running in the United States is the Presidential Election,2016. So, we decided to take the most highlighted topic in the field to predict the sentiment of the public after carefully examining, experimenting and analyzing them. We are excited about how the public views the election candidates. Feelings toward candidates fluctuate swiftly as interviews, debates, responses to world routine, and different disorders come to the front. To reap a colossal, diverse dataset of current public opinions on the candidates, we made up our minds to use Twitter. Twitter supplies us with residing entry to opinions about the election across the globe. We are aiming for a better, extra primary evaluation of sentiments toward candidates than the media and polls have. We are expanding on the Positive, negative and neutral emotions.

Objective:

- Predicting sentiments of the US public about the Presidential election 2016.
- Multi-label classification allows us to categorize the tweets into multiple distinct classes to handle a more specific range of emotions. We modeled the task as a multi-label classification problem to account for an array of different sentiments that people could realistically feel regarding the election candidates
- This will be performed using algorithms "SVM", "KNN" and "Maximum Entropy classifiers" to come up with the most efficient approach for analyzing.
- In addition, word clouds are formed using various emotions of the public towards the candidates.

2. Data

For our input data, we downloaded tweets using the Twitter API Keywords like "PRESIDENTIAL ELECTION 2016"

Collecting Twitter data

- Total number of Tweets -10000
- Unique Tweets – 6000
- Removed Attributes- 7

3. KDD

3.1. Data Mining Processes

Sentimental Analysis:

Sentiment Analysis is an area of Natural Language Processing which is focused on identifying and extracting subjective information from human language. Sentiment Analysis usually tries to identify the attitude of the owner of the analyzed text with respect to some topic. Sentiment analysis is a classification technique which performs either Binary Classification or Multi-class categorization.

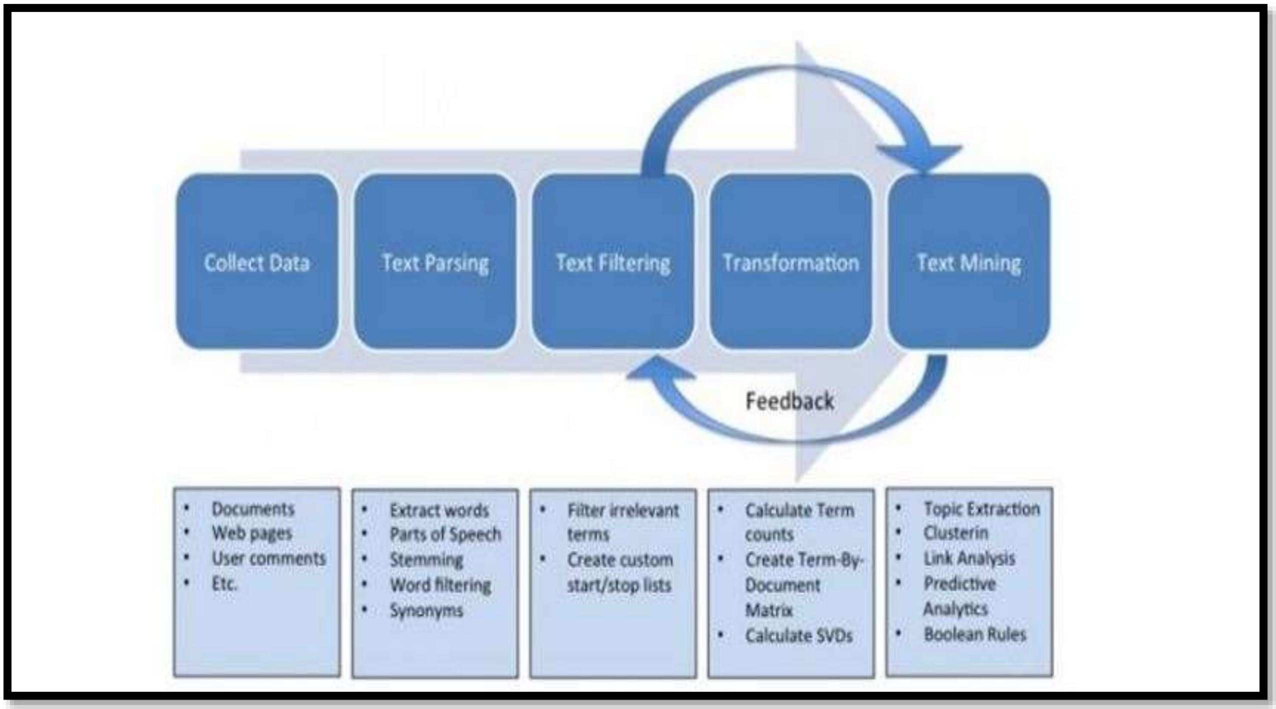
- Binary classification is also known as polarity classification which classifies text as either positive or negative.
- Multi-class categorization classifies text into positive, negative or neutral.

The sentiment is a view based on emotion instead of reason. It is a subjective impression of the sentiment feeling in art and literature. Sentiment analysis is the computational technique for extracting, classifying, understanding and determining the opinions expressed in various contents. It attempts to identify the opinion / sentiment that holds towards an object. sentiment analysis aims to determine the state of mind of a speaker or a writer with respect to some topic or the overall tonality of a document.

Sentiment analysis is widely used in online reviews, recommendations, blogs, user opinion towards political candidates, etc.

Text Mining:

Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text by usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database, deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling ,knowledge discovery process.



Text Preprocessing of data is the process of preparing and cleaning the data of dataset for classification. Pre-processing is a technique which is performed to reduce the noise in the text to improve the performance of the classifier and speed up the classification process.

