National College of Ireland

Predicting Highest Facebook Reaction Count through News Articles

MSc Research Project
Data Analytics

Saad Siddique
x15040704

School of Computing
National College of Ireland

Supervisor: Dr. Cristian Rusu

National College of Ireland
Project Submission Sheet – 2015/2016
School of Computing

Student Name:	Saad Siddique
Student ID:	x15040704
Programme:	Data Analytics
Year:	2017
Module:	MSc Research Project
Lecturer:	Dr. Cristian Rusu
Submission Due Date:	11/12/2017
Project Title:	Predicting Highest Facebook Reaction Count through News Articles
Word Count:	6175

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	9th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting Highest Facebook Reaction Count through News Articles

Saad Siddique
x15040704
MSc Research Project in Data Analytics

9th December 2017

Abstract

News organizations have been instrumental in holding governments and other powerful figures accountable for their actions. They have change social perspectives and guide their readers towards the desired objectives. The relative decline of traditional media, such as television and paper, that once attracted millions of people, has forced news organizations to shift their focus towards social media to increase viewership. Using 'like', share and comment counts to understand the readers has been fruitful, but the news organizations should utilize the new reaction buttons introduced on Facebook. My objective is demonstrate how these news organizations can analyze their own, and their competitors, new articles to predict which highest reaction they are likely to receive. The conclusions drawn from my dissertation is that it is possible. However, caution must be drawn for scenarios when predicting controversial news article.

1 Introduction

Social media has become an integral part of daily life. It was initially used for communicating with friends, and family, as well as dating. Overtime, it expanded to encompass e-commerce, political discussions, brand promotion and connecting those with common interests with one another (Udanor et al.; 2016). In addition, many news organizations use social media sites to increase their viewership, by posting news article links, and videos. This has resulted in users generating hundreds of terabytes of data daily.

Facebook introduced reaction buttons, to complement their 'Like' button, at the end of February 2016 (Stinson; 2016). These new reactions are 'Love', 'Haha', 'Wow', 'Sad' and 'Angry'. These new reaction buttons have given people more freedom in expressing their opinion with a single press and swipe. It has also opened a new resource for researchers and organizations to utilize towards generating content that would get a desired reaction.

With the advantage of attracting more viewers online than through traditional media, news organizations compete with one another vigorously. In order to gain advantage over competitors, it is important for a news organization to understand their readers, and how

the reader reactions towards their published article online. It is also important to know how their competitors‘ readers would react to their news articles. In addition, knowing how the reader would react could also allow the marketing department to create improved targeted ads that would utilize the readers emotional state upon reading the article.

Pilon (2016) provided a guide to businesses on how to interpret reactions. Peterson (2015) and Greenberg (2016) have stressed the benefits that marketers and advertisers can gain by evaluating these reactions. Turnbull and Jenkins (2016) explains how media campaigns can utilize these reactions to better understand how users react towards posted content. Given these potential applications of reactions, Badache and Boughanem (2017), Pool and Nissim (2016) and others have used them in their research. They find the reactions to have significant implications in their studies.

For the purpose of this dissertation, the main objective is to analyze if text analysis of news articles can be used to predict user reactions on Facebook. This objective is divided into 3 research questions, in order to better understand the main objective:

1. Can news articles predict which (Love, Haha, Wow, Sad, Angry) reaction gets the highest count from Facebook?

2. Can news articles predict which groups of (Love And Haha, Wow and Haha, Sad and Wow, etc) reactions gets the highest count from Facebook?

3. Does the first research question differ for articles posted by different news organizations?

In order to answer these questions, articles posted by the following news organizations are used: Yahoo News, CNN, New York Times, Fox News, NBC News, Washington Post, The Guardian, ABC News, BBC News, USA Today and Los Angeles Times. The official Facebook posts and news articles of these 11 news organizations are extracted, cleaned and analyzed. The data extracted are from March 2016 till June 2017.

Research is conducted to check if reactions can be predicted, with regards to news articles posted on Facebook by different news organizations. More than one years worth of posts, along with the reactions, are extracted from news organizations official pages on Facebook. Using the article links in the post, the main content of the article is extracted from the news organizations site. After cleaning the extracted news articles, text analysis through LIWC is conducted. The results of the text analysis are joined with the original Facebook post. Finally, Support Vector Machine is used to address the above research questions. Using the results of the classifier, each of the questions are separately analyzed and conclusions drawn.

The rest of this paper is organized as follows: Section 2 covers related works and provides a literature review, where other research, addressing similar questions, is discussed in detail. Section 3 discusses the Research Methodology, while section 4 discusses implementation. Section 5 address the research questions stated above. Section 6 gives the conclusion and discusses future work.

2 Related Work

2.1 Introduction

As of the third quarter of 2017, there are more than 2 billion active monthly users on Facebook[1]. With Facebook connecting large numbers of people, it has provided many academic researchers and organizations with the opportunities to analyze and understand social media users. They utilize Facebook to understand human behaviour. As such, Machine learning techniques have been extensively used to analyze people, and their interactions on Facebook.

2.2 Limited Collection of User Reactions: Like

The Facebook 'Like' button has been extensively researched by both academics and organizations. Most journal articles show that the like button has a positive impact on the post. People having greater connection to the post have a higher chance of pressing the like button (Pelletier and Blakeney Horky; 2015) than those who do not have the connection. In addition, people use the like button for a wide range of reasons, like maintaining a conversation and social ties, and dating efforts (Eranti and Lonkila; 2015). Even journal articles that received one or more likes on Facebook are likely to have a higher number of citations (Ringelhan et al.; 2015). The number of likes on a post about movies, before their release date, has a wide impact on the box office performance of movies (Ding et al.; 2017).

However, while there are many different benefits associated with 'like', it is not universally acknowledged as the ideal candidate to focus upon. Brands that respond to complaints quickly, develop loyal relationships with customers, and understand the fact that people have greater control over their brand image online (Avery et al.; 2017), can gain greater online presence and followers than many others. In addition, Winter et al. (2015) have demonstrated that positive comments or likes do not influence a readers perspective on a news story or its content.

While researchers do not always agree on the significance, or insignificance, of the 'like' button, they do agree that 'like' button limits the users interaction towards 'affirmation' of the post. This limited interaction makes it difficult to 'like' posts with negative contexts, like natural disasters, deaths or crime. With the addition of five different reaction buttons, this limitation has been significantly mitigated. As such, it allows many researchers to improve their existing studies.

2.3 Extended Collections of User Reactions

For years, many studies have been conducted on Facebook that focused on 'Like', share and comment counts. This is common to studies conducted on Twitter, YouTube and other social media networks. The Facebook Reactions provide the opportunity to conduct a wider range of studies. Search Engines use social signals (e.g. like, +1, rating) as a source of evidence for their retrieval ranking systems. Ranking functions based on

[1] `https://s21.q4cdn.com/399680738/files/doc_financials/2017/Q3/Q3'17-Earnings-Release.pdf`

social and sentic features of users feedback in YouTube, perform better than those based on basic features (e.g. title, tags) (Orellana-Rodriguez et al.; 2014). As such, Badache and Boughanem (2017) studied the impact of Facebook reactions, and concluded that incorporating Facebook reactions can also improve the retrieval ranking performance. In addition to analyzing each reaction individually, they also demonstrate that analyzing combinations of reactions is also useful for analysis.

Research has been conducted to detect emotion from a Facebook post. Pool and Nissim (2016) used Facebook reactions to train a support vector machine to detect emotion. By comparing their results with other emotion detection methodologies, they achieve competitive results without relying on handcrafted resources. Tian et al. (2017) analyze the correlation between emoji usage and Facebook reactions, and conclude that Facebook reactions and comments can be used to investigate a user's emotional attitude towards a post. In addition, they demonstrate that it is possible to detect positive emoji usage under negative context, and vice versa. Guzman et al. (2016) were able to predict reactions by analyzing only the headlines posted on a Facebook post. They have identified Support Vector Machine as the most ideal classifier when using Facebook reactions, though it has been stated that other classifiers can perform better by the changing parameters of the classifiers and feature models.

