A Neutral Analysis: Using Sentiment in Deducing the Origin of Facebook Confessions Forums Data

Adrian Naranjo

Abstract

This paper attempts to answer the question: “can a classifier be built to categorize the country of origin of particular posts on Facebook Confessions Forum using only sentiment?” Only the English speaking countries of United States, Canada, and United Kingdom are considered. In conclusion, the accuracy of models using only sentiment variables ranges between 84% - 90%. Should other characteristics beyond sentiment be included, such as number of words and characters per word, a nearly perfect categorization can be achieved. However, given the perfection of this latter classifier, there is concern that information from the response variables is somehow contained or transferred to the sentiment features used.

Introduction

Over the years, Facebook has collected people’s thoughts on a range of topics. However, this information is mostly public and is presented by users knowing that their posts will be read by others. On the other hand, Facebook Confession Boards (FCB) contain information about or surrounding personal facts that “someone would prefer to not be shared under their true identity” (2). Much research has been done on public Facebook posts, but not as much on these anonymous pages which offer insights into personal matters that users would not be willing to share openly (otherwise they could simply share it on Facebook).

Our goal is to determine if a classifier can be built to categorize the country of origin of FCB posts using only sentiment. It is our belief that the sentiments between posts of different countries are sufficient to classify between countries. The classification of country is important as it can provide insight to differences between communities. The purpose isn’t necessarily to understand where the posts are coming from, but to eventually provide insight into what these differences tell us about those communities (e.g. is one community more positive or negative?) While country may be too high level of an aggregation to give meaningful understanding into a community, it is a good technical starting point from which more granular classifications (e.g. regions, states, or universities) can proceed.

I will begin by describing the raw data and the modifications made to it. Then I’ll proceed to discuss our classifier, it’s anomalies, attempts to reduce the concern surrounding those anomalies, and its results. Finally, I’ll discuss implications and future work.

Data

I have elected to use data from Facebook Confession Boards (FCB) which I extracted using a website provided to us by Arindam Paul (1). The raw data contains 8 columns (Table 1) with approximately 150,000 observations spanning from 2012 thru 2017 for the countries of Canada, United States, and United Kingdom. The data is heaving weighted towards North America with Canada and the United States accounting for 97% of the data (approximately 140,000 observations).

Table 1. Original Raw Features.

Country	Datetime	Number of Comments	Number of Likes	Post	Region	State	University

Table 2. New Features Extracted From the Raw Data.

Day	Characters per Word	I-Word Count	I-Word Percent of Total Words	I-Word to U-Word Ratio	Month	Number of Characters
Number of Words	Region/State	Stop Word Count	Stop Word Percent of Total Words	You-Word Count	You-Word Percent of Total Words	Year

In addition to the variables contained by the raw data, 11 other variables were added (Table 2). The I-Word and You-Word variables are based on lists of words constructed to identify the use of posts referencing the singular self (I, I've, I'd, I'll, I'm) and singular other (you, you've, you'd, you'll, you're). Our belief in introducing these lists is that the use of speaking about oneself versus speaking of others, particularly as a singular pronoun, provides more information about an individual. Furthermore, several built-in R sentiment dictionaries were imported and used to constructed 13 more variables (not shown in Table 2). Two of these variables measured the average negative or positive rating of the words in a given post on a numerical scale (Afinn and Bing dictionaries) while the other 11 represented a specific sentiment from the NRC dictionary (anger, anticipation, disgust, fear, joy, negative, neutral, positive, sadness, surprise, trust).