BASIC TERMS IN TEXT MINING

- **Model** refers to an algorithm as applied to a dataset, complete with its settings many of the algorithms have parameters which the user can adjust.

- **Observation** is the unit of analysis on which the measurements are taken a customer, a transaction, etc.; also called the case, record, pattern or row. Each row typically represents a record, each column a variable.
- **P(A|B)** is the conditional probability of event A occurring given that event B has occurred. Read as "the probability that A will occur, given that B has occurred."
- **The pattern** is a set of measurements on an observation e.g., the height, weight, and age of a person.
- **Prediction** means the prediction of the value of a continuous output variable; also called estimation.
- **Predictor** usually denoted by X is also called a feature, input variable, independent variable, or, from a database perspective, a field.
- **Feature** denotes a representative term or set of terms selected from a corpus for the purpose of applying data mining techniques.
- **Clustering** denotes the task of grouping documents or words according to the similarity. Similarity measures usually used are Cosine Similarity, Euclidean distance etc.
- **The response** usually denoted by Y, is the variable being predicted in supervised learning; also called dependent variable, output variable, target variable or outcome variable.
- **Score** refers to a predicted value or class. "Scoring new data" means to use a model developed with training data to predict output values in new data.
- **Success class** is the class of interest in a binary outcome, e.g., "purchasers" in the outcome "purchase/no-purchase"
- **Supervised learning** refers to the process of providing an algorithm like, logistic regression, regression tree, etc. with records in which an output variable of interest is known and the algorithm "learns" how to predict this value with new records where the output is unknown.

- **Test data** refers to that portion of the data used only at the end of the model building and selection process to assess how well the final model might perform on additional data.
- **Training data** refers to that portion of data used to fit a model.
- **Unsupervised learning** refers to analysis in which one attempts to learn something about the data other than predicting an output value of interest E.g., clusters.
- **Validation data** refers to that portion of the data used to assess how well the model fits, to adjust some models, and to select the best model from among those that have been tried.
- **Variable** is any measurement on the records, including both the input (X) variables and the output (Y) variable.

4.) Approach:

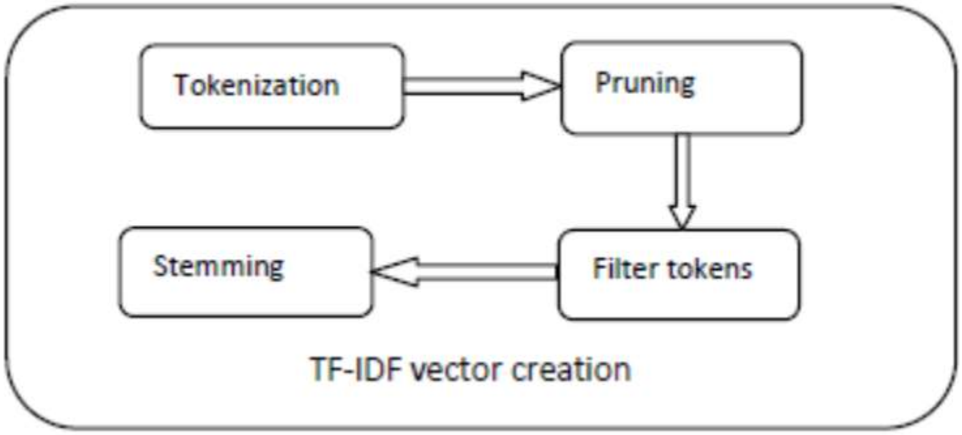
1. **A collection of data:**

- Using Twitter Search API keywords like "PRESIDENTIAL ELECTION 2016" tweets will be collected for Election Candidates.

  Total number of Tweets -10000
  Unique Tweets – 6000

2. **Text Preprocessing**

Preprocessing approach plays the main role in textual content mining strategies and functions. It's the first step in the text mining process. The tweets content – text data which are used as information is totally in an unstructured format, in order to perform analysis on the collected data, the data is primarily cleaned and structured.



- **Tokenization**

Tokenization is the approach of breaking a circulation of text into words, phrases, symbols, or different meaningful elements known as tokens. The intention of the tokenization is the exploration of the words in a sentence. The record of tokens becomes input for further processing equivalent to parsing or textual content mining Textual information is best a block of characters at the commencing. All strategies in information retrieval require the phrases of the knowledge set. Therefore, the requirement for a parser is a tokenization of files. This will likely sound trivial because the textual content is already saved in machine-readable codes. Nonetheless, some issues are still left, just like the removal of punctuation marks. Other characters like brackets, hyphens, and many others require processing as well. Furthermore, tokenizer can cater for consistency in the files. The principal use of tokenization is identifying the meaningful keywords. The inconsistency can be exclusive number and time formats

- **Pruning**

The dataset was pruned to ignore the too frequent and too infrequent words. Absolute pruning scheme was used for the task. Two parameters were used for the pruning task namely, prune below and prune above. The value of these parameters was set as: pruned below =4 and pruned above = 2000.

-Stop-words:

Many words in records recur very generally, however, are very nearly meaningless as they are used to become a member of words collectively in a sentence. It is almost always understood that discontinue words do not contribute to the context. Due to their excessive frequency of occurrence, their presence in text mining offers a drawback in figuring out the content material of the documents. Stop words are very most often used original phrases like ‘and’, ‘are’, ‘this’ etc. They are not valuable in the classification of documents. So they need to be removed. Nevertheless, the development of such stop phrases record is elaborate and inconsistent between textual sources. This process additionally reduces the text information and improves the approach performance. Each text document offers with these phrases which are not imperative for text mining purposes.