The first research question is inspired by Guzman et al. (2016) research. In their research, they analyze headlines posted on Facebook to predict the reactions. However, in first research question, the whole news article text, extracted from news organizations websites, is analyzed to predict reactions. Badache and Boughanem (2017) inspired the second research question. They demonstrated the usefulness of analyzing combinations of reactions.

2.4 Linguistic Inquiry and Word Count Text Analysis

Linguistic Inquiry and Word Count (LIWC) is a text analysis software which is extensively used by many researchers. It analyses text provided by the user and returns psychologically meaningful categories in numeric form(Tausczik and Pennebaker; 2010). LIWC provides over 95 psycho-linguistic categories, including Functional Words, Perceptual Processes, Biological Processes, Drives, Effect, Social, Cognitive Process etc..

In order to investigate how citizens select election news online, Jang and Oh (2016) used LIWC to analyze the news headline and subhead to get the positive and negative emotional percentage. To understand the relationship between political ideology of U.S. politics and mind-body focus, Robinson et al. (2017) used LIWC to analyze news articles, presidential state of union address and layperson writing sample. Their conclusion demonstrated that text written liberal ideologies score positively in mind-body terms, while conservative ideologies score negatively in mind-body terms. Hwong et al. (2017) uses LIWC to analyze if social media messages that focus on space science posts, have unique psycho-linguistic features.

Due to its extensive usage by many researchers and extensive text analysis capabilities, LIWC is used to analyze the news article to address all three research questions.

3 Methodology

For the dissertation, R Programming language is primarily used. It is an open source statistical software, that incorporates features from both object oriented programming and functional programming (Matloff; 2011). It has a dedicated online community, that contributes libraries and updates that are essential for a data analyst. It is used for data extraction, cleaning and transforming, as well as implementing classifiers, such as Support Vector Machine (SVM).

There are many different text analysis libraries and software available. For the dissertation, LIWC is used to analyze the news articles. Unlike many other text analysis libraries, LIWC is robust and does not require an extensive amount of text cleaning prior to analyzing. For analysis of the results of SVM for each research question, Tableau is used to create graphs and complement the evaluation.

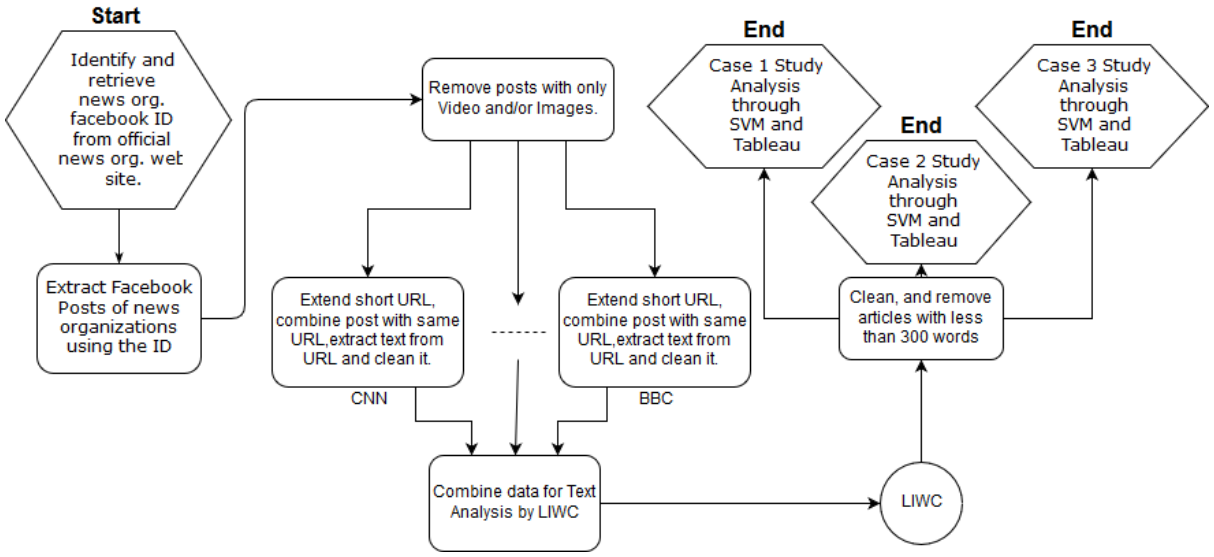

Figure 1: Flow Chart

3.1 Facebook Posts Data

In order to ensure the best possible analysis for this dissertation, news organizations with largest number of followers, as of August 2017, are selected[2]. To ensure that 'Fake News' does not have any influence over the analysis, the Facebook ID of the original news organizations is extracted from their official web site. No unofficial news organizations Facebook ID is included in the Facebook post extraction process.

As this dissertation is about analyzing the Facebook reactions, news organizations' Facebook post extraction starts from the 1st March 2016. This is when Facebook reactions were introduced world wide (Stinson; 2016). The extraction ends before June 2017. This is because Facebook introduced a new reaction, Pride, to celebrate LGBT pride month[3]. As the analysis requires Facebook posts data containing the reactions buttons,

[2] `http://fanpagelist.com/category/news/world/`

[3] `https://www.ajc.com/news/facebook-introduces-rainbow-reaction-celebrate-lgbt-pride-month/93HZe0ZEX8FJ0bStVgl9CJ/`

that spans more than a year, the new Pride reaction is not included. As the Pride reaction would interfere with the analysis, the Facebook posts extraction is limited till 31st May 2017.

To extract the news organizations posts from Facebook, Rfacebook[4] library is used. The benefit of using this API is that, even with free Facebook developer credentials, it ensures that it extracts as many posts as possible on a given day. Due to the limitation of 25 post returns on free Facebook developer credentials, the API will call the extraction service multiple times in the same day. As there is a limit to calling the extraction service 200 times per hour, this causes the API to generate errors when it calls the extraction service beyond 200 time per hour. By using a combination of 'Sys.sleep' and 'tryCatch' functions of R, it is possible to make the extraction process completely automated. The data retrieved is saved in CSV files periodically, in case the 'tryCatch' function is unable to capture the error generated.

As new organizations Facebook posts contain multimedia content, they need to be cleaned. This is done by removing the posts that only contained either videos or images. Only the posts with news article links remain.

3.2 News Article Data

The links to access the news article are provided by Facebook posts. However, the URL provided are shortened. This causes some issues when accessing the website for article extraction. To resolve this issue, longurl[5] library is used to expand the short URL to it original format. The original URL format is also checked to make sure it redirects towards the actual article, and not towards the main page of the news organization.

There have been cases of some Facebook posts pointing towards the same news article. To resolve this issue, the original URL links are compared with one another. If they are the same, the original URL link posts are combined. The reaction counts will add with their respective reactions counts in other posts.

In order to extract the news articles, the library boilerpipeR[6] is used. Unlike the Facebook extraction process, there are no limitations to extraction. However, the API does generate java related errors in some cases. To ensure that the errors are minimized and handled appropriately, 'tryCatch', 'withCallHandlers' and 'withRestarts' functions of R are used.

3.3 News Article Analysis

Unlike the unstructured text commonly available online, the news articles are relatively clean. However, they are still required to undergo text cleaning before analysis. The cleaning involves removing text that does not contribute towards the news article, like 'Click here to read more' or 'Newsletter Sign Up'. Each news organization article had to

[4] `https://cran.r-project.org/package=Rfacebook`
[5] `https://cran.r-project.org/package=longurl`
[6] `https://cran.r-project.org/package=boilerpipeR`

be cleaned separately from other news organization articles.