It should be emphasized that the predictive models below are trained and tested on the meta characteristics of each individual post (i.e. variables from tables 1 and 2), not the posts themselves, nor the aggregate summaries of each country. In other words, I trained on 80% of the observations, though the training observations were in fact pure meta characteristics of the raw data (with the exception of the Number of Likes and Number of Comments). Also, before creating the sentiments for each post, preprocessing on the posts did take place and included:

- Removal of punctuation, stopwords, numbers, and whitespace
- Stemming
- Removal of terms which were 99% spare or greater

Models & Results

To predict the country of each observation I focused on Tree Based methods (in particular boosted trees, random forest, and bagging). It is not that other methods may not be as successful, but due to the initial high accuracy of these models and the added benefit of having built-in feature importance, other models were not needed. For Boosted Trees 10-fold cross-validation was used with a learning rate of .3 on a maximum of 10,000 trees. An early stopping was also implemented, so that if the error did not decrease within 10 iterations, no more trees would be grown. For Random Forest/Bagging out-of-bag error was used on 500 trees.

Because the accuracy of our models was high, I implemented each of our tree models five times (Table 3). Each run used a different set of features, with each sequential run having less features than the previous. The first run used all possible variables, while the last was strictly based on sentiments from the NRC library and also excluded the Neutral sentiment. All models excluded the variables of Datetime (including Year, Month, Day, and Hour), Post, University, Region, and State as some of these characteristics implicitly contain information about the origin country.

Table 3. Model Results.

Run	Number of Variables	Which Variables were used?	Boosted Accuracy	Random Forest Accuracy	Bagging Accuracy
1	24	All	100%	100%	100%
2	20	Removed Counts	99.7%	99.8%	99.8%
3	14	Only Sentiments	99.4%	99.3%	99.4%
4	11	Only NRC Sentiments	93.9%	91.1%	93.9%
5	10	NRC Sentiments except Neutral	84.7%	84.4%	84.7%

Note: to view the exact variables used during each run, please refer to the Appendix below.

More specifically, the initial run included all variables and led to an accuracy of 100% (Run 1). Given this highly suspicious outcome, the importance of the variables were reviewed and displayed several non-sentiments as being most important (Figure 1). It was decided to remove these variables from the model and re-run (Run 2). This still led to highly suspicious accuracies near 100%. Looking at the importance of variables for these models, all non-sentiment variables were removed (e.g. variables associated with stopwords and I/You-words) which led to accuracies slightly above 99% (Run 3). The variables were then trimmed to only those produced by the NRC dictionary (Run 4). With only these variables, the accuracies dropped to around approximately 91% - 94%. Finally, the Neutral sentiment was removed and lead to accuracies of approximately 84% - 85% (Run 5).

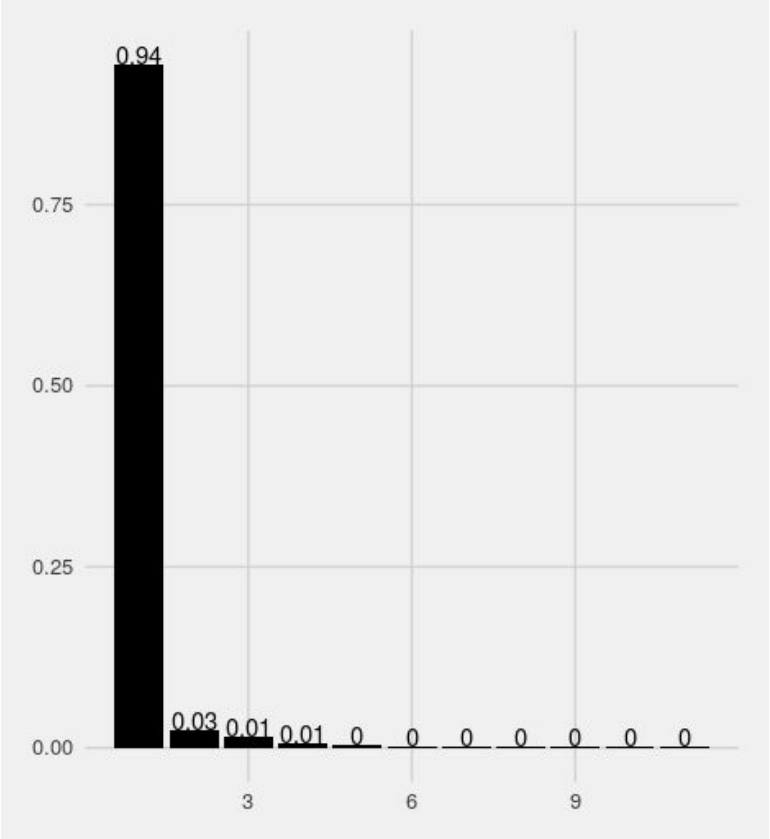
Table 4. Top 3 Features of each Run by Importance.