- **Filtering Tokens:**

Length established filtration scheme used to be applied for decreasing the generated token set. The parameters used to clear out the tokens are the minimal length and maximum size. The parameters outline the variety for deciding on the tokens. Within the proposed model the minimal size used to be set to four characters and maximum length to 30 characters

- **Text Transformation**

Stemming defines a procedure that is used to search out the foundation or stem of a word. The filtered token set undergoes stemming from shrinking the length of words unless a minimum length is reached. This resulted in lowering the unique grammatical forms of a phrase to a single term. Two procedure for stemming

casting off the endings:

the general principles for shedding the endings from words comprise:

i. If a word ends in „es‟ drop the s.

Remodeling the phrases:

The words can also be transformed to a couple other grammatical kind using a suite of outlined principles.

The stemming process increases the effectivity and effectiveness of the information retrieval

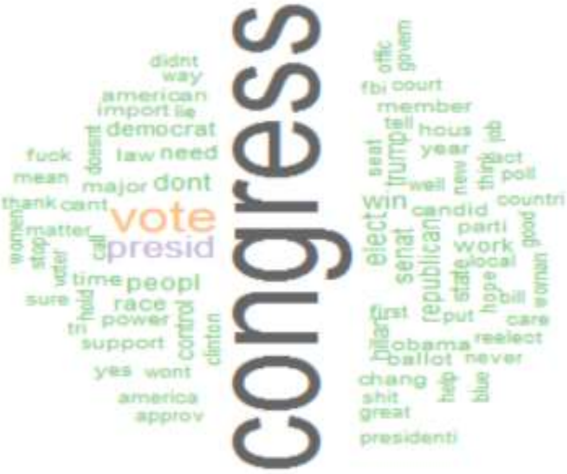
and text mining tactics.

Feature Selection:

In textual content classification, the characteristic resolution is the procedure of picking out a targeted subset of the terms of the educational set and using handiest them within the classification algorithm. The characteristic selection procedure takes situation before the educational of the classifier. Feature determination is the study of algorithms for lowering dimensionality of information to support computer finding outperformance. For a dataset with N elements and M dimensions (or elements, attributes), characteristic determination aims to scale down M to M′ and M′ ≤ M. It is a foremost and greatly used procedure to dimensionality reduction. Yet another robust approach is characteristic extraction.

Word Cloud:

A word cloud is a graphical representation of word frequency. For example, the word congress is used most number of times and all those words inside the word are the maximum frequency and now we can have the central topic from looking at the word cloud



- **Labeling the Tweets**

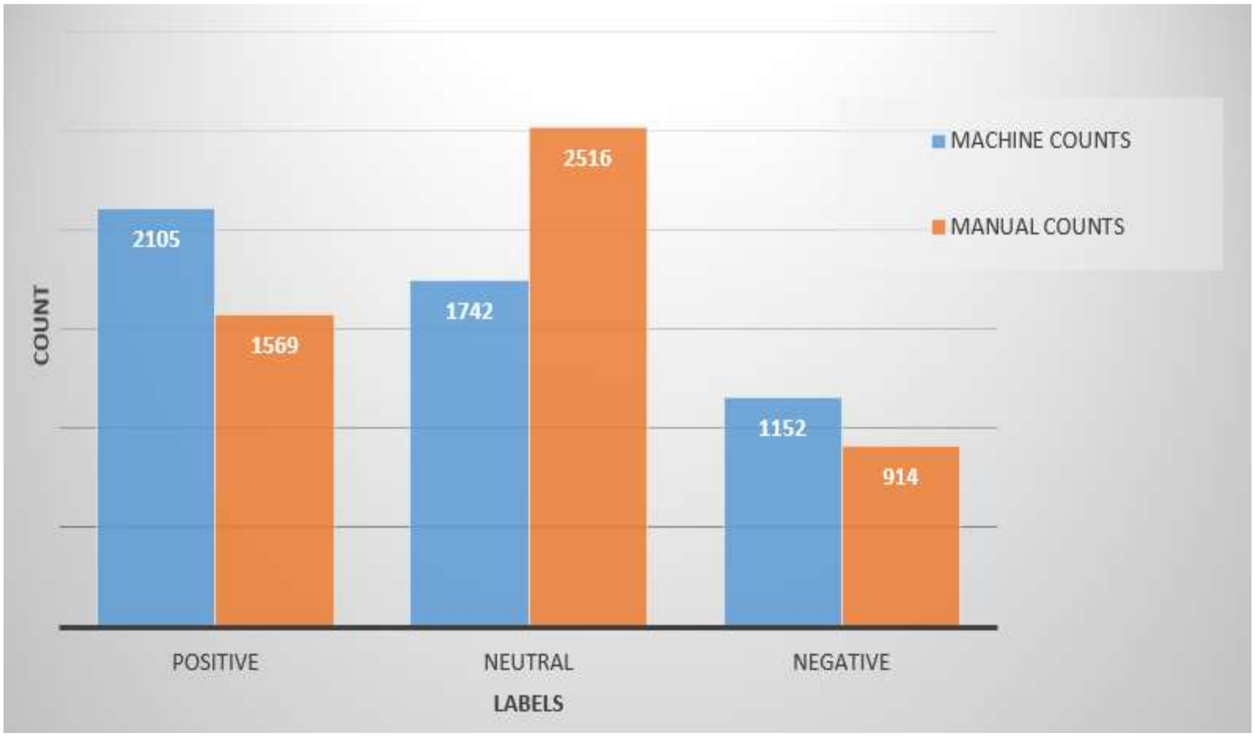
There were two types of labeling

1) manual labeling:

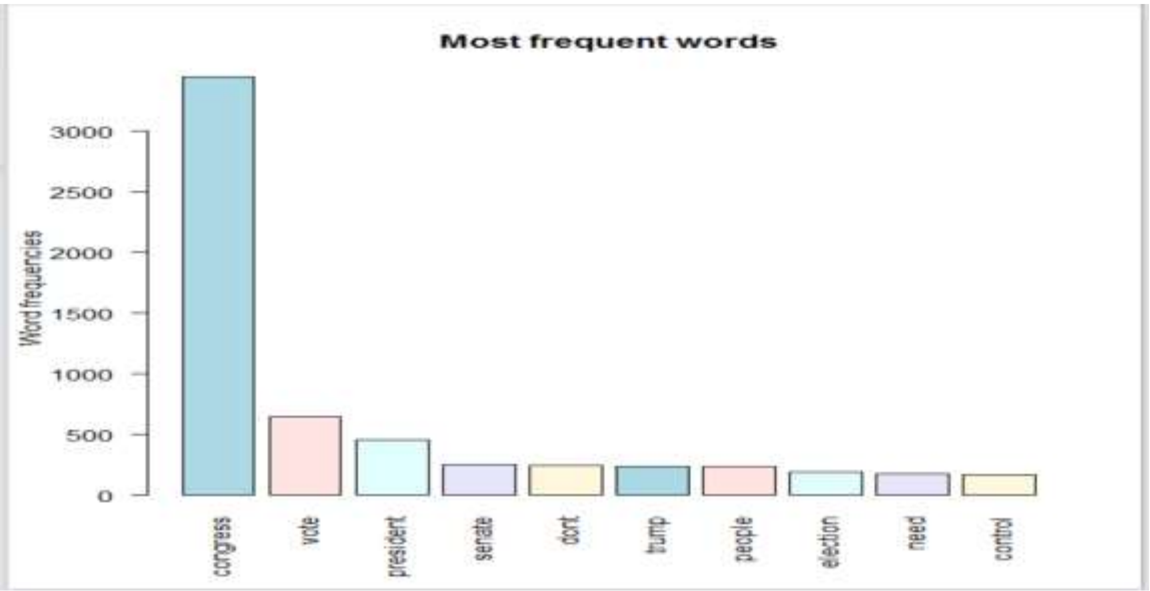
In Manual Labelling, when we came across the text we analyzed it and labeled it POSITIVE, NEGATIVE, and NEUTRAL. It works on what emotion or response a person is reacting to a particular sentence. A sentence will have grammar, emotions, reactions for which human can analyze and understand it emotionally and react to it. So on basis of that, we got sentiments and plotted a graph shown below.

2) machine labeling:

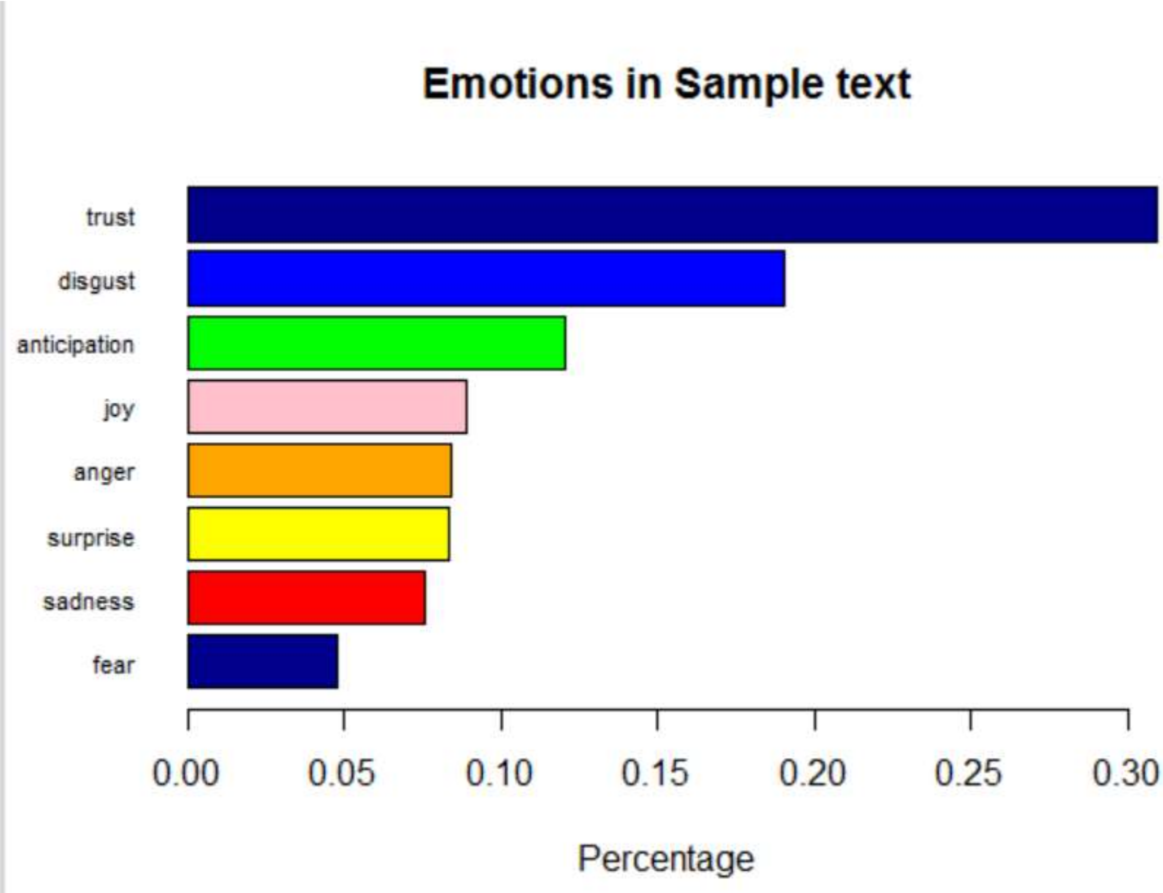
The important words which were segregated and was put in the bag of words model. this model is the one which is going to help us in predicting the sentiments of the public. The working taking place like the words which are in the model is compared with an already predefined set of positive and negative words. Now the polarity of the words is checked and we try to predict which polarity label is more, hence the label is done without the understanding of grammar, the pos tag, etc.



