WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

SPEECH ANALYSIS OF THE AMERICAN PRESIDENTIAL ELECTION

Marimuthu M*, Deepak Kumar C. R. *, Guhanesvar M*, Saagarika S*, Swetha Srinivasan*, Tilak Vijayaraghavan*

**M.Sc Decision and Computing Sciences, Coimbatore Institute of Technology, Coimbatore*

ABSTRACT

This paper shows the application of novel text mining techniques to analyze the fundamental differences of the campaign strategies between the Democratic and Republican parties by focusing on the campaign speeches of the 2012 American Presidential Election which pitted the incumbent Democrat Barack Obama against the Republican Mitt Romney. By analyzing the speeches made by both men in the course of the campaign the strategies and government objectives of each side are determined. With the use of text mining techniques such as text categorisation and similarity analysis techniques coupled with visual aids of the significant words and phrases, the fundamental differences between both parties become apparent

Keywords*: text mining, R language, sentiment, presidents*

1. INTRODUCTION

Every four years the American Presidential Election captures the imagination of not just the American public, but the entire global community. Such is the perceived influence of the elected leader of the 'Free World' in world economic, military and political issues, that in the year long run up to the election, blanket international media coverage is given to this divisive showdown between America's red and blue states. This practical paper analyses the fundamental differences between the Democratic and Republican parties by focusing on the campaigning strategies of the 2012 American Presidential Election which pitted the incumbent Democrat Barack Obama against the Republican Mitt Romney. By analyzing the speeches made by both men in the course of the campaign the strategies and government objectives of each side are determined. This not just highlights the key issues of the campaign but by using similarity measures it denotes if either side is espousing a clear programme for government or is giving particular sound bites which appeal to a certain sector of the electorate. Dissimilarity measures emphasize the opposing views of each side, as to how to take the country forward over the next four years. These measures also determine whether the nominee acts reactively to his opponent or takes an independent proactive stance on set issues. With the use of text mining applications such as text categorisation and similarity analysis techniques coupled with visual aids of the significant words and phrases, the fundamental differences between both parties become apparent.

2. DATA COLLECTION

The data for this paper was initially obtained from two different sources. The speeches, which totalled twenty five per candidate, were extracted using various web crawling and web extraction processes through the R language. The primary data for both presidential nominees was collected from the http://www.presidency.ucsb.edu/index.php website. Additional data for Mitt Romney's speeches which were subsequently not used due to a large number of duplications were extracted from the http://mittromneycentral.com/ website. Complications arose when trying to collect the data from the http://www.presidency.ucsb.edu/index.php website. A number of factors which may have contributed to this difficulty were the url structure combined with the number of speeches on the website. The website contains numerous amounts of previous election documents all of which were formatted with the same url structure as the specific speeches required. The collective size of these documents restricted the webcrawler's extraction ability even with adjusting the properties of the web crawler to deal with the maximum number of webpages, the depth and the maximum page size. To overcome this a .csv file was created for each candidate which contained the url of each specific speech required.

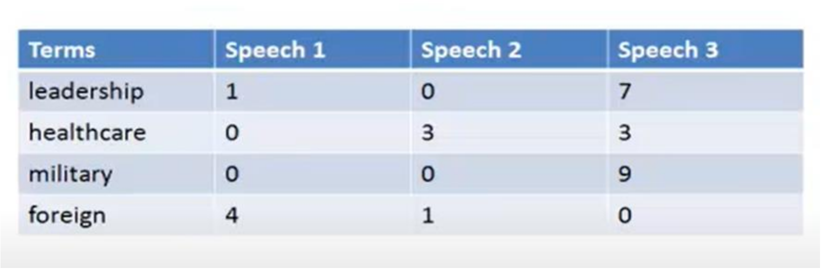
3. METHODOLOGY

To determine if different strategies were adopted by both the Democrat and Republican nominees, the following analysis techniques were performed:

1. Speech Categorization using Machine Learning Models
2. Speech Similarity
3. Word Usage
4. Sentiment Analysis
5. Emotion Summary

Speech Categorisation:

Initially to obtain an understanding of the different campaign approaches by both parties a categorization task was performed. This process also determined if it were possible to categorize additional unseen speeches by each candidate. Text categorization models were generated using various machine learning models. This technique can also be used to provide a word list detailing the frequency each candidate used various words (or tokens) in their campaign. A term document matrix is a way of representing the words in the text as a table (or matrix) of numbers. The rows of the matrix represent the text responses to be analysed, and the columns of the matrix represent the words from the text that are to be used in the analysis.To perform this we use a TermDocumentMatrix-function.