Importance	**Run 1**	**Run 2**	**Run 3**	**Run 4**	**Run 5**
First	Number of Likes	Percentage of I-Words	Neutral	Neutral	Anticipation
Second	Characters per Word	Neutral	Avg Afinn	Anticipation	Positive
Third	Number of Words	Average Afinn	I-Word to You-Word Ratio	Positive	Trust

Note: to view the the full importance plot for each run, please refer to the Appendix below.

To be sure of the importance of the Neutral sentiment, I conducted a Principal Components Analysis (PCA) of Run 4. The PCA indicates the first two principal components account for 97% of the variance, with the first component alone accounting for 94%. Within the first component, Neutral receives a loading of .98, which is over 10x larger, on an absolute scale, than the next loading (please see “Summary of Principal Components Analysis of Run 4” in the Appendix). This indicates the Neutral sentiment is indeed a driver of the classification using the NRC sentiment variables only. Beyond the Neutral sentiment, no particular sentiment appears to dominate, though anticipation is slightly more useful than either positive or trust, both of which edge out negative.

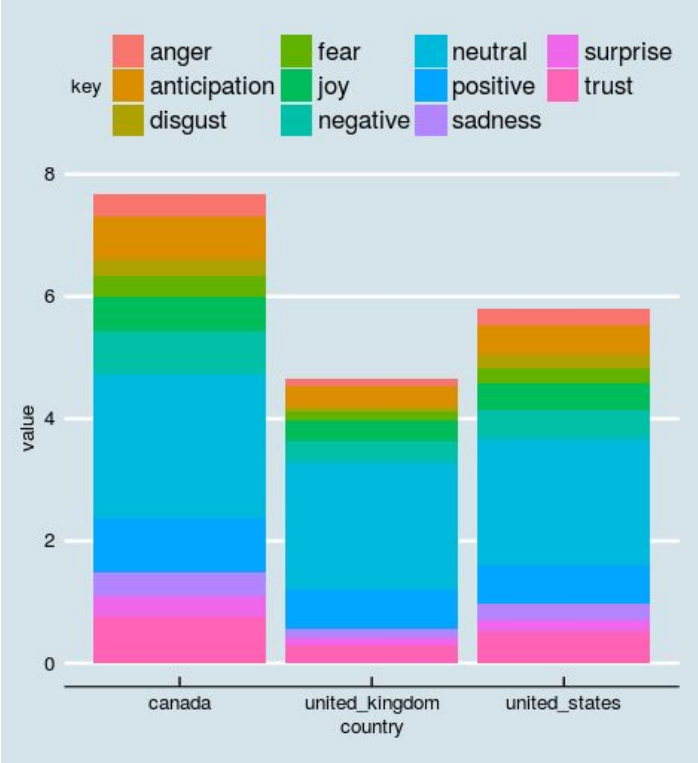
Figure 1. Variance from Principal Components Analysis



Discussion

While it is concerning that our full model (Run 1) and nearly full models (Runs 2 & 3) achieve nearly perfect accuracies, models using only NRC sentiments (Runs 4 & 5) performed reasonably well with 84%+ accuracy and appear much less likely to contain information from the response in the predictors. So, what can explain such good accuracy using only sentiment? One particularly interesting aspect lies in the descriptive sentiment analysis of each of the countries (Figure 2. However, for more detailed plots, please view “Sentiments by Country...” in the Appendix). That is, there is a clear hierarchy with regards to the expression of emotion between Canada, the United States, and the United Kingdom, even after standardizing counts to address the uneven representation of North America within the data. For any given sentiment, Canada clearly expresses more emotion than the United States, which clearly expresses more emotion than than the United Kingdom. While this may not generalize to each individual post (which can be attested to by the fact that the accuracy is not perfect), this trend holds well at a country level and could a reasonable explanation for our models better than average accuracy.

