IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

A Case Study on Text Classification using Classification Algorithms and Latent Semantic Analysis

Shekhar Tanwar[1], **Shalini L.**[2]

Alumni, School of Computing Science and Engineering, VIT, Vellore, India[1]

Assistant Professor (Senior), School of Computing Science and Engineering, VIT, Vellore, India[2]

Abstract: Data Mining techniques are helpful in finding out patterns between data attributes and results in probalistic prediction of the label attributes. Keeping Predictive Modeling as center of attention, this paper focuses on application of analytics on dataset comprising of real world text messages. The Classification techniques i.e. Decision Tree and Random Forest combined with Bag Of Words Model, Latent Semantic Analysis, Singular Value Decomposition and Feature Engineering helps in meticulously predicting and classifying the dataset into two distinct parts i.e.legitimate text messages HAM and SPAM. The paper presents a thorough study and analysis of the techniques applied for classification and prediction, and also discusses the application of Vector Space Modelin making the dataset feasible for the application of the prediction and classification algorithms.

Keywords: Bag of Words Model, Document Frequency Matrix, Stop Words, Stemming, Cross Validation, Decision Tree, Random Forest, TF-IDF, Documents, Terms, Corpus, Vector Space Model, Latent Semantic Analysis,Singular Value Decomposition, n-gram, Feature Engineering.

I. INTRODUCTION

The main objective of paper is to study the impact of different classification algorithms in the prediction of unknown label attributes. Thiscase study is judged by using Accuracy, Sensivity, Specification and Confusion Matrix readings as parameters.These parameters effectively helps in determining the prediction power of the model developed in both the Training and the Test dataset.This paper is brochured as follows. Section II presents the methodology used in the Case Study and discusses the aspects of classification algorithm and respective datasets. Section IIIdevelops on the Experiments and summarizes the results produced by the model when different techniques are applied on the Test and Train data sets sequentially, and finalizes the results produced by the model. Section IV presents the conclusion.

II. METHODOLOGY

The following steps are included in the classification process in this paper,

- A dataset is chosen which comprises of text conversations between individuals
- The dataset is divided in the ratio of 7:3 between Train and Test datasets
- With the help of the Quanteda Package, The Bag Of Words Model is applied to tokenize the train dataset, and after stemming and removing all stop words the results are stored in Document Frequency Matrix. The Efficiency of the Model developed is checked using Decision Trees.
- Term Frequency (TF), Inverse Documents Frequency (IDF) and TF-IDF matrices calculated and Decision Tree is again applied on the modified dataset to measure accuracy
- To boost the predictive power of the model developed, n-gram is applied, and the modified Dataset is used for Latent Semantic Analysis and Singular Value Decomposition.
- Finally Vector Space model is applied to compute similarity between documents and the model is trained using Random Forest.

A. Dataset Used

The Data set used is offered from Kaggle and is originally from University Of California Irvine and comprises of real world text messages with SPAM messages present among them. The dataset has 5572 instances and 3 variables, initially. However, as the model developed progresses the variables expand to 5704, then to 29095 and finally after application of Singular Value Decomposition are reduced to 300. There were no missing values present in the dataset.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

B. Classifiers Used

1) Decision Tree:

Decision Trees are supervised learning models, which serve the dual purpose of either classifying an item into its target value (classification) or to predict the continuous value of the item under consideration(regression). Using Tree-structure as base, this classifier extracts knowledge from the large data at disposal, and then effectively classifies new data.In data mining the target variable (Y) is dependent on the input variables (X_1, X_2, $X_3 \ldots X_n$).

2) Random Forest

Random forest is a collection of decision trees. It is presented independently with some controlled modification. Trees and the results included in Random Forest are based on majority of accurate output. Random forest is the best classifier for large datasets. 1) If 'n' is the number of cases in the training set, then 'n' cases are to be sampled randomly but with replacement, from the original data. This sample will act as a training set for growing the tree. 2) If input variables are 'M' in number, a number mM is specified such that at each node, m variables randomly selected out of the 'M' input variables and among all these 'm',the best split is used to split the node. The value of m is kept constant during the forest growing. 3) Each tree is made to grow to the largest extent possible. Pruning is restricted just to get more accuracy compromising increased execution time in [7]

C. Other Techniques applied for Making Data Sets feasible for application of Classifiers

1) Bag Of Words Model

The Bag Of Words model is used to store all the Terms (words) in the documents (each row) in the data set in a Document Frequency Matrix (DFM).