Terms	Speech 1	Speech 2	Speech 3
leadership	1	0	7
healthcare	0	3	3
military	0	0	9
foreign	4	1	0

Fig 1: Term Document Matrix

After cleaning the text data, the next step is to count the occurrence of each word, to identify popular or trending topics. Using the function TermDocumentMatrix() from the text mining package, you can build a Document Matrix – a table containing the frequency of words.

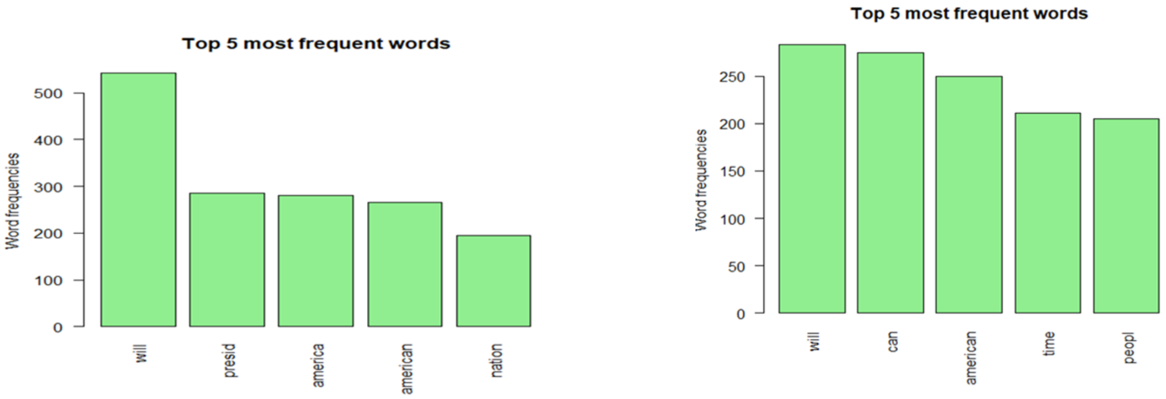


Fig 2: Term Frequency Plot of Obama and Romney's speeches

The data was then fit into three machine learning models namely:

1. KNN
2. Random Forest
3. Decision Tree

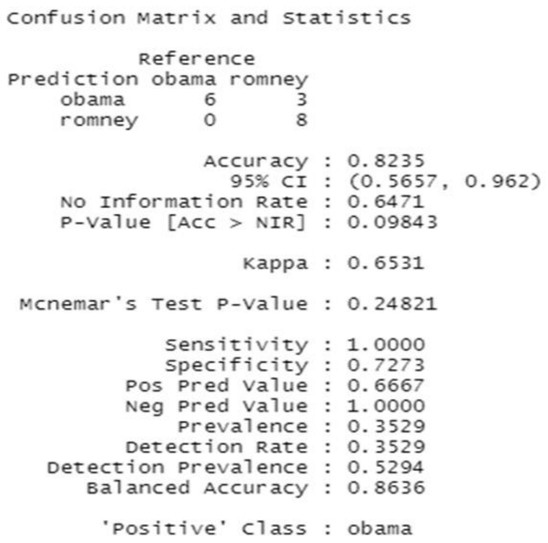
3.1 KNN Algorithm:

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. It is a versatile algorithm also used for imputing missing values and resampling datasets. As the name (K Nearest Neighbor) suggests it considers K Nearest Neighbors (Data points) to predict the class or continuous value for the new datapoint. The algorithm's learning is:

1. Instance-based learning: Here weights are not learnt from training data to predict output (as in model-based algorithms) but use entire training instances to predict output for unseen data.