The above diagram depicts the difference between the labeling done to the 5000 tweets manually as well with the machine. These are some of the most important words which were used for sentimental analysis and these are some of the most occurred words in the tweets.



By Analyzing most frequently used words, the word ‘Congress’ was found to be most frequently used the word and varied emotions by the public were calculated for the word ‘Congress’ as pictured below.



Performing Classification algorithms

- Machine learning finding out tactics simulate the best way humans be taught from their past experiences to accumulate expertise and practice it in making future decisions. These learning approaches are broadly utilized in Text classification. The classification utilizing machine
- learning will also be summed up in two steps:
- 1. Studying the model utilizing the training dataset
- 2. Making use of the educated model to the scanned dataset.
- Sentiment analysis is a textual content classification trouble and therefore any present supervised classification method can also be applied. Our work uses the highest entropy, Support Vector and KNN for classifying the sentiments and compares the results got using the three approaches.

Classification Methods Used:

- **Support Vector Machines**

Support Vector Machines are supervised learning methods used for classification and regression. Support Vector Machines makes use of certain kernels in order to transform the problem. Using Kernel equation, we can apply linear classification techniques to non-linear data. The kernel equations contain hyperplane in multi-dimensional space that will arrange the data instances of one kind from those of another.

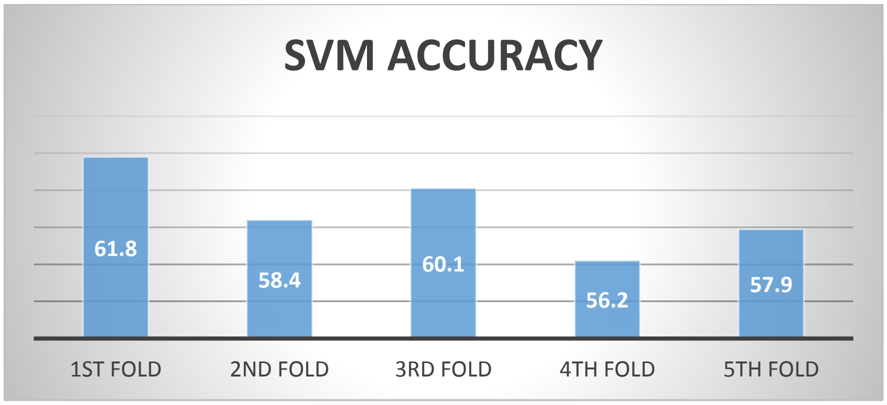
The kernel equations could basically be any function that transforms the linearly non-separable data in one domain into another domain where the instances become linearly separable.
we divide the data into two distinct categories, into separate two types of instances. The hyperplane decides the target variable value for future predictions. We select the hyperplane which maximizes the margin between the support vectors on either side of the plane. Support vectors consist of instances that are either side of the separating planes.

SVM classification is to find a maximum margin hyperplane that separates the document vector in one class from the other with maximum margin. SVM use overfitting protection, which does not necessarily depend on the number of features, they have the potential to handle these large feature spaces.

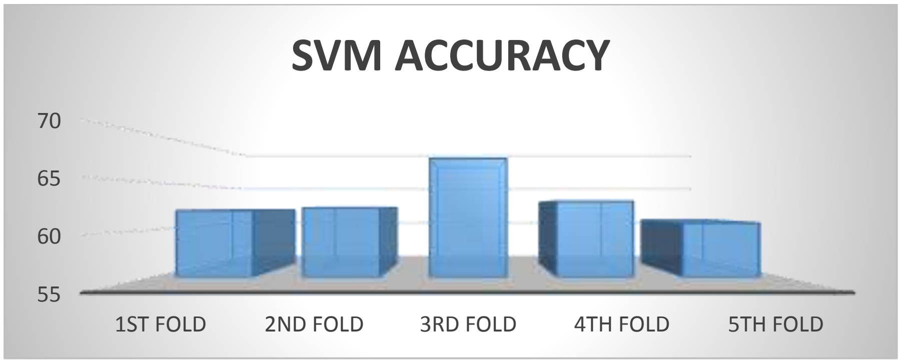
SVM are well suited for problems with dense concepts and sparse instances, Support vector Machines supports and works on binary data only, even if the text is not binary data the algorithm will consider the data as binary and does complete binary assessments on the data.

We performed 5 fold validations for SVM algorithm in two methods and accuracies are calculated

METHOD 1: Manual Counts



METHOD 2: Machine Counts



- **KNN Classifier:**

The k-Nearest Neighbors algorithm *KNN* is a non-parametric method used for classification and regression.