The Bag of Words Model iterates through all the documents in the dataset and converts the documents in lower case, while simultaneously removing all special symbols and daily English Language words or stop words. The algorithm then picks up each term present in a document and calculates its frequency of occurrence in that document. Finally the terms are made new input variables and their corresponding frequency of occurrence is stored in the cells of the Document Frequency Matrix. The order of words is lost in this model, for which n-gram model comes handy.

2) TF-IDF

The TF-IDF addresses two key problems present in the Bag Of Words Model. Firstly, Longer documents will tend to have higher word counts. Secondly, the Bag Of Words Model doesn't account for the fact that the terms that appear frequently across the corpus aren't as important, and contribute minimally towards prediction and classification. TF-IDF counters these problems by normalizing the documents based on their length and penalizes terms frequently across the corpus, this enhances the documents term frequency matrix.The TF-IDF is calculated as the product of the Term Frequency and the Inverse Document Frequency of a matrix,

$$\textbf{TF-IDF = TF * IDF}$$

The Term Frequency and the Inverse Document Frequency are described as follows:

A. Term Frequency

Let freq (t,d) be the count of the instances of the term t in document d, also let TF(t,d) be the proportion of the count of the term t in document d. For n number of distinct terms in document d:

$$\textbf{TF (t, d)} = \frac{\textbf{freq(t,d)}}{\sum \textbf{(i to n)}\ \textbf{freq(t (i) ,d)}}$$

B.Inverse Documents Frequency

Let N be the count of the distinct terms in the corpus, and if count(t) be the number of documents in the corpus which the term t is present, the IDF is the log of how many times a term appears in the documents across the corpus, and is calculated as :

$$\textbf{IDF (t) = log } \left(\frac{\textbf{N}}{\textbf{count(t)}}\right)$$

The IDF penalizes terms, which show up frequently across the corpus. In particular, if a term shows up in every document across the corpus, the above mathematical equations gives it a weight of [0].

The TF-IDF however demands transformation of the new data set, using the IDF matrix, for the model to work on it.

3) n-gram

The n-grammodel is a continuous sequence of N items from a given sequence of text or speech. The items can be either of syllables, letters, words or base pairs, depending purely on the application of the model. The n-grams are typically collected from the dataset or the corpus in study. The n-gram model allows the bag of words model to include word ordering and thereby increases the prediction power of the classification model developed.

A n-gram model with n equal to 2 is called a bi-gram, and is a sequence of two adjacent elements from a string of

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

tokens. Bigrams use the concept of conditional probability, and thus provide the probability of a token given the preceding token, when the relation of the conditional probability is applied :

$$\mathbf{P\ (W_n \mid W_{n-1}) = \frac{P(Wn-1,Wn)}{P(Wn-1)}}$$

That is, the probability of a token given the preceding token is equal to the probability of their bigram, or the co-occurrence of the two tokens, divided by the probability of the preceding token. Using the bigram model increases the prediction power, however tremendously increases the size of the matrix, which further result in scalability issues.

4) Vector Space Model

The dimensionality and scalability problem introduced by the bi-gram model is handled by the Vector Space Model. This model represents documents as vectors of numbers. Represented this way, it enables to handle the documents geometrically, which in turn provides an intuition regarding alikeness of documents with each other. Given **document-term** frequency matrix, the dot product of the documents is indicative of document correlation given the set of matrix terms

A.Latent Semantic Analysis

Building on the Vector Space Model, this case study uses Latent Semantic Analysis (LSA). The technique focuses on extracting relationship between documents and terms assuming that terms which are close in meaning will appear in similar (i.e. correlated) pieces of text. LSA leverages a Singular Value Decomposition (SVD) factorization of **term-document** matrix to extract these relationships.

$$\mathbf{SVD\ of\ X = X = U\sum V^T}$$

Where:

U contains the eigenvectors of the term correlations, XX^T

V contains the eigenvectors of the document correlations, X^TX

$\sum$ contains the singular values of the factorization

The Latent Semantic Analysis often provides a solution for the dimensionality problem in text analytics:

- The Matrix factorization combines columns thereby potentially enriching signal in the data.
- By selecting a fraction of the most important singular values, the LSA can dramatically reduce the dimensionality.