2. Lazy Learning: Model is not learned using training data prior and the learning process is postponed to a time when prediction is requested on the new instance.

3. Non -Parametric: In KNN, there is no predefined form of the mapping function.

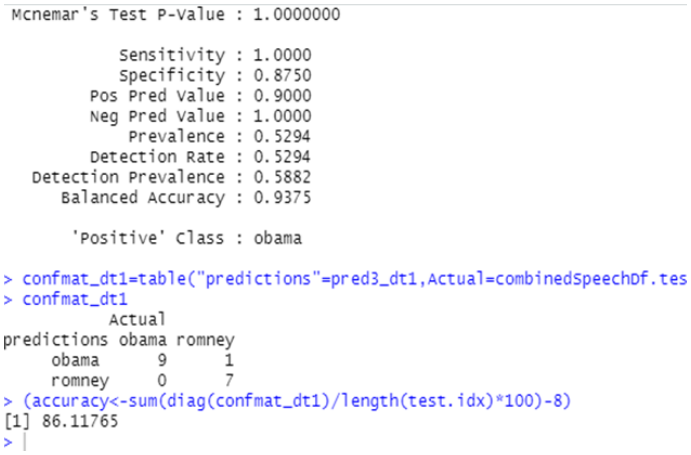


```
Confusion Matrix and Statistics

          Reference
Prediction obama romney
    obama      6      3
    romney     0      8

               Accuracy : 0.8235
                 95% CI : (0.5657, 0.962)
    No Information Rate : 0.6471
    P-Value [Acc > NIR] : 0.09843

                  Kappa : 0.6531

 Mcnemar's Test P-Value : 0.24821

            Sensitivity : 1.0000
            Specificity : 0.7273
         Pos Pred Value : 0.6667
         Neg Pred Value : 1.0000
             Prevalence : 0.3529
         Detection Rate : 0.3529
   Detection Prevalence : 0.5294
      Balanced Accuracy : 0.8636

       'Positive' Class : obama
```

Fig 3: Results of KNN Model

3.2 Decision Tree:

Decision tree is a graph to represent choices and their results in the form of a tree. The nodes in the graph represent an event or choice and the edges of the graph represent the decision rules or conditions. It is mostly used in Machine Learning and Data Mining applications using R. Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation. Decision tree learning or induction of decision trees is one of the predictive modeling approaches used in statistics, data mining and machine learning. It uses a decision tree to go from observations about an item to conclusions about the item's target value.

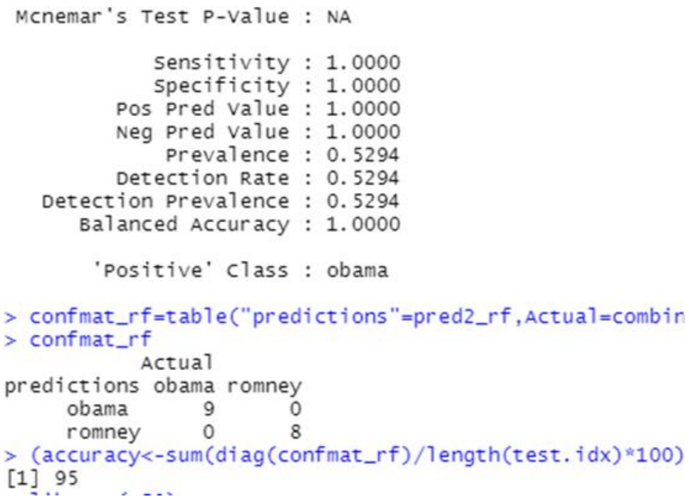


```
 Mcnemar's Test P-Value : 1.0000000

            Sensitivity : 1.0000
            Specificity : 0.8750
         Pos Pred Value : 0.9000
         Neg Pred Value : 1.0000
             Prevalence : 0.5294
         Detection Rate : 0.5294
   Detection Prevalence : 0.5882
      Balanced Accuracy : 0.9375

       'Positive' Class : obama

> confmat_dt1=table("predictions"=pred3_dt1,Actual=combinedSpeechDf.tes
> confmat_dt1
           Actual
predictions obama romney
     obama      9      1
     romney     0      7
> (accuracy<-sum(diag(confmat_dt1)/length(test.idx)*100)-8)
[1] 86.11765
> |
```

Fig 4: Results of Decision Tree Model

3.3 Random Forest:

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.