The input in *k* consists of the K closest training examples in the feature space. The output depends on whether *K*NN is used for classification or regression:

*K*NN is generally known to be as a non - parametric lazy learning algorithm. As it does not use the training data points to do any generalization. This means the training phase is very fast as it keeps all the training data. More exactly, all the training data is needed during the testing phase. *K*NN – makes decision-based on the entire training data set.

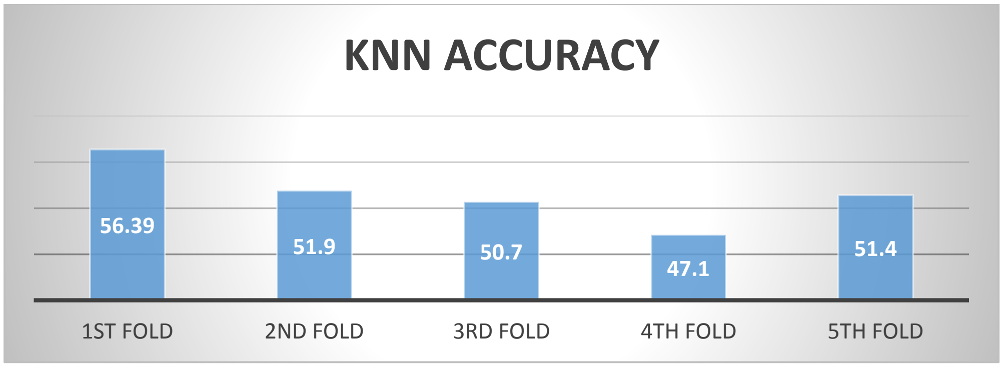
The *KNN* classifier is based on the assumption that the classification of an instance is most similar to the classification of other instances that are nearby in the vector space. Compared to other text categorization methods such as Bayesian classifier, *KNN* does not rely on prior probabilities, and it is computationally efficient.

The main computation is the sorting of training documents in order to find the *k* nearest neighbors for the test document.

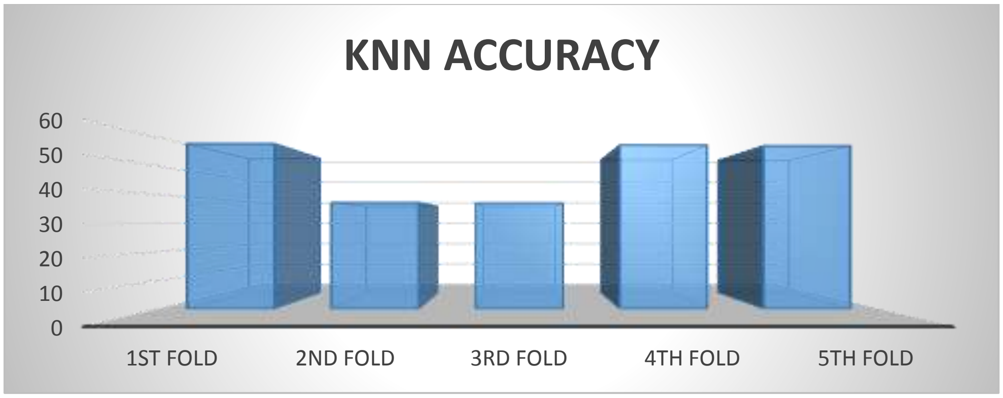
The performance of this algorithm greatly depends on two factors, that is, a suitable similarity function and an appropriate value for the parameter k.

We performed 5 fold validations for KNN algorithm in two methods and accuracies are calculated

METHOD 1: Manual Counts



METHOD 2: Machine counts



- **Maximum Entropy:**

The Maximum Entropy classifier is a probabilistic classifier which belongs to the class of exponential models.

The Maximum Entropy classifier is a discriminative classifier commonly used in Natural Language Processing, Speech and Information Retrieval problems.

The Maximum Entropy classifier can be used to solve a large variety of text classification problems such as language detection, topic classification, sentiment analysis and more.

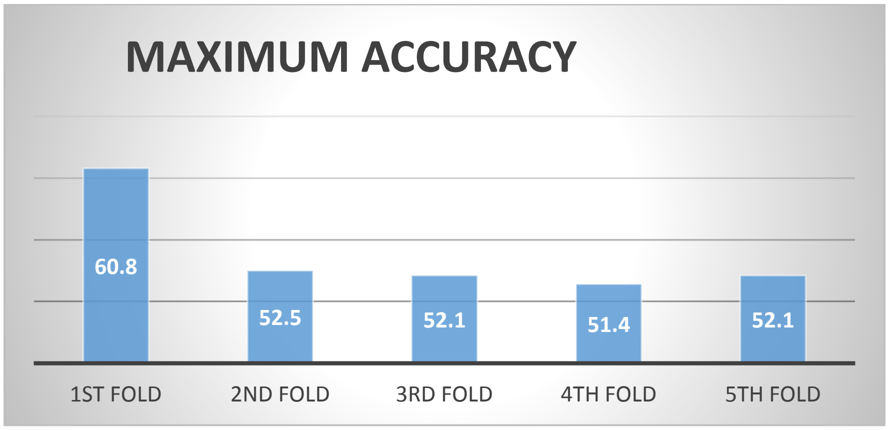
Maximum Entropy classifier is used when we can't assume the conditional independence of the features.

Maximum entropy is a general technique for estimating probability distributions from data.