However, performing SVD is computationally intensive, also the new data needs to be projected into the semantic space for the predictive model to effectively work on it.

B.Projecting New Data

As with TF-IDF the use of SVD will require that new data be transformed/projected before predictions can be made. The following presents a high level process for projection:

- Normalize the document vector (i.e. row)
- Complete the TF_IDF projection using the TF-IDF algorithm
- Apply the SVD projection on the document vector

Mathematically the SVD projection for documents d is :

$$\mathbf{\hat{h} = \sum^{-1} U^T d}$$

D. Factors Considered For Calculating Performance Of Classifiers

1) Accuracy

Accuracy is calculated as the percentage of the total number of instances present for classification which were correctly classified by the model developed. In this case study accuracy would be measure on a scale of 0 to 100, and not from 0 to 1.

2) Confusion Matrix

The confusion matrix depicts the number of instances of either type correctly and incorrectly classified by the model developed.

		Reference	
		ham	spam
Prediction	ham	a (TP)	b (FP)
	spam	c (FN)	d (TN)

Referring the confusion matrix table , the accuracy of the model developed would be:

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

$$\textbf{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} = \frac{a+d}{a+b+c+d}$$

3) Sensitivity
Sensitivity would be calculated as the percentage of the **ham** messages (legitimate) correctly classified
From the above table:

$$\textbf{Sensitivity} = \frac{TP}{TP+FN} = \frac{a}{a+c}$$

4) Specificity
Specificity would be calculated as the percentage of the **spam** messages (illegitimate) correctly classified
From the above table:

$$\textbf{Specificity} = \frac{TN}{FP+TN} = \frac{d}{b+d}$$

5) Similarity in Vector Space
Using cosine similarity helps us to find out which documents in the matrix are closer to each other. Cosine has advantages over other trigonometric functions, and is thus used. Documents which are closer to each other in the vector space model would have higher cosine values or higher similarity values, whereas documents which are farther from each other are would have lower similarity or lower cosine values

III.EXPERIMENTS AND RESULTS

The data set is broken into Train and Test in the ration 7:3, while keeping the percentage of HAM and SPAM intact in either distribution. The Bag Of Words Model is applied on the Train dataset and the resultant tokens are stored in the Data Frequency Matrix, which is then converted to a new Train Dataset. To the first Bag Of Model developed, 10-fold Cross Validation is applied 3 times on the Train Data set followed by Decision Tree Classification Algorithm.To increase the efficiency of the model developed, TF-IDF model is applied and the results of classification are measured by again using the 10-fold Cross Validation applied 3 times, followed by Decision Tree classification technique. Next, the Bag Of Words Model is enhanced by application of the n-gram model, which preserves word ordering, along with the compromise on computation time and issues related to scalability. The efficiency of the model is checked again using Division Trees preceded by 10-fold Cross Validation, for the last time. A reduction in efficiency calls for an improvement, and thus the Vector Space Model comprising of the Latent Semantic Analysis and finally Singular Value Decomposition is applied on the dataset. The accuracy of the model is now checked using Random Forest. Finally the model is enhanced by using the cosine similarity on dataset.

TABLE I: MODELS ANDACCURACIESONTRAINDATA

Model Used	Accuracy (%)
Bag of Words & Decision Tree	95.33
TF-IDF & Decision Tree	94.76
n-grams& Decision Tree	94.57
LSA - SVD& Decision Tree	93.36

Table I: Presents a collective sum of the accuracies of the model developed after each operation was applied on the train dataset and Decision Tree classification algorithm was used to train the mode. A comparison between accuracies of the model using Decision Trees and Random Forest suggest that Random Forest are clearly the better choice out of the two for classification in this case study.