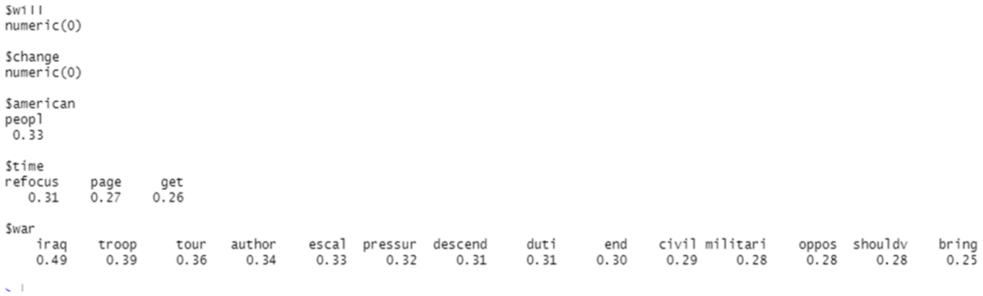
```
Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.5294
         Detection Rate : 0.5294
   Detection Prevalence : 0.5294
      Balanced Accuracy : 1.0000

       'Positive' Class : obama

> confmat_rf=table("predictions"=pred2_rf,Actual=combin
> confmat_rf
           Actual
predictions obama romney
     obama      9      0
     romney     0      8
> (accuracy<-sum(diag(confmat_rf)/length(test.idx)*100)
[1] 95
```

Fig 5: Results of Random Forest Model

Based on the results of the model, it can be inferred that the Random Forest model is the best fit for the data owing to high accuracy. Using a Random Forest model for this data will provide efficient results.

Speech Similarity:

To establish if a clear government programme and campaign approach was adopted, similarity techniques were generated to determine if any comparisons can be drawn from the speeches. These techniques were also explored by [6] and [7]. Speeches which are similar to each other would indicate a more clear and consistent campaign approach, whereas speeches which are dissimilar would point towards an inconsistent and ad-hoc campaign process. Three similarity analysis techniques were generated; the similarity of Barack Obama's speeches, the similarity of Mitt Romney's speeches and a similarity measure of both nominees' speeches. The final analysis provides more insight into determining if both parties placed similar emphasis on the same topics. All similarity models were generated using the same operators as can be seen in Table 2 These models consisted of extracting the speeches from the website using different .csv files for each party. A combined .csv file was created to determine the similarity of both candidates' speeches together with text files 1-26 denoting Obama's speeches while text files 27-52 signifying Romney's speeches. Given the general focus of identifying similarities we have only used the Cosine Similarity measure which measures the cosine of the angle between two vectors of an inner product space. Correlation is a statistical technique that can demonstrate whether, and how strongly, pairs of variables are related. This technique can be used effectively to analyze which words occur most often in association with the most frequently occurring words in the survey responses, which helps to see the context around these words



```
$will
numeric(0)

$change
numeric(0)

$american
peopl
 0.33

$time
refocus    page    get
   0.31    0.27   0.26

$war
    iraq    troop     tour   author    escal  pressur  descend     duti      end    civil militari    oppos  shouldv    bring
    0.49     0.39     0.36     0.34     0.33     0.32     0.31     0.31     0.30     0.29     0.28     0.28     0.28     0.25
```

Fig 6: Results of Word Similarity

Word Usage:

A word cloud is one of the most popular ways to visualize and analyze qualitative data. It's an image composed of keywords found within a body of text, where the size of each word indicates its frequency in that body of text. Using the word frequency data frame (table) created previously to generate the word cloud. The following word clouds were analyzed to infer the President's performance.