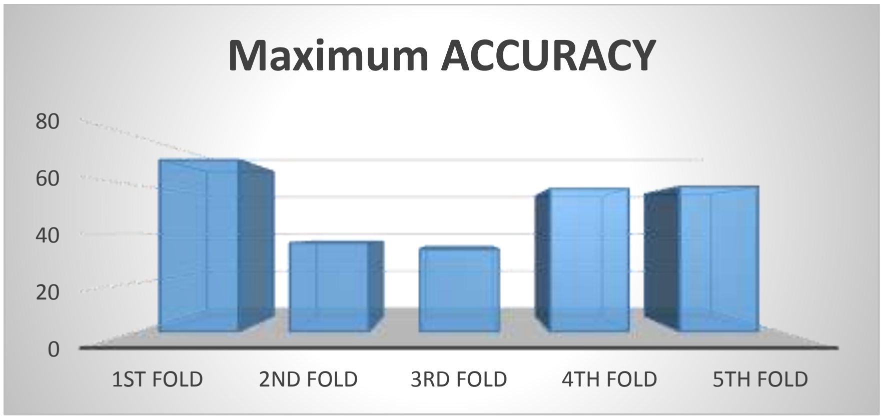
The over-riding principle of maximum entropy is that when nothing is known, the distribution should be as uniform as possible, that is, have maximal entropy.

We performed 5 fold validations for Maximum Entropy algorithm in two methods and accuracies are calculated

METHOD 1: Manual Counts



METHOD 2: Machine counts



5. Evaluations and Results

5.1. Evaluation Methods

Text categorization systems employed in the project are susceptible to make mistakes. So in order to know which classifier is better, we compare our model build in different text classifier algorithms and check for evaluation patterns to understand which does sentimental analysis better and future predictions better.

Our model the measures the performance on one multi-label category, it aggregates per-category measures, to give an overall performance.

Here values are done labels-wise. In this method for each label the precision, recall is computed and then these label-wise metrics are aggregated.

We used K-fold cross validations, Taking k =5.

The important intent is that the k-fold cross-validation estimator has a cutback variance than a single maintain-out set estimator, which will also be very primary if the amount of knowledge available is restricted. When you have a single maintain outset, the place 80% of knowledge are used for training and 20% used for testing, the experiment set may be very small, so there shall be various variant within the performance estimate for special samples of knowledge, or for one of a kind partitions of the data to kind training and test sets. K-fold validation reduces this variance through averaging over okay one-of-a-kind partitions, so the performance estimate is much less sensitive to the partitioning of the data. That you could go even additional through repeated k-fold pass-validation, where the cross-validation is carried out making use of exclusive partitioning’s of the information to form k subsets, and then taking the usual over that as good.
Ensuring good efficiency.

Precision:

Precision measures the exactness of a classifier. A higher precision means less false positives, while a lower precision means false positives. As precision increases, recall decreases.

Precision: $tp/tp+fp$

Recall:

Recall measures the completeness, or sensitivity, of a classifier. Higher recall means less false negatives, while lower recall means a false negative. As Recall increases, precision decreases.

Recall: $tp/tp+fn$

F1- Score:

F1 – score also known as F-measure or balanced F-score is a measure that combines precision and recall. It is also known as the harmonic mean of precision and recall. Commonly used in information retrieval, it measures accuracy using the statistics precision p and recalls r.