Even though LIWC is more robust than other text analyses, it still requires the cleaning and transforming of text before analysis. Grammar, capitalization and sentence structure do not need to be corrected. However, meaningful abbreviations, like month 'Dec', should be December. Common abbreviations like 'Dr.', 'Ms.' or 'Mr.' need to have their periods removed. Common internet notations, like email address, URL address, hash tag and twitter handler need to be changed. Other common problems like 'w/' or 'b/' also need to be changed.

Once the text is ready for analysis on LIWC, a csv files is created, containing all the news articles extracted. The csv file is analyzed by the LIWC software, and a new csv file is generated. The new csv file contains 95 LIWC text analysis components of each news article.

The new csv file is read by R and preprocessed for the main analysis. The results of the text analysis conducted by LIWC are based on the total number of words the text contains. Text with 10000 words give far more reliable results than 100 words[7]. So, in order to conduct analysis with reliable results from LIWC, news articles with word count less than 300 are removed.

3.4 Main Analysis

To evaluate the research question, multiple datasets are made from the main data for analysis. The 'like' reaction is removed from the analysis, as it has the highest count across the entire data, as shown in Fig 2, and would cause issues with the evaluation.

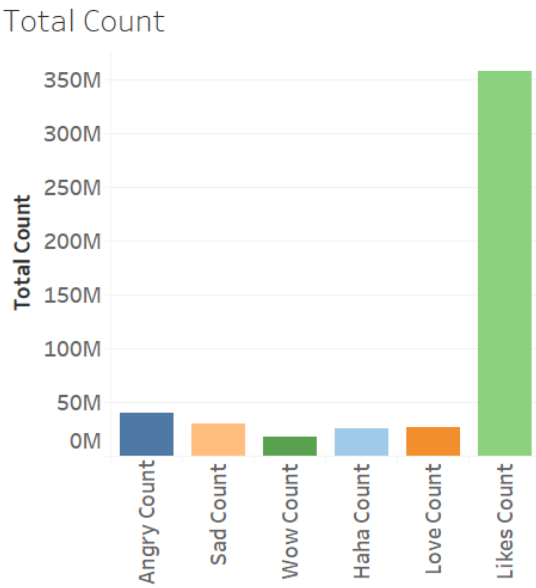

Figure 2: Total Count

For each research question, multiple datasets are created to handle each reaction. In the case of second research question, multiple datsets are created to handle each group of

[7] https://liwc.wpengine.com/how-it-works/

reactions. Each dataset is given an additional column, 'Category', which identifies which news articles have the highest particular reaction, relevant to that dataset. For example, in a dataset handling 'Angry' reaction, the 'Category' is set to 1, if that news article has 'Angry' count greater than other reactions count. Otherwise, the 'Category' is set to 0. For the second research question, a dataset handling 'Wow' and 'Haha' group, the 'Category' is set to 1, if the news article has 'Wow' and 'Haha' count greater than others reactions. Otherwise, the 'Category' is set to 0. Please note that count between 'Wow' and 'Haha' reactions is not considered, only that they are individually greater than others reactions count.

To conduct the research question analysis, the following four libraries are used. The 'splitstackshape'[8] library is used to stratify the data, ensuring that the data analyzed are balanced. The 'caret'[9] library is used to divide the data into test and train sets, and calculate the confusion matrix of the classifier, which is analyzed in the Evaluation section of the dissertation. The 'e1071'[10] library is used for conducting the analysis through Support Vector Machine (SVM).

In order to reduce the bias associated with one-shot random sampling, each analysis is conducted 10 times with different sets of stratified data and different test and train sets. The resulting data generated is transferred to Tableau, where the average is calculated, and the standard deviation is also calculated to ensure consistency between each analysis.

In addition to SVM, other classifiers were also considered for main analysis. Guzman et al. (2016) also used Stochastic Gradient Descent and Multinomial Naive Bayes classier but found SVM to perform the best. Decision Tree and Artificial Neural Network are tested for the dissertation, but did not perform as well as SVM. Hence, SVM is the main classifier used for all three research questions.

4 Implementation

4.1 Facebook Post Retrieval

In order to retrieve data from Facebook, a developer account is created on Facebook, and an application is created. The application ID and secret ID are retrieved.

After deciding on which news organizations to cover, their official Facebook account IDs are retrieved from their official websites. Using the Rfacebook library, the application account ID and secret ID are authenticated. A date list is then created, using the 'seq' function to retrieve each day between 1st March 2016 till 1st June 2017.

A 'for' loop is used to sequentially go through all the days. Inside the loop, the 'getPage' function of the Rfacebook API is used to retrieve all the posts posted on that particular day, using the specified news organization ID. The 'getPage' function will call the a post retrieval few time, in order to get all the posts for that day. On average, the 'getPage' function will call the post retrieval function 3 to 4 times, to retrieve all

[8] `https://cran.r-project.org/package=splitstackshape`
[9] `https://caret.r-forge.r-project.org`
[10] `https://cran.r-project.org/package=e1071`

the posts posted by the news organization in a day. As there is a limit of 200 calls per hour on Facebook, the 'getPage' function is called between 60 to 70 times before the 'Sys.sleep' function is called to hold the loop for 1 hour. The 'tryCatch' will handle the errors generated by the 'getPage'.

Among the information retrieved from Facebook, it also provides 'type'. It gives 3 categories regarding the post - video, photo and link. As per requirement of the dissertation, only the posts with links are saved and the rest are discarded.

4.2 News Article Retrieval

On Facebook, the news organizations save short URL links. In half the cases, directly retrieving the news articles from the short URL link causes problems with boilerpipeR library. To resolve this issue, the 'longurl' library is used to convert the short URL to the original URL, by using the 'expand_urls' function.

Once the original URL has been retrieved, it is checked to make sure that they redirect towards the actual article, and not the main web page of the news article. All the URLs below a certain size are collected and visually inspected. If they redirect towards the home page, they are discarded.

There have been cases, while inspecting the articles retrieved, that news articles were repeated across multiple posts. Using R code, these article original URLs are identified, and their relevant posts are combined. This eliminates article repetition, and still preserves their relevant Facebook information.

The library boilerpipeR is used to retrieve the news articles from the original URLs, with 'ArticleExtractor' function. While the library does not have the same limitations as the Facebook retrieval process did, it still generated errors that needed to be handled.

To handle scenarios where the extraction process toke too long, the 'evalWithTimeout' function was used to give the 'ArticleExtractor' 4 seconds to retrieve the article. Otherwise, it would interrupt the process and generate an error. BoilerpipeR library, which is provides an interface to the Boilerpipe Java Library, generates java related errors in some scenarios. As 'tryCatch' is unable to handle java related errors, a combination of 'withCallHandlers' and 'withRestarts' functions is incorporated with 'tryCatch' functions to handle the java related errors. In order to ensure that the article is extracted, the same original URL is called 3 times if the first attempt at retrieval fails.

4.3 News Article Cleaning, and Analysis

Once all the articles are extracted from a news organization website, they undergo a cleaning process before they are analyzed by LIWC. The articles have to be grouped by news organization, as each news organizations articles require a different cleaning process. Using combinations of 'sapply' and 'gsub' functions, the news article text is cleaned. The same combination is also used to clean the text as required by LIWC Operator Manual[11].

[11] `https://liwc.wpengine.com/wp-content/uploads/2015/11/LIWC2015_OperatorManual.pdf`

When the news article texts are cleaned, they are all combined. After they are combined, they are saved in .csv file. The LIWC software uses the .csv file to analyze all the news articles extracted, and generates a new .csv file with additional text analysis columns. The new csv file is then cleaned in R, and articles with word count less than 300 are removed.

4.4 Main Analysis

Initially, 'set.seed(15040704)' is set in the R code. This is to ensure that the results of the analysis conducted on R are reproducible.

Once the data reactions has been categorized, columns of data that are not relevant to the analysis are removed. The 'Category' column is factorized for classifier analysis purposes. In case of first and second research questions, two 'for' loops are used to analyze the categorized data. The external 'for' loop selects the specific reaction/s data to analyze, and the inner 'for' loop is to run the same analysis 10 times. In case of the third research question, three 'for' loops are used. The first 'for' loop is used to select the specific reaction data to analyze, the second 'for' loop is used to select the news organization, and the third 'for' loop is used to run the same analysis 10 times.