Figure 2. Sentiment Comparison by Country.



Future Work

Are sentiments from the NRC library enough to predict other countries or even universities within the data? Future work should focus on precisely that. Predicting Region and State is not a far state from predicting country and can potentially shed considerable light on whether accuracy can be maintained. With regards to other countries, while data was available for countries outside of the three which I selected for the above analysis, I elected to focus on only English speaking countries as I was not sure of how to pre-process posts in other languages.

It is also worth mentioning that future work must be conducted to understand if information from the response variable is indeed contained in some of the predictors which did not make it to models 4 and 5. After reviewing our analysis I do not believe I've made a mistake (however, this is certainly not out of the question), yet accuracies of 100% are rare and highly unlikely.

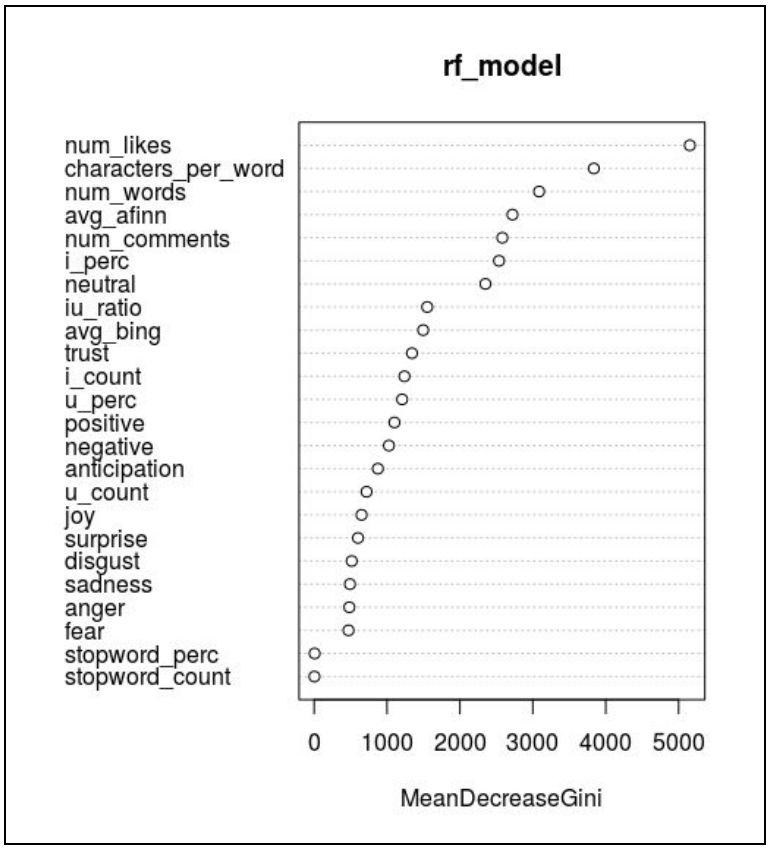
References

(1) http://sentic.net/wisdom2016paul.pdf
(2) https://en.wikipedia.org/wiki/Confessions_page

Appendix

Variables used in Run 1: Average characters per Word, Number of Comments, Number of I-Words, Number of You-Words, Number of Likes, Number of Stopwords, Percentage of Stopwords, Number of Words, Percentage of I-Words, Percentage of You-Words, Ratio of I-Words to You-Words, NRC Sentiments (Anger, Anticipation, Disgust, Fear, Joy, Negative, Neutral, Positive, Sadness, Surprise, Trust), Average Afinn Sentiment, Average Bing Sentiment