$F1=2prp+r$

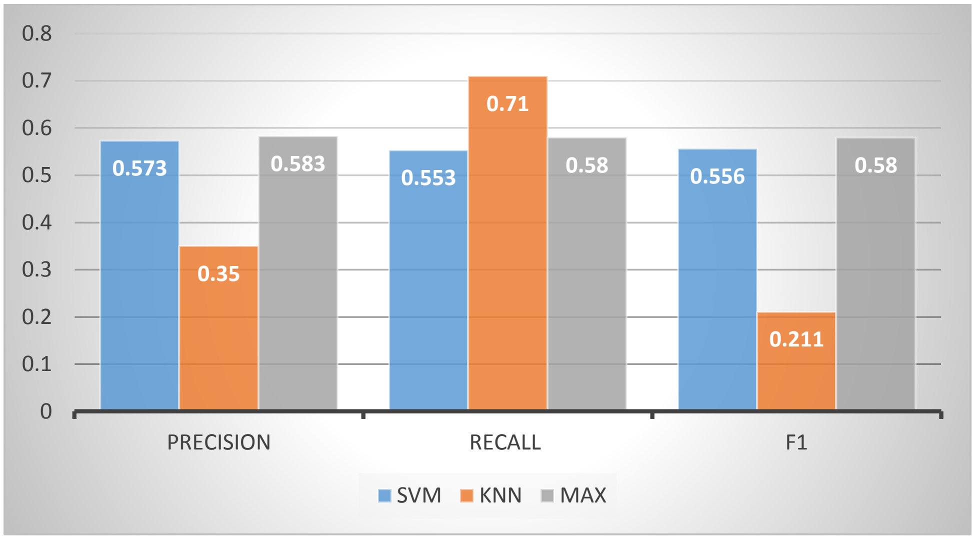
where p=tptp+fp, r=tptp+fn

5.2. Results and Findings

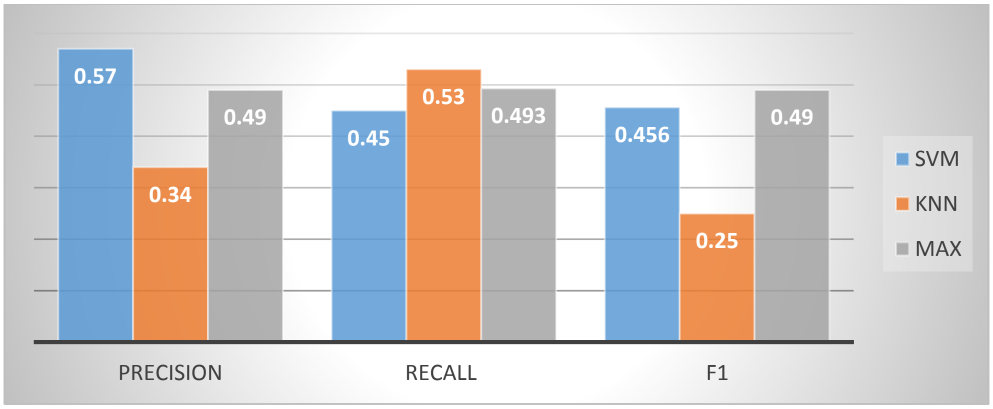
We have calculated Precision, Recall, and Accuracy for Manual Counts and Machine counts using KNN, SVM, and Maximum Entropy algorithms method.

Recall, Precision, and F1 Chart:

Method 1: Manual Counts

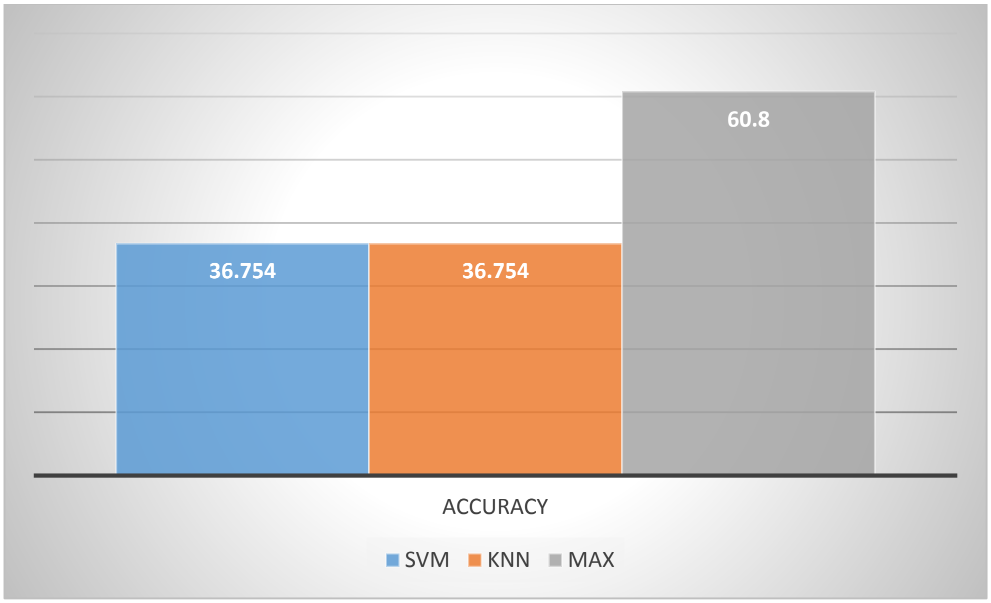


Method 2: Machine Counts

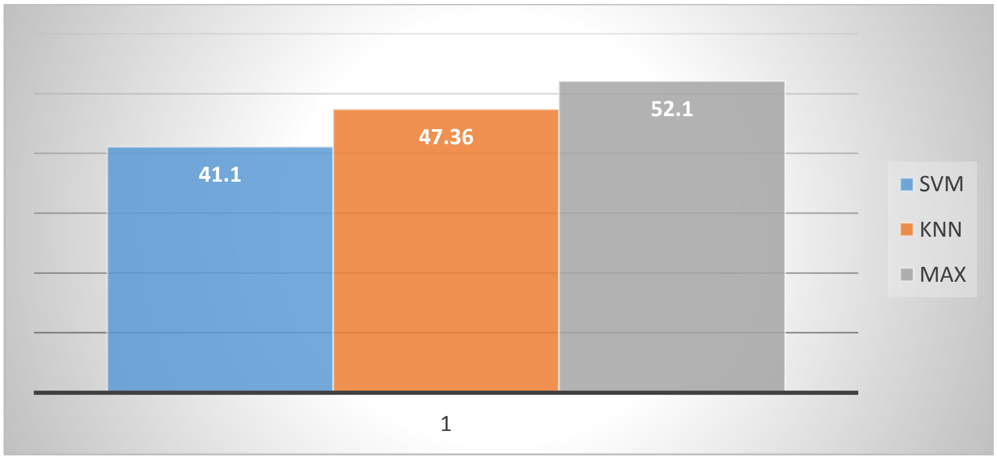


Accuracy Chart:

Method 1: Manual Counts



Method 2: Machine Counts



We find from the above-calculated charts that Maximum Entropy method has Better Accuracy and Precision values for both Manual and Machine counts method.

6. Conclusions and Future Work

6.1. Conclusions

- Sentiment analysis by comparing the different classification methods in combination with various feature selection schemes.
- The accuracy of the Maximum Entropy model stands greater than the other models compared in our project, So We conclude that Maximum Entropy model analyzes Sentiments better.

6.2. Limitations

- Ambiguity problem arises due to complex multi-lingual data collected from tweets.
- Analyzing the type of language tone like Sarcasm, ironic statements etc. are very difficult for a machine to comprehend and predict exact sentiment.
- The computational taken for analyzing a huge dataset in R language.

6.3. Potential Improvements or Future Work

This model can be enhanced for a better understanding of Sarcastic methods.

NLP techniques could be used for the better working of the model.