TABLE II SUMMARY OF RESULTS FOR MODELS USED AND ACCURACY ON TRAIN AND TEST DATA

Model used	Accuracy	Sensivity	Specificity	True Positive	True Negative	Total Instances
LSA- SVD & Random Forest	96.82	96.65	98.31	3371	406	3901
LSA – SVD & Random Forest	97.1	96.84	99.28	3375	413	3901
Cosine Similarity & Random Forest	97.9	97.99	97.22	3365	454	3901
LSA- SVD &Random Forest (with spam similarity feature)	85.69	100	0	1447	0	1671
LSA- SVD & Random Forest (without spam similarity feature)	96.47	100	73.66	1447	165	1671

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

Table II presents a summarized view of the operations carried out on the Train dataset along with the parameters used to measure the efficiency of the model developed. After using LSA- SVD on the Test data set along with Spam Similarity as one of the features, the efficiency of the model significantly reduces , which suggests overfitting. In the next iteration, the Spam Similarity feature is reduced from the model, and then the efficiency of the model is checked again using the Test dataset. As is evident above, there is a considerable jump from the previous check. Also the Sensivity and Specificity values which became 100 and 0 in the first check for Test dataset finally adjusts to 100 and 73.66 suggesting that the model performed reasonably well on unseen data, here in this case that is the Test dataset.