Within the inner most loop, the selected reaction data is stratified using the 'stratified' function. The function takes the data, 'Category' as the column to identify for grouping and the total number of 'Category' that are identified as 1. This is to ensure that the resulting dataset is balanced and has equal numbers of 'Categories' with 1 and 0. This is also the primary reason why entire main analysis is run 10 times. The total number of 'Category' that are identified as 1 are small, when compared to the entire dataset. Hence the need to run it 10 time so that there is greater chances of selecting 'Category' identified as 0 that were not select in the previous analysis. Thus ensuring that the end result is consistent, and unambiguous.

The 'createDataPartition' function is used to create datasets for testing and training the SVM. 'Category' is identified as the vector to considered when creating the testing and training sets. 75% is the training set size, and 25% is the testing set size of the total stratified dataset. The List parameter is set to false. Running the main analysis 10 times also allows the function to used different sets of data for each analysis, within the percentage set.

The 'svm' function is then used to run the SVM analysis using the training dataset. The function is used is set to default settings. The 'Category' is set as the dependent variable, and the rest of the data is set as independent variables. The data is scaled internally. The type used is categorical classification. Cache size is set to 40. Tolerance of termination criteria is set to 0.1. Shrinking Heuristics is set to TRUE.

The 'predict' function is then used to calculate the predictability of the SVM using the training dataset. The 'confusionMatrix' function calculates the variables related to classifier 'predict' function. The resulting variables of'confusionMatrix' are placed into csv files, which are read by the Tableau. The averages and standard deviation are generated and their respective graphs are generated. These graphs are discussed in detail in the

Evaluation section of dissertation.

5 Evaluation

For evaluation of each research question, the averages of accuracy, kappa, specificity, and sensitivity of the SVM are evaluated. In the third research question, the standard deviation will be considered, as it has a considerable impact over the other variables considered for evaluation. The first and second research questions have standard deviation of less than 0.02, thus having negligible effect over their conclusion.

$$Accuracy(\%) = \frac{(TruePositive + TrueNegative)}{(TruePositive + FalseNegative) + (FalsePositive + TrueNegative)}$$

The accuracy defines the capability of the model to predict the reaction of the news article. However, it cannot be solely relied upon to give an overall evaluation of the research questions. Other variables of the analysis are also considered to provide definitive evaluation.

$$KappaStatistic = 1 - \frac{(1 - Po)}{(1 - Pe)}$$

Po = Relative Observe Agreement among Rater (similar to Accuracy)
Pe = Hypothetical Probability of Chance Agreement

The Kappa statistic evaluates the model by analyzing its accuracy, while taking possibility of correct prediction by chance into account. It will help explaining, how the model performs better than the probability of chance. Evaluating the Kappa value is arbitrary, and largely dependent upon the analysis being conducted. The Kappa Statistic Table 3 displays the common bases of evaluating Kappa values.

Figure 3: Kappa Statistic

Kappa Statistic	Strength of Agreement
Less than 0.20	Poor Agreement
0.20 - 0.40	Fair Agreement
0.40 - 0.60	Moderate Agreement
0.60 - 0.80	Good Agreement
0.80 - 1.00	Very Good Agreement

$$Sensitivity = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$Specificity = \frac{TrueNegative}{FalsePositive + TrueNegative}$$

Sensitivity calculation indicate if the model is able to better predict the correct reaction from the actual correct reaction. Specificity calculation indicate if the model is able to better predict the incorrect reaction from the actual incorrect reaction. These two variables are needed to judge if the current model is suitable for use by news organizations and marketing teams.

5.1 Research Question 1: Can news articles predict which (Love, Haha, Wow, Sad, Angry) reaction gets the highest count from Facebook??

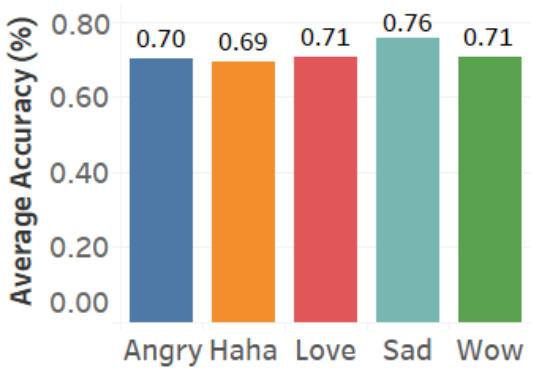

Figure 4: Average Accuracy

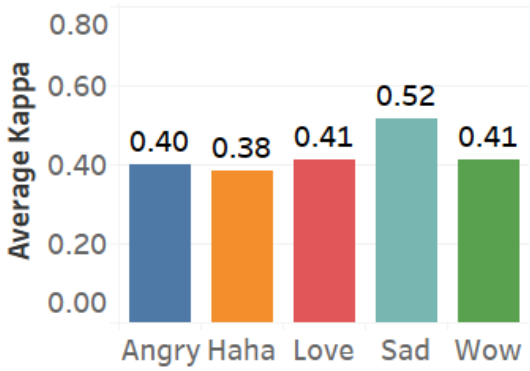

Figure 5: Average Kappa

The average accuracy displayed on Fig. 4 indicates that the model gives a good prediction. The average kappa values in Fig. 5, on the bases on Table 3, demonstrate fair to moderate agreement.

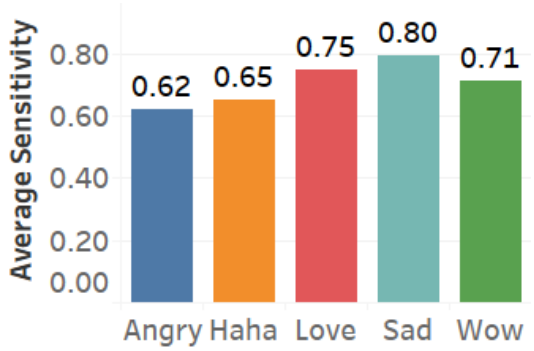

Figure 6: Average Sensitivity

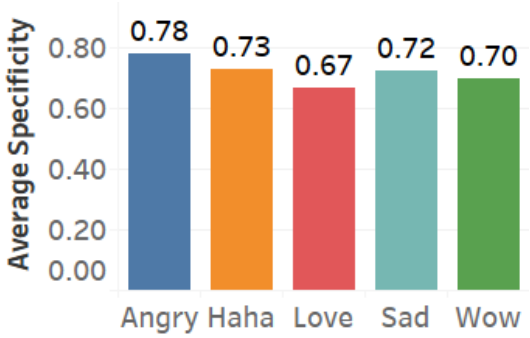

Figure 7: Average Specificity

The sensitivity in Fig.6 and specificity in Fig.7 show moderate to good level.

As the data analyzed is based on the social media, attaining accuracy above 0.90 is hard. However, it is possible to improve this model accuracy by incorporating other text analysis systems, in addition to LIWC. The increase in accuracy will subsequently improve the Kappa value. The additional parameters will also improve the sensitivity and specificity.

Based on the accuracy, kappa, specificity and sensitivity, it is possible to use the SVM to predict which reaction will get the highest count on Facebook, based on the analysis of news article.

5.2 Research Question 2: Can news articles predict which groups of (Love And Haha, Wow and Haha, Sad and Wow, etc) reactions gets the highest count from Facebook??