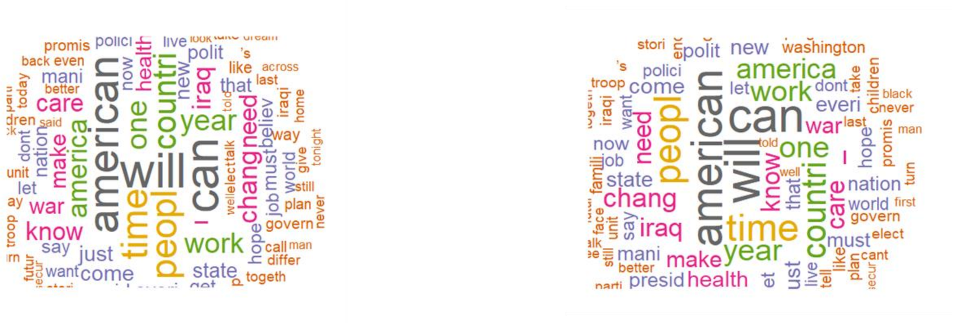
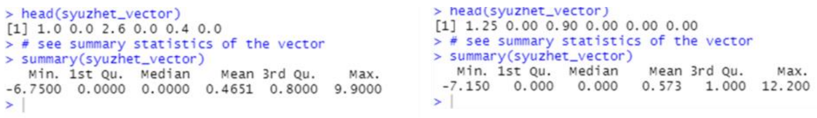


Fig 7: Word Cloud generated for Obama and Romney's speeches

Sentiment Analysis:

Sentiment analysis (or opinion mining) is a natural language processing (NLP) technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs. Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helps a business to understand the social sentiment of their brand, product or service while monitoring online conversations. However, analysis of social media streams is usually restricted to just basic sentiment analysis and count based metrics. This is akin to just scratching the surface and missing out on those high value insights that are waiting to be discovered. Sentiments can be classified as positive, neutral or negative. They can also be represented on a numeric scale, to better express the degree of positive or negative strength of the sentiment contained in a body of text. The Syuzhet package for generating sentiment scores, which has four sentiment dictionaries and offers a method for accessing the sentiment extraction tool developed in the NLP group at Stanford.



```
> head(syuzhet_vector)
[1] 1.0 0.0 2.6 0.0 0.4 0.0
> # see summary statistics of the vector
> summary(syuzhet_vector)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-6.7500  0.0000  0.0000  0.4651  0.8000  9.9000 
> |
```

```
> head(syuzhet_vector)
[1] 1.25 0.00 0.90 0.00 0.00 0.00
> # see summary statistics of the vector
> summary(syuzhet_vector)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-7.150   0.000   0.000   0.573   1.000  12.200 
> |
```

Fig 8: Results of Sentiment Analysis

Emotion Summary:

Emotion classification is built on the NRC Word-Emotion Association Lexicon (aka EmoLex), is "The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The annotations were manually done by crowdsourcing."

To understand this, explore the get_nrc_sentiments function, which returns a data frame with each row representing a sentence from the original file