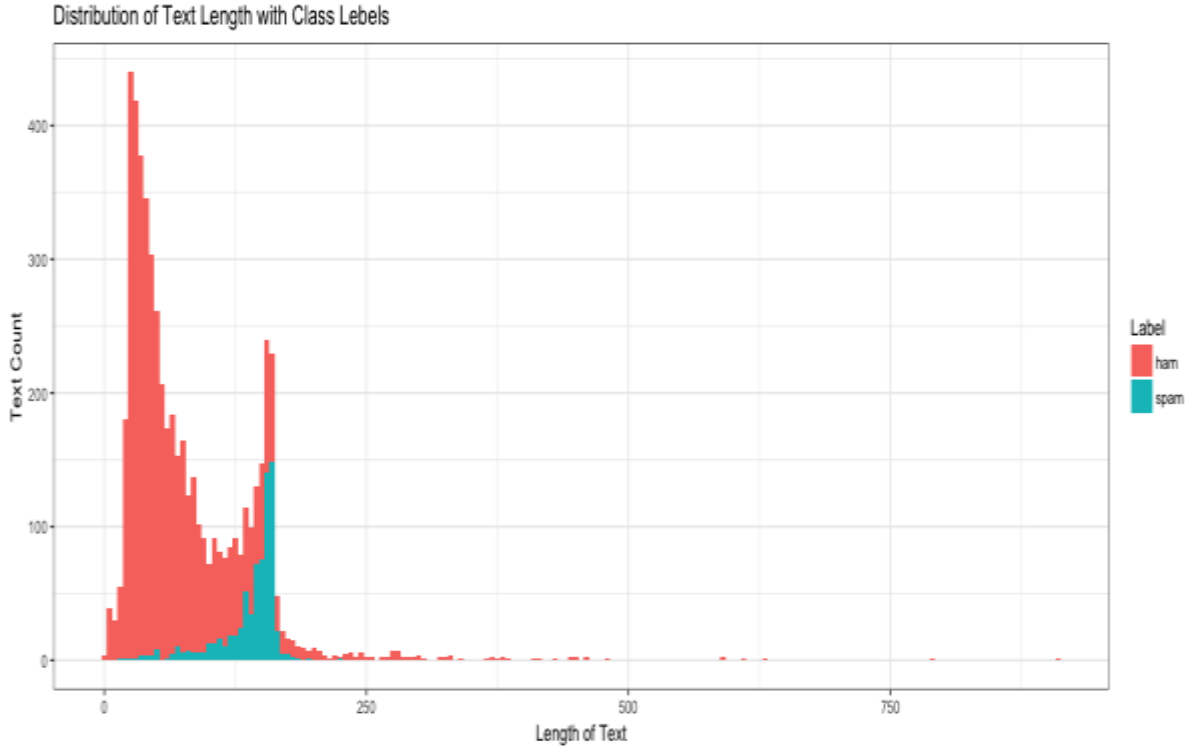

Fig. 1Comparison of Text length and Classification of labels

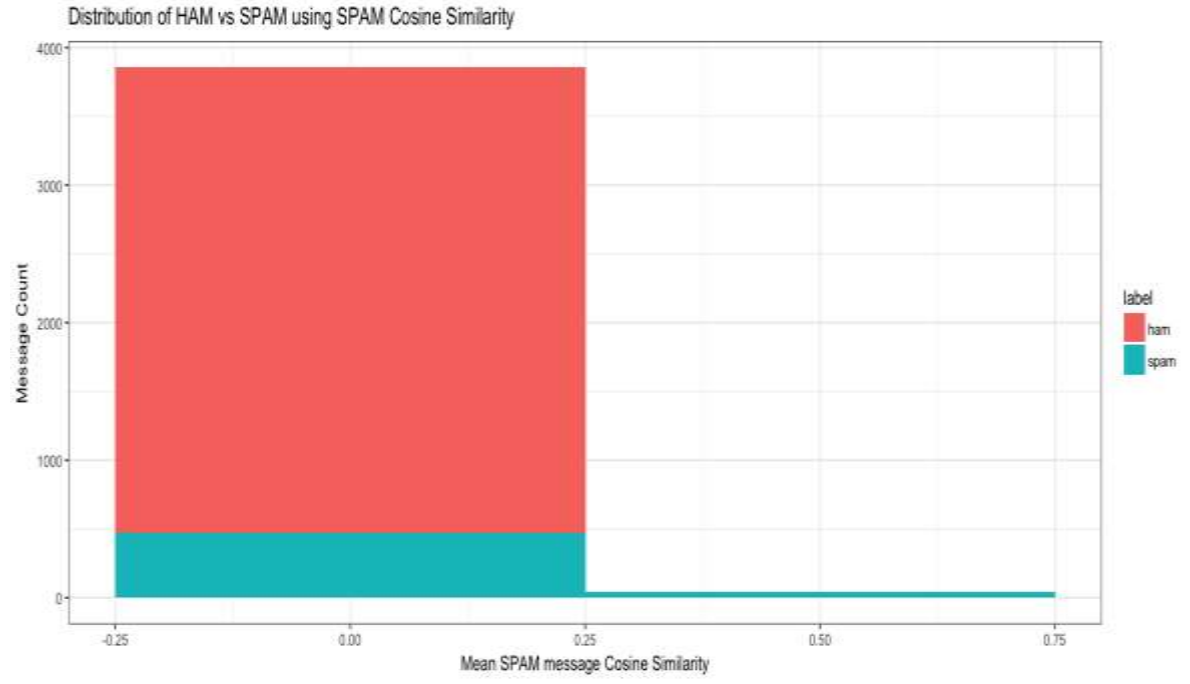

Fig. 2Mean spam message Cosine similarity vs message count

DOI10.17148/IJARCCE.2017.6819