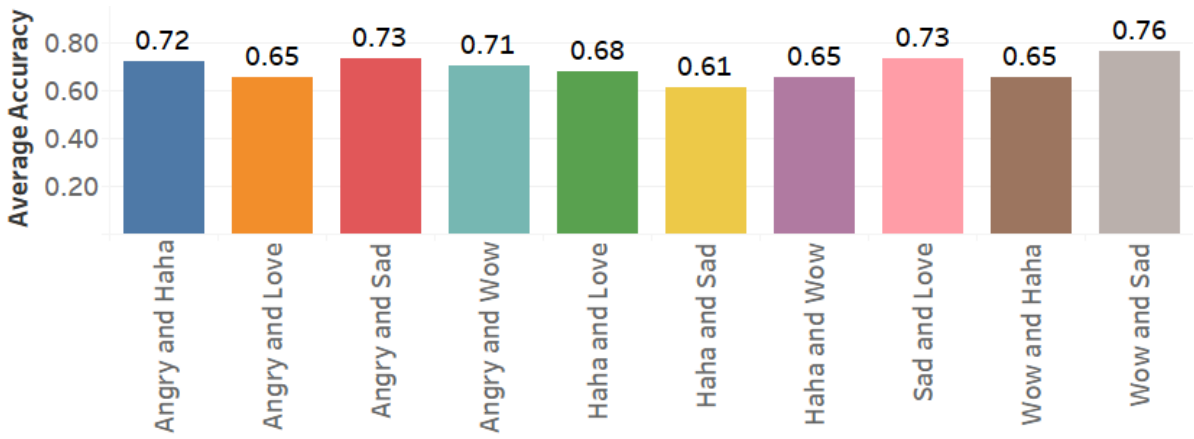

Figure 8: Average Accuracy

The accuracy shown in Fig.8 demonstrates that all 10 groups of reactions have moderate to good level of prediction. Compared to Fig.4, the overall accuracy is lower in group reactions than individual reactions. However, the variation between each group of reaction is higher than between reaction.

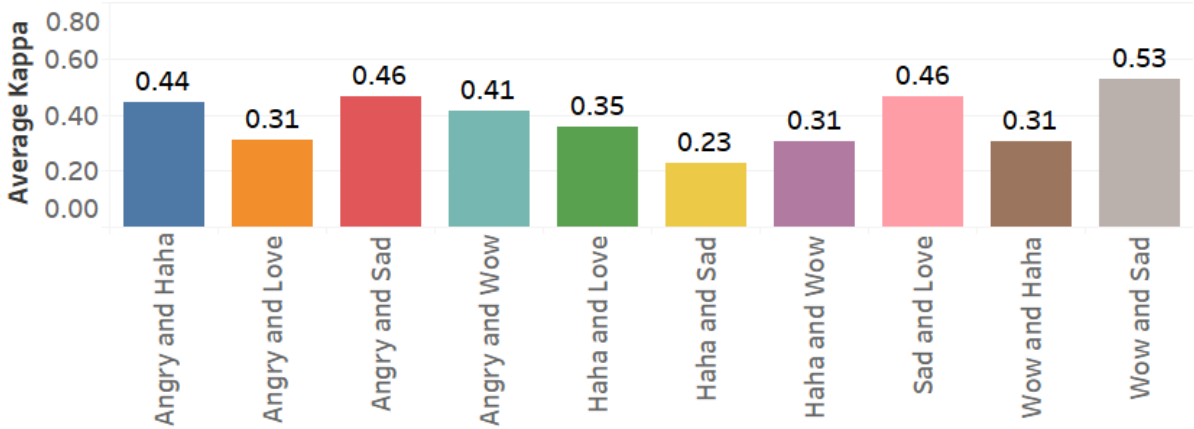

Figure 9: Average Kappa

The Kappa values in Fig.9 have varying numbers of Kappa statistics, ranging from low fair agreement to moderate agreement.

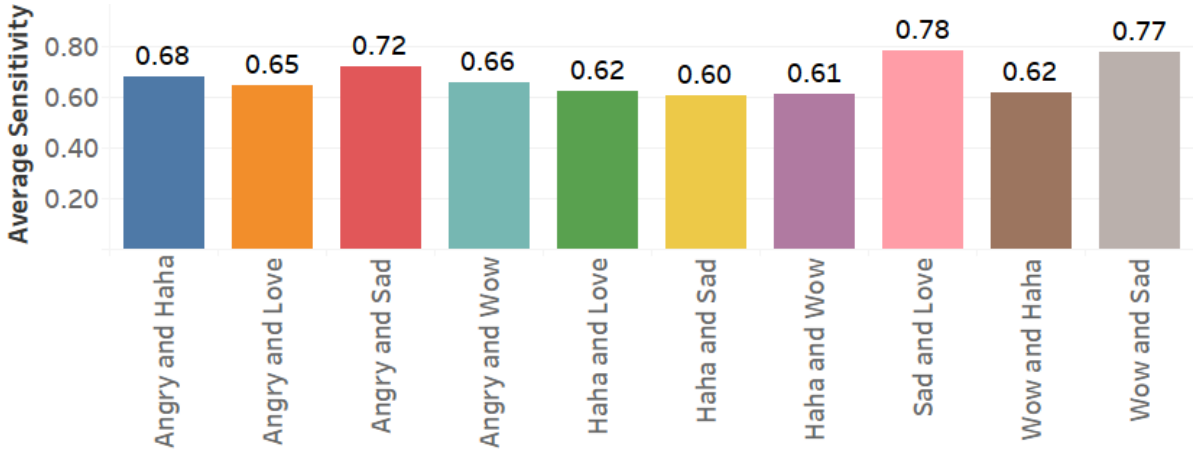

Figure 10: Average Sensitivity

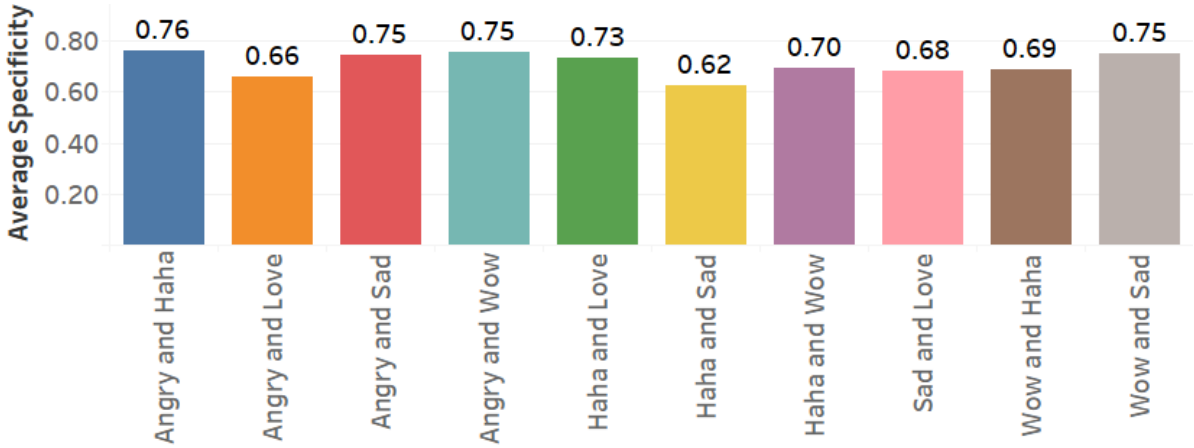

Figure 11: Average Specificity

The sensitivity in Fig.10 and specificity in Fig.11 show moderate to good level.

Based on the variables shown, it is possible to use the SVM to predict which groups of reactions will get the highest count on Facebook, based on the analysis of news article. However, in comparison with the first case study, it performs a little worse.