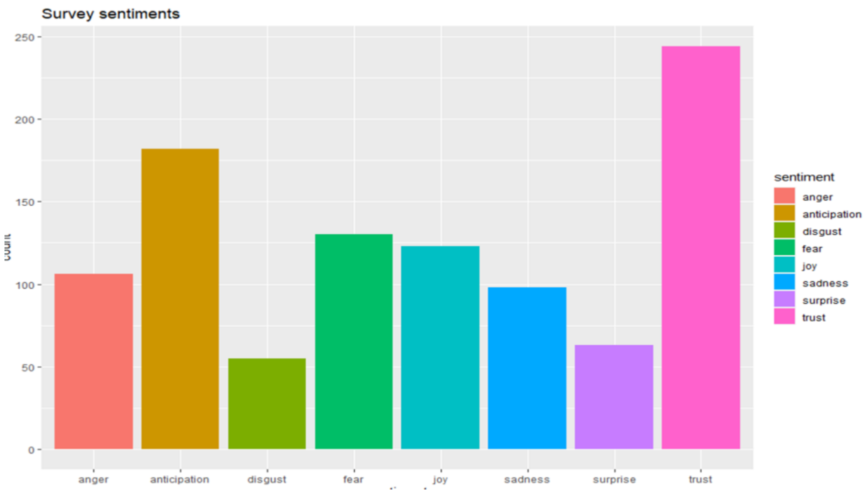


Fig 9: Survey Sentiments for Obama's speech

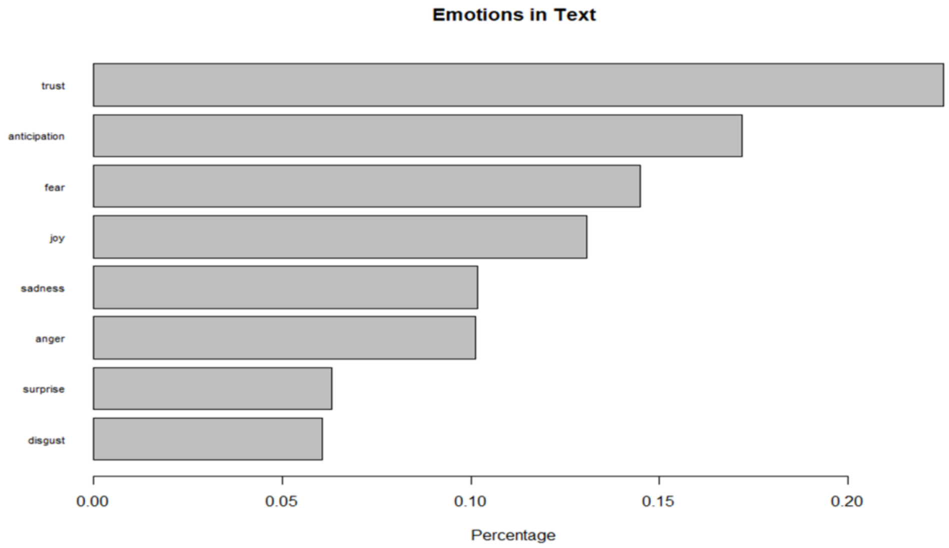


Fig 10: Emotions in Text for Obama's speech

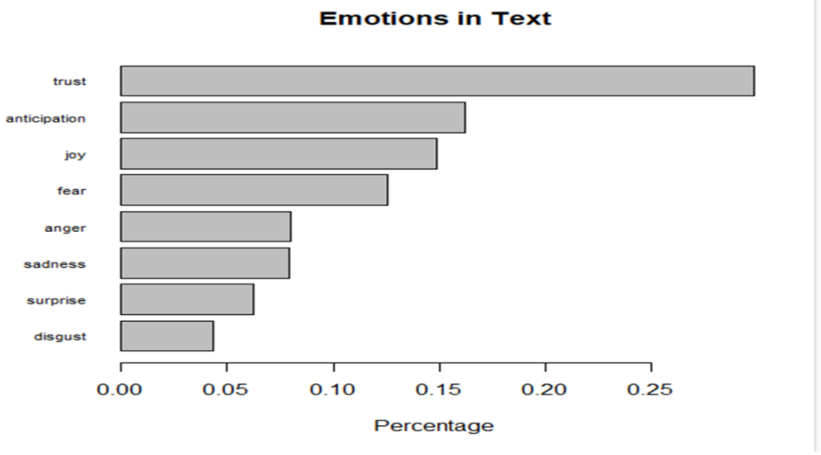


Fig 11: Survey Sentiments in Romney's speech

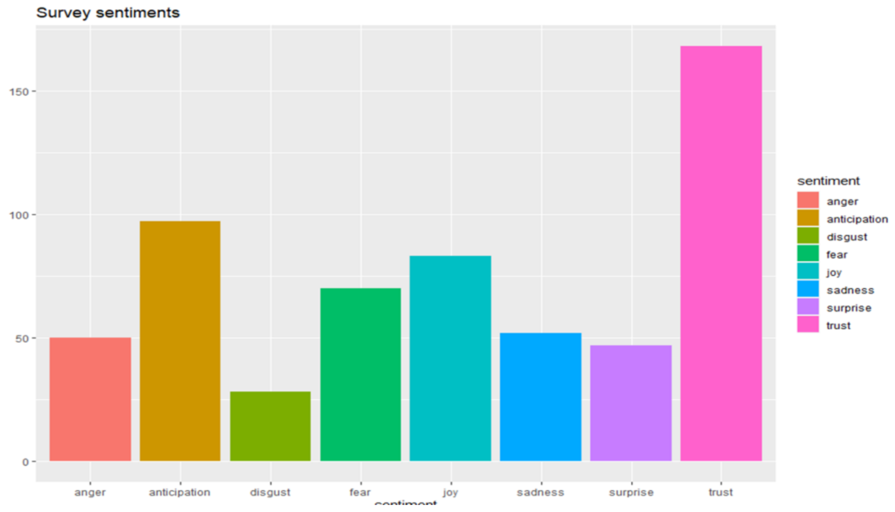


Fig 12: Emotions in Romney's Speech

4. CONCLUSION

This paper provided an analysis of some of the campaign speeches of the 2012 American Presidential Election outlining the fundamental differences between the Democratic and Republican parties. By utilising text mining applications such as text categorisation, similarity analysis and visual aids, the significant words and phrases used by Barack Obama and Mitt Romney were made apparent. The analysis of the Obama/Romney speeches using the techniques above clearly identified the key policy issues for each side, the style of the respective campaign and the character of the strategy. In summation Obama focussed solely on the domestic economy, job creation and appealing to the middle classes, by stating ”if you vote for me I will help you, I will fight for you in congress”. He also emphasised the diverse nation of the country but that the strength of the country and the hopes of the future lay in unifying all peoples. His strategy was consistent and cohesive, focussing on the strength of a unified people in difficult economic times. The analysis presented here shows that he did mention many of the keywords identifying current problems in the country and mentioned them more frequently than his rival, demonstrating a perceived willingness to address these issues. Romney’s strategy was to attack Obama and his policies from the outset. As a result of his target audience he jumped from issues of the threat of state assisted welfare states, to the nuclear threat posed by Iran and Cuba, to the need to build up a foreign defence system with allies such as Israel. He spoke about everything that was wrong