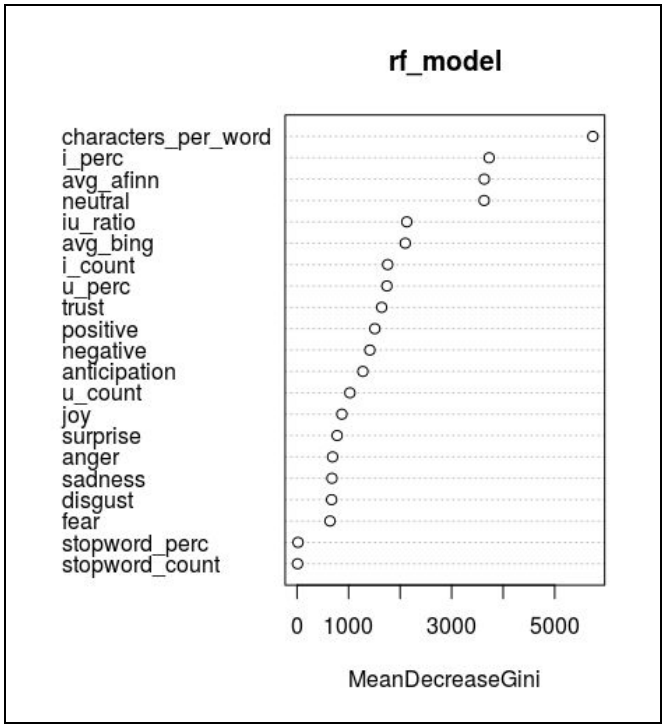
Variable Importance Plot for Run 1.



Variables used in Run 2: Average characters per Word, Number of I-Words, Number of You-Words, Number of Stopwords, Percentage of Stopwords, Percentage of I-Words,

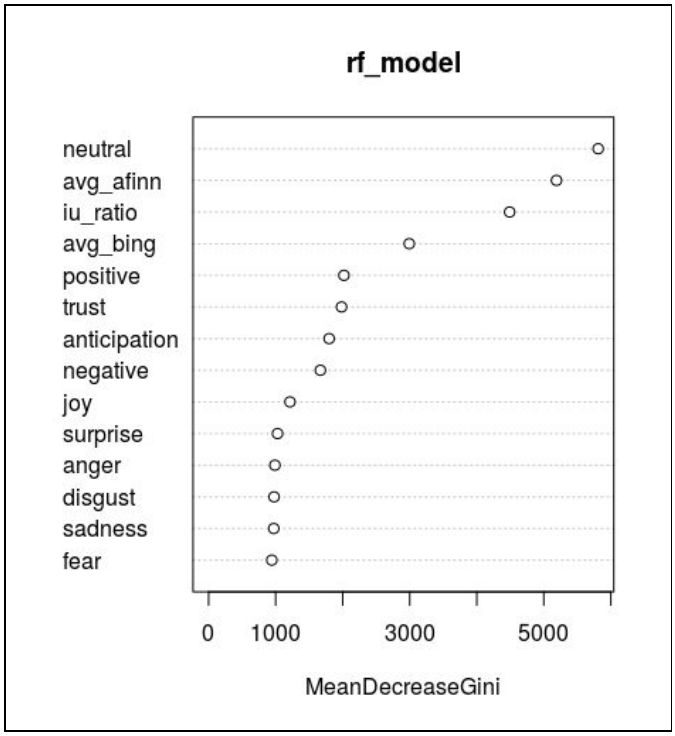
Percentage of You-Words, Ratio of I-Words to You-Words, NRC Sentiments (Anger, Anticipation, Disgust, Fear, Joy, Negative, Neutral, Positive, Sadness, Surprise, Trust), Average Afinn Sentiment, Average Bing Sentiment

Variable Importance Plot for Run 2.