5.3 Research Question 3: Does the first research question differ for articles posted by different news organizations?

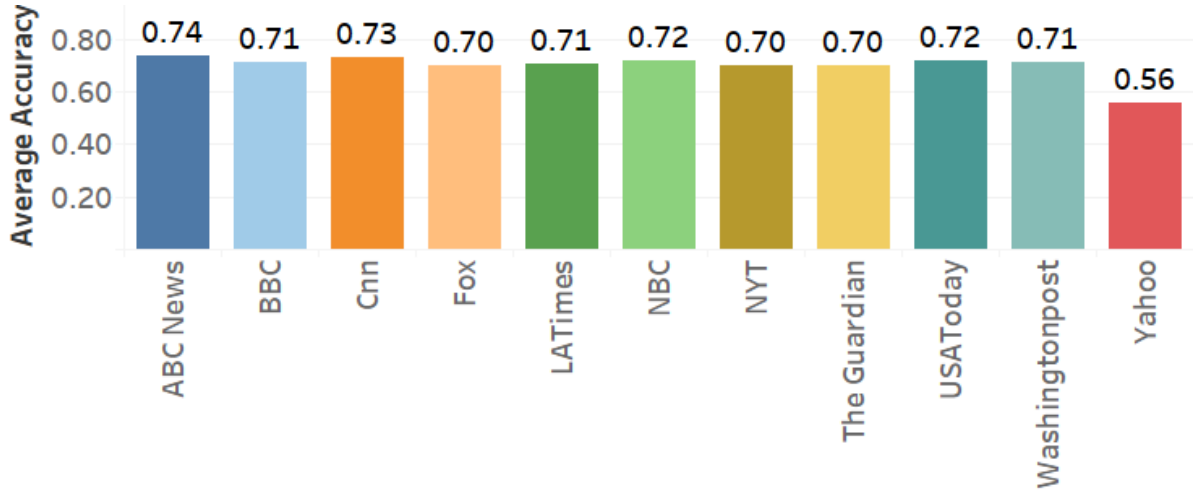

Figure 12: Average Accuracy

The accuracy shown in Fig.12 demonstrates that 10 groups of news organizations have good level of prediction, between 70% to 74%. Yahoo, however, has a low accuracy compared to others. Compared to Fig.4 and Fig.8, the overall accuracy is higher, excluding Yahoo.

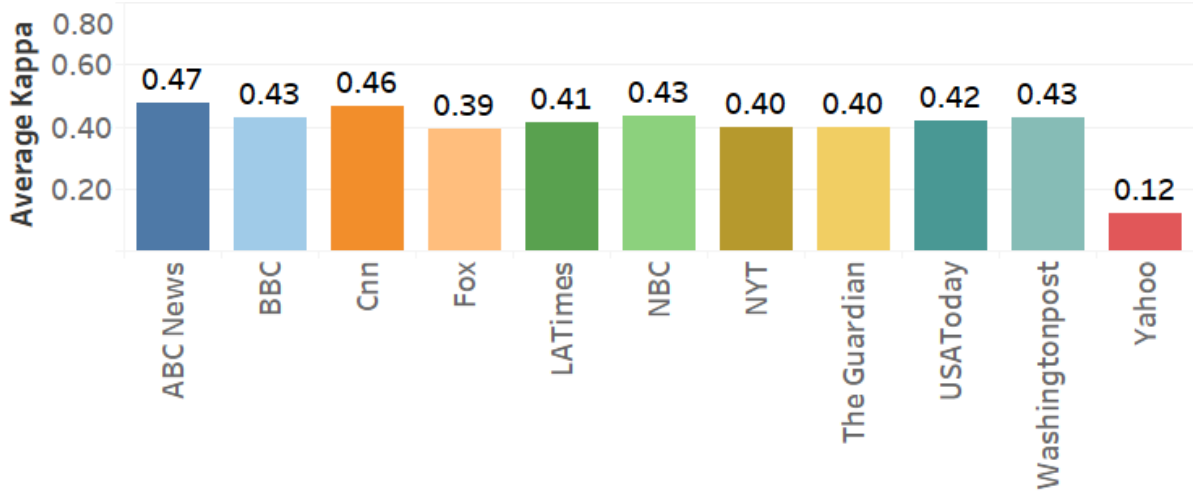

Figure 13: Average Kappa

The Kappa values in Fig.13 have varying numbers of Kappa statistics, ranging within low fair agreement. Yahoo Kappa has a poor agreement.

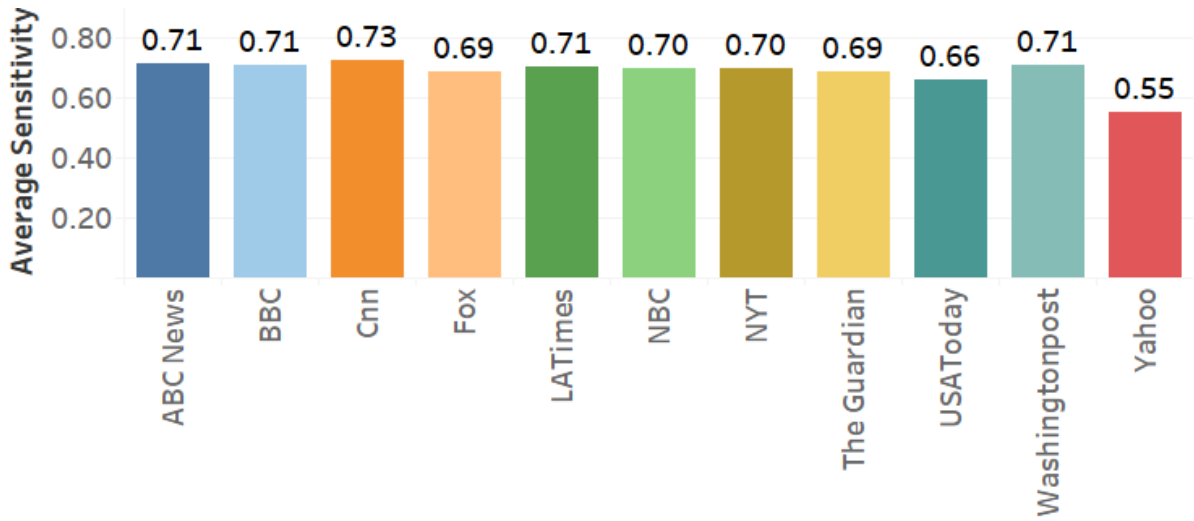

Figure 14: Average Sensitivity

The Sensitivity in Fig.14 high-moderate and low-good level, while Yahoo has a low sensitivity of 0.55.

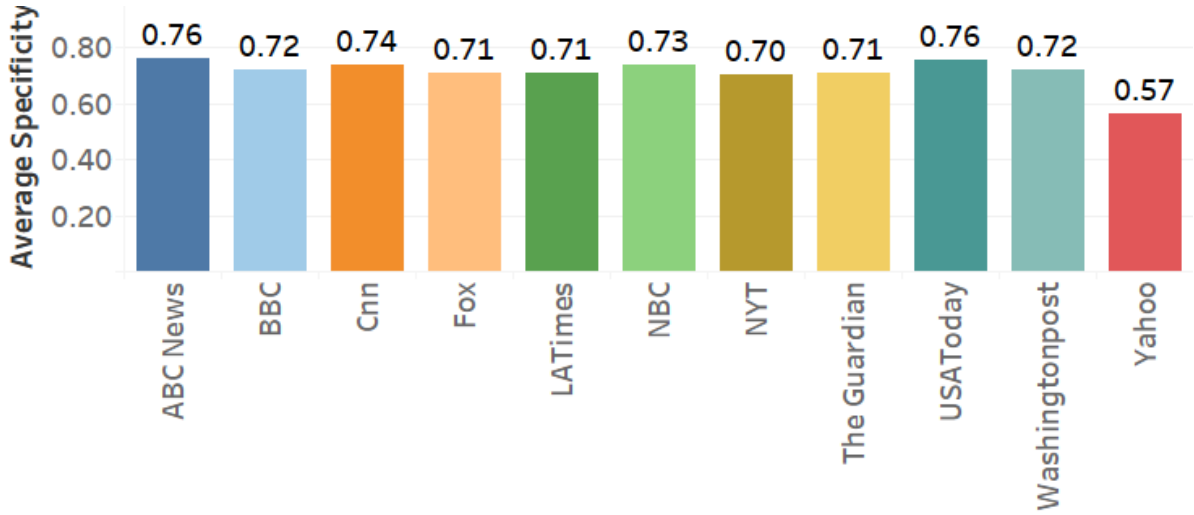

Figure 15: Average Specificity

The specificity in Fig.15 has good level, while Yahoo has low specificity of 0.57.