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

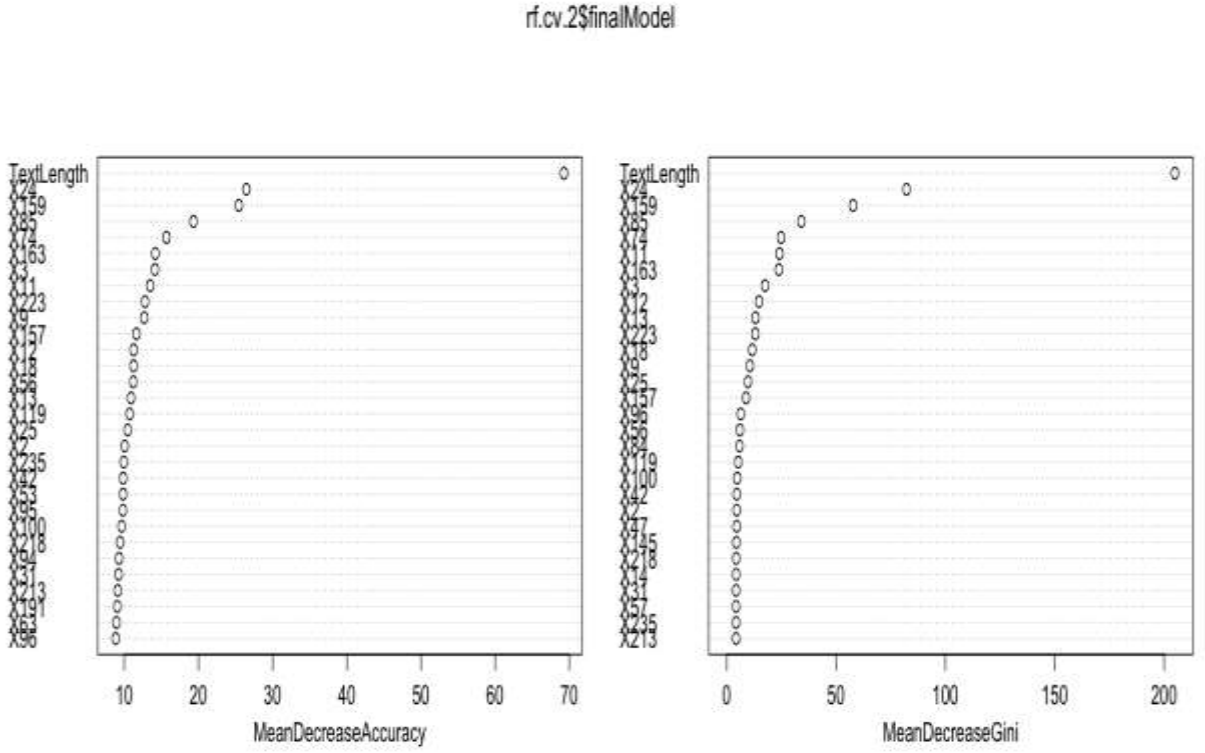

Fig 3: Shows the importance of the old parameters and Text length, in prediction and classification, the farthest one from Y axis being more important