with Obama and his policies and the deficit, but offered no concrete solutions. Romney was not proactive in his speeches but reactive to everything he perceived the Obama administration to be about. The dissimilarity in the speeches highlights the polar opposite views of both sides. It also illustrates the consistency of the Obama strategy and the fragmented and alienating nature of the Romney campaign, reinforcing the fundamental differences between both sides.

REFERENCES

[1] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. Yale: Rapid prototyping for complex data mining tasks. In Lyle Ungar, Mark Craven, Dimitrios Gunopulos, and Tina Eliassi-Rad, editors, KDD ’06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 935–940, New York, NY, USA, August 2006. ACM.

[2] Murphy Choy, Michelle LF Cheong, Ma Nang Laik, and Koo Ping Shung. A sentiment analysis of Singapore presidential election 2011 using twitter data with census correction. arXiv preprint arXiv:1108.5520, 2011.

[3] Scott P Robertson. Changes in referents and emotions over time in election-related social networking dialog. In System Sciences (HICSS), 2011 44th Hawaii International Conference on, pages 1–9. IEEE, 2011.

[4] Adam Bermingham and Alan F Smeaton. Classifying sentiment in microblogs: is brevity an advantage? In Proceedings of the 19th ACM international conference on Information and knowledge management, pages 1833–1836. ACM, 2010.

[5] Kevin Coe. George w. bush, television news, and rationales for the iraq war. Journal of Broadcasting & Electronic Media, 55(3):307–324, 2011.

[6] Marco R Steenbergen, Andr´e B¨achtiger, Markus Sp¨orndli, and J¨urg Steiner. Measuring political deliberation: a discourse quality index. Comparative European Politics, 1(1):21–48, 2003.

[7] Robert Klemmensen, Sara Binzer Hobolt, and Martin Ejnar Hansen. Estimating policy positions using political texts: An evaluation of the wordscores approach. Electoral Studies, 26(4):746–755, 2007.