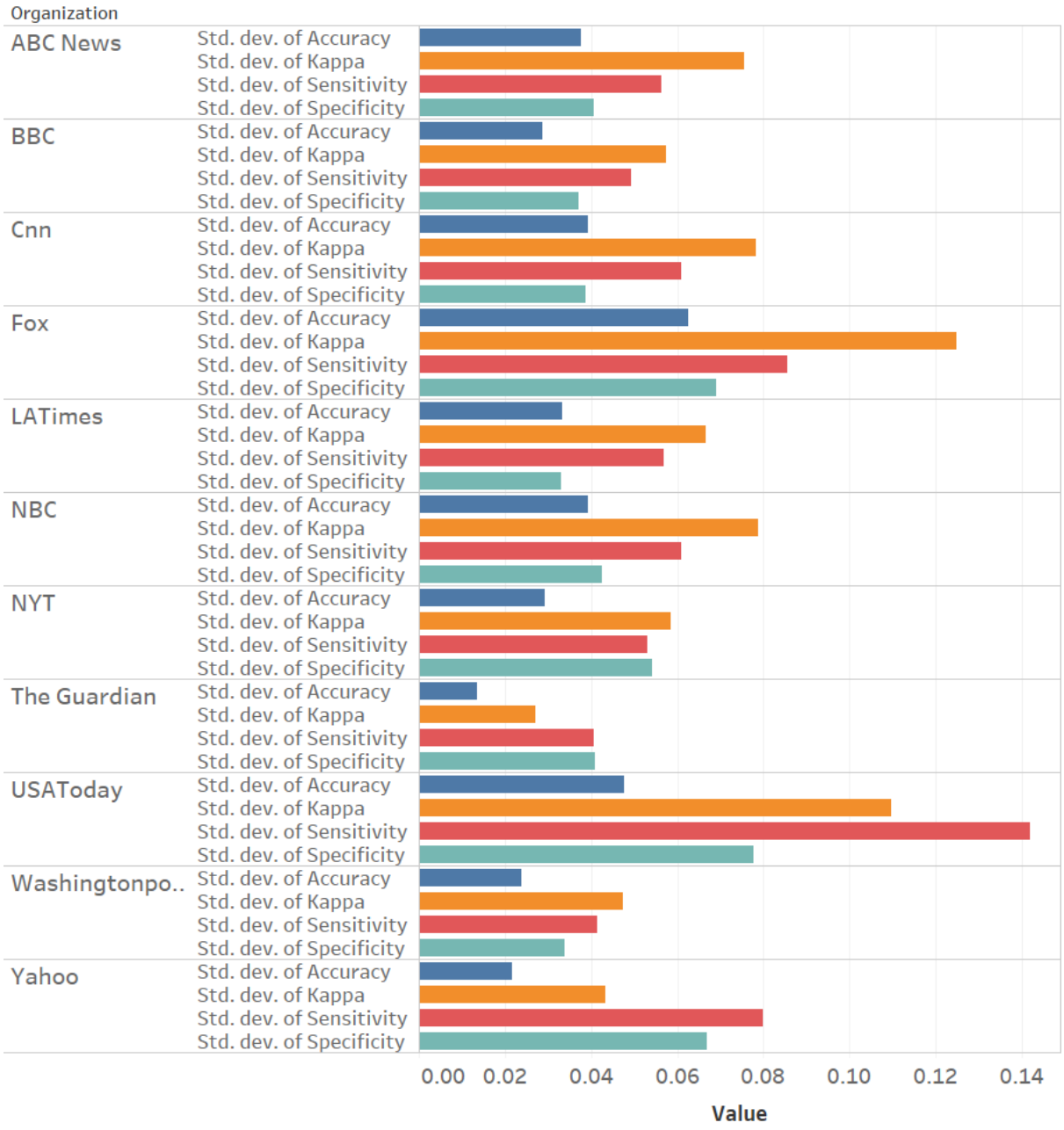

Figure 16: Standard Deviation

The standard deviation is prominent in the third research question. The cause for this is that all reactions data has been averaged, after being grouped by news organizations. The first two research question show large variations between reaction(s) in terms kappa, sensitivity and specificity. Thus concluding that for better analysis, each news organization reactions should be analyzed individually to get a better evaluation.

Based on the variables shown, there is no definitive conclusion. It needs further evaluation by analyzing articles of each individual news organization.

5.4 Discussion

Based on the analysis of this dissertation, it is possible to identify which reaction(s) will get the highest count by the text analysis of news articles. The first research question is comparable with the study conducted by Guzman et al. (2016). They reported the highest accuracy of 0.7254, for love reaction. They provided a colored confusion matrix graph, which demonstrates love, haha and sad achieving moderate to high accuracy, but angry and wow achieving very low accuracy. The highest accuracy achieved in first research question is 0.76, for sad reaction, and demonstrated high accuracy across all 5 reactions. In addition, the current model includes a larger dataset of 16 months data, while Guzman et al. (2016) used 4 months of data. The first research question concludes that the current model performs better in terms of accuracy than used by Guzman et al. (2016). However, Guzman et al. (2016) did not report any other variables regarding their model, like kappa, sensitivity and specificity.

The second research question is inspired by the recommendation of Badache and Boughanem (2017). They recommended that, in addition to analyzing individual reactions, group of reactions should also be analyzed. The second research question also performed as good as the first research question, but not as well. This is due to significantly large variations in the kappa values(0.3) in Fig. 9 of second research question, when compared to first research question kappa values(0.14) in Fig. 5.

The third research question shows how the reactions perform on the basis of news organization. Other than Yahoo, all other news organizations perform well, as per the variables. However, the standard deviation shows that some news organizations have large variations between some of the variables. For example, Fox and USA Today kappa value has a large standard deviation, around 0.13 and 0.11. The largest standard deviation is the USA Today sensitivity, which is more than 0.14.

When the dissertation was started, the news organizations were selected that had the largest number of followers on Facebook[12], as of August 2017. By checking the political leaning of the news organizations covered, through All Sides[13], Media Bias Fact Check[14] and Pew Research Centre[15] websites, the large standard deviation of Fox and USA Today become clearer. All news organizations covered are either left or left center wing of political spectrum. Fox is the only right wing news, and USA Today is the only center wing news. The large kappa standard deviation of Fox and USA Today is because the model is trained primarily by the left wing news articles bias. Fig 17 shows that articles with highest sad counts is similar to left wing news bias regarding sad reaction. In Fox case, the other reactions have very little in common with left wing bias reactions. In USA Today case, this is only true for love reaction. It could also explain why there is a large standard deviation in sensitivity of USA Today, as displayed in Fig 18.

[12] http://fanpagelist.com/category/news/world/

[13] https://www.allsides.com

[14] https://mediabiasfactcheck.com

[15] http://www.journalism.org/interactives/media-polarization/

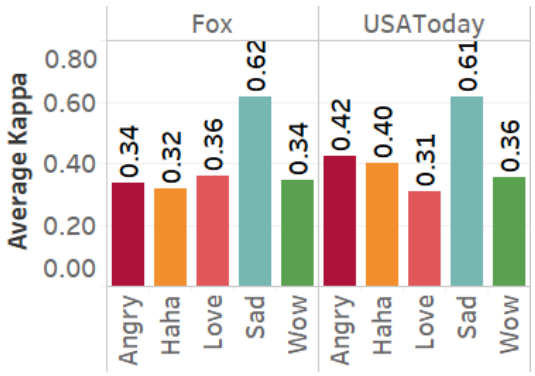

Figure 17: Average Kappa of Fox and USA Today

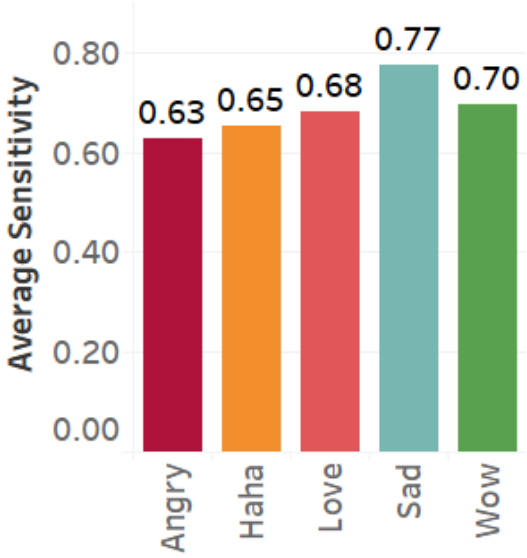

Figure 18: Sensitivity of USA Today

The third research question also indicates that the Yahoo is the outlier of all the news organizations used for the dissertation. In addition to USA Today and Fox data, it would explain why there is large variation in kappa, sensitivity and specificity in the first two research questions. Removing these three news organizations data from the entire dataset can improve the performance of the model.