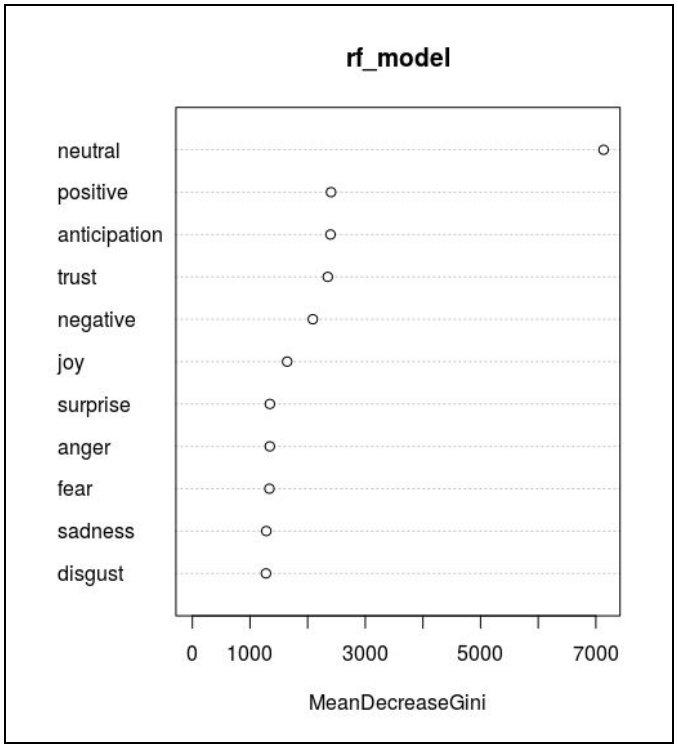
Variables used in Run 3: Ratio of I-Words to You-Words, NRC Sentiments (Anger, Anticipation, Disgust, Fear, Joy, Negative, Neutral, Positive, Sadness, Surprise, Trust), Average Afinn Sentiment, Average Bing Sentiment

Variable Importance Plot for Run 3.



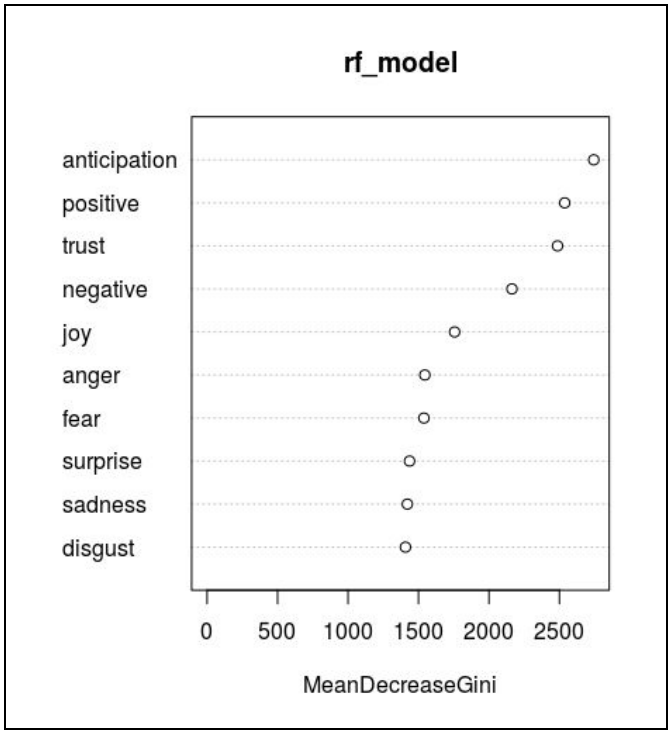
Variables used in Run 4: Only NRC Sentiments (Anger, Anticipation, Disgust, Fear, Joy, Negative, Neutral, Positive, Sadness, Surprise, Trust)

Variable Importance Plot for Run 4.



Variables used in Run 5: All NRC Sentiments except Neutral (Anger, Anticipation, Disgust, Fear, Joy, Negative, Positive, Sadness, Surprise, Trust)