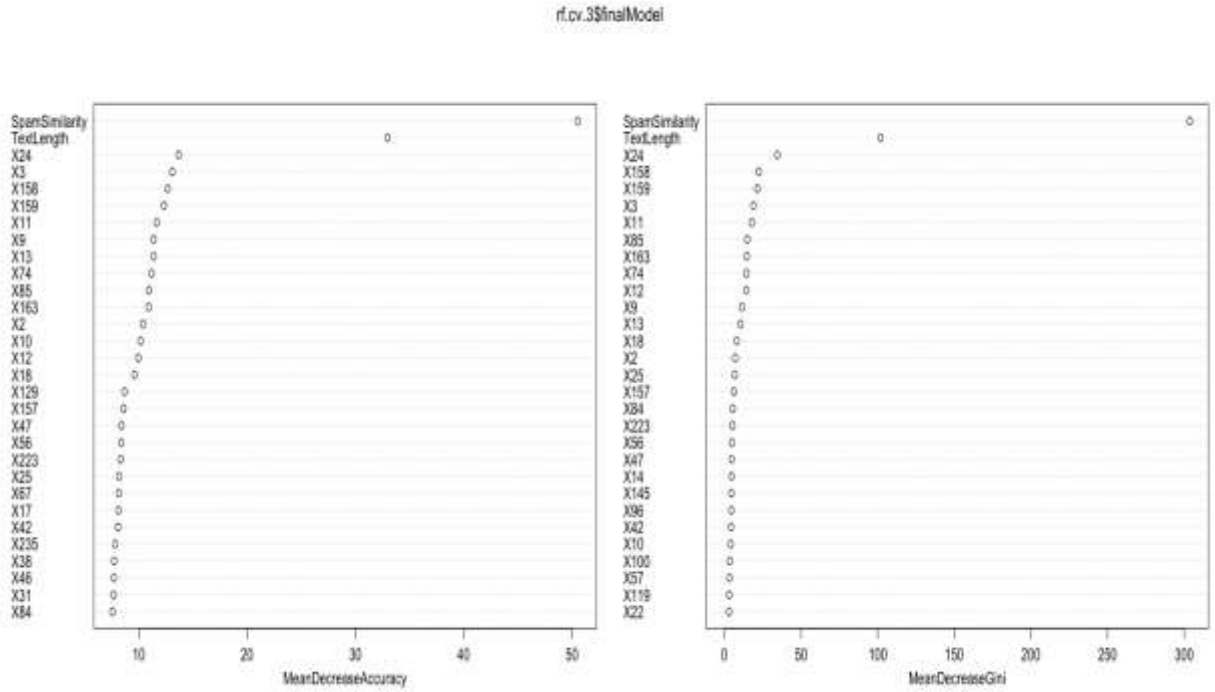

Fig 4: Feature Comparison Importance with Spam Similarity and Text length as additional feature

IV. CONCLUSION

In this study , random forest proved reasonably better than Decision Trees. Through a series of iterations, it can be concluded that Text Length plays a significant role in prediction and classification of text and spam messages. The

additional feature SPAM similarity which was initially thought to be key towards classification, as was evident from Fig 4., actually resulted in overfitting of the model. Finally, when this additional feature was removed, the model's efficiency improved significantly.

ACKNOWLEDGEMENT

Every venture requires the combined efforts of all those working on it. Working under the supervision of My Professor MS. ShaliniL.N. Rao has been very fruitful for me and I offer my sincere gratitude towards her. I would also like to thank my colleagues at work who have offered their insight in this field and have helped me and motivated me in achieving this goal.

REFERENCES

[1] ShahrukhTeli, PrashastiKanikar " A Survey on Decision Tree Based Approaches in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, 2015.
[2] Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood," Random Forests and Decision Trees", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.
[3] Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam," Comparative Study of Classification Algorithms used in Sentiment Analysis", International Journal of Computer Science and Information Technologies, Vol. 5 (5), 2014.
[4] UCI Machine Learning Repository, " http://archive.ics.uci.edu/ml/machine-learning-databases/credit-screening/ ".
[5] Applied Predictive Modelling : Kjell Johnson, Max Kuhn
[6] Introduction to Statistical Learning :http://www-bcf.usc.edu/~gareth/ISL/, Larry Wasserman, Matthew Richey
[7] W. B. Cavnar. ``N-gram-based text filtering for TREC-2." Proceedings of TREC-2: Text Retrieval Conference 2, Donna Harman, ed. National Bureau of Standards, August 1993.
[8] Claudia Pearce and Charles Nicholas. ``Using n-gram analysis in dynamic hypertext environments." In Proceedings of the Second International Conference on Information and Knowledge Management (CIKM '93), November 1-5 1993. (This paper was also released as UMBC technical report CS-93-10.)
[9] [Ding] Chris H. Q. Ding, A Similarity-Based Probability Model for Latent Semantic indexing Proc. of SIGIR'99, Berkeley, August 1999
[10] [Hull] D. Hull, Improving text retrieval for the routing problem using Latent Semantic Indexing, in Proceedings of the Seventeenth Annual International ACM-SIGIR Conference, 1994
[11] D. R. Cutting, J. O. Pedersen, D. R. Karger, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In Proceedings of the ACM SIGIR, 1992.
[12] A. Hotho, S. Staab, and G. Stumme. Word net improves text document clustering. In Proceedings of the SIGIR Semantic Web Workshop, Toronto,, 2003.
[13] Y. Zhao and G. Karypis. Evaluation of hierarchical clustering algorithms for document datasets.In Proceedings of the International Conference on Information and Knowledge Management, 2002.
[14] D. R. Cutting, J. O. Pedersen, D. R. Karger, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In Proceedings of the ACM SIGIR, 1992.

BIOGRAPHIES

Shekhar Tanwar is B.Tech. In computer Science and Engineering from Vellore Institute of Technology and is working as a Network Analyst at Accenture Services. Pvt. LTD. His areas of interest are Big Data, Machine Learning, Data Science, Text Analytics &Natural Language Processing.

Prof. Shalini. L is working as Assistant Professor (Senior) at School of computer Science and Engineering, VIT University. Her area of interest are data structure and algorithm, machine learning, natural language processing & Theory or computation. She has been with VIT for more than 10 years and has a total of 18 years of teaching experience.