Excluding the Yahoo news, the sensitivity and specificity readings in research questions indicate the current model has moderate credibility, ranging from 0.62 to 0.80 of both sensitivity and specificity. However, the Kappa values indicates fair to moderate agreement. This indicate that the model can be used by news organizations to see how the users would react, but only in supporting role. If the kappa values had been higher than 0.70, then the model can be used as primary source for predicting the reactions. The marketing team can utilize the model to create targeted ads that would take advantage of the user reactions. However, since the model does not have excellent credibility, it would be a risk to use the model for controversial new articles. This is especially true for marketing team, as the ads they post with the news article can receive backlash. Examples of this include how Cheerios tried to honour Prince death with a tribute that was of poor taste and it backfired, harming the company profile [16].

6 Conclusion and Future Work

Social Media has increased in size and prominence greatly in last couple of years. More people are joining social media to connect with one another. Dwindling News Organizations and Advertisement are embracing social media to reach wider audience than they did through traditional methods. Understanding how their audience, and their competitor audience, would react towards news articles is important. This dissertation is the initial step towards understanding audience reactions in Facebook.

[16] https://www.netbase.com/blog/dos-donts-celebrity-deaths-trendjacking/

Based on the conclusion of the research questions, it is possible to predict which reaction will get the highest count, by analyzing the text of news article published online. However, there are many ways this can be further improved. Using a different classifier than SVM can improve performance. Using Deep learning methods, like deep neural networks, would perform better than SVM. It is also possible to test a scenario, in second research question, where larger groups of reactions can be predicted. Ranking within the group reaction can also give a better understanding of the reactions. For example, angry and sad are grouped, but the angry count is greater than sad, and sad is greater than the other reactions. Using a different text analysis software, like Open Information Extraction, might give a different insight into predicting reactions, and improve overall performance of the model. It is possible to study the differences of audience reactions to news organizations with a different political leaning.

Overall, there are many different branches of Data Analysis that can expand the current research.

References

Avery, J. J., Deighton, J. A., Gupta, S. and John, L. K. (2017). 'likes' lead to nothing-and other hard-learned lessons of social media marketing, *HBS Working Knowledge* . Accessed on 9 Dec. 2017.
URL: *https://hbswk.hbs.edu/item/don-t-express-sympathy-with-a-cheerio-and-other-hard-learned-lessons-of-social-media-marketing*

Badache, I. and Boughanem, M. (2017). Emotional social signals for search ranking, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, ACM, New York, NY, USA, pp. 1053–1056.

Ding, C., Cheng, H. K., Duan, Y. and Jin, Y. (2017). The power of the like button: The impact of social media on box office, *Decision Support Systems* **94**: pp. 77–84.

Eranti, V. and Lonkila, M. (2015). The social significance of the facebook like button, *First Monday* **20**(6): pp 53.

Greenberg, J. (2016). Advertisers dont like facebooks reactions. they love them, *Wired* . Accessed on 9 Dec. 2017.
URL: *https://www.wired.com/2016/02/advertisers-feel-facebooks-new-reactions-%F0%9F%98%8D/*

Guzman, R., Ochoa-Luna, J., Cruz-Quispe, L. and Vera-Cervantes, E. (2016). Predicting reactions to blog headlines, *CEUR Workshop Proceedings* **1743**: 43–47. Accessed on 9 Dec. 2017.
URL: *https://www.researchgate.net/publication/312117234*$_{Predicting_{R}eactions - to_{B}log_{H}eadlines}$

Hwong, Y.-L., Oliver, C., Kranendonk, M. V., Sammut, C. and Seroussi, Y. (2017). What makes you tick? the psychology of social media engagement in space science communication, *Computers in Human Behavior* **68**(Supplement C): pp. 480 – 492. Accessed on 9 Dec. 2017.
URL: *http://www.sciencedirect.com/science/article/pii/S0747563216308172*

Jang, S. M. and Oh, Y. W. (2016). Getting attention online in election coverage: Audience selectivity in the 2012 us presidential election., *New Media Society* **18**(10): pp. 2271 – 2286.

Matloff, N. (2011). *The art of R programming: A tour of statistical software design*, No Starch Press.

Orellana-Rodriguez, C., Nejdl, W., Diaz-Aviles, E. and Altingovde, I. S. (2014). Learning to rank for joy, *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, ACM, New York, NY, USA, pp. 569–570.

Pelletier, M. J. and Blakeney Horky, A. (2015). Exploring the facebook like: a product and service perspective, *Journal of Research in Interactive Marketing* **9**(4): pp. 337–354.

Peterson, T. (2015). How will brands feel about facebook 'reactions'?, *Advertising Age* **86**(19): pp. 6. Accessed on 9 Dec. 2017.
URL: *http://search.ebscohost.com/login.aspx?direct=trueAuthType=ip,cookie,shibdb=edsgaoAN=eds livescope=sitecustid=ncirlib*

Pilon, A. (2016). Small business trends: How to use new facebook reactions for emotional content ideas, *Newstex Entrepreneurship Blogs* . Accessed on 9 Dec. 2017.
URL: *https://smallbiztrends.com/2016/03/new-facebook-reactions.html*

Pool, C. and Nissim, M. (2016). Distant supervision for emotion detection using facebook reactions, *arXiv preprint arXiv:1611.02988* .

Ringelhan, S., Wollersheim, J. and Welpe, I. M. (2015). I like, i cite? do facebook likes predict the impact of scientific work?., *PLoS ONE* **10**(8): pp. 1 – 21.

Robinson, M. D., Boyd, R. L., Fetterman, A. K. and Persich, M. R. (2017). The mind versus the body in political (and nonpolitical) discourse: Linguistic evidence for an ideological signature in u.s. politics., *Journal of Language Social Psychology* **36**(4): pp. 438.

Stinson, L. (2016). Facebook reactions, the totally redesigned like button, is here, *Wired* . Accessed on 9 Dec. 2017.
URL: *https://www.wired.com/2016/02/facebook-reactions-totally-redesigned-like-button/*

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods, *Journal of language and social psychology* **29**(1): pp. 24–54.

Tian, Y., Galery, T., Dulcinati, G., Molimpakis, E. and Sun, C. (2017). Facebook sentiment: Reactions and emojis, *SocialNLP 2017* p. 11.

Turnbull, S. and Jenkins, S. (2016). Why facebook reactions are good news for evaluating social media campaigns., *Journal of Direct, Data Digital Marketing Practice* **17**(3): pp. 156.

Udanor, C., Aneke, S. and Ogbuokiri, B. O. (2016). Determining social media impact on the politics of developing countries using social network analytics, *Program* **50**(4): pp. 481–507.

Winter, S., Brckner, C. and Krmer, N. C. (2015). They came, they liked, they commented: Social influence on facebook news channels., *CyberPsychology, Behavior Social Networking* **18**(8): pp. 431 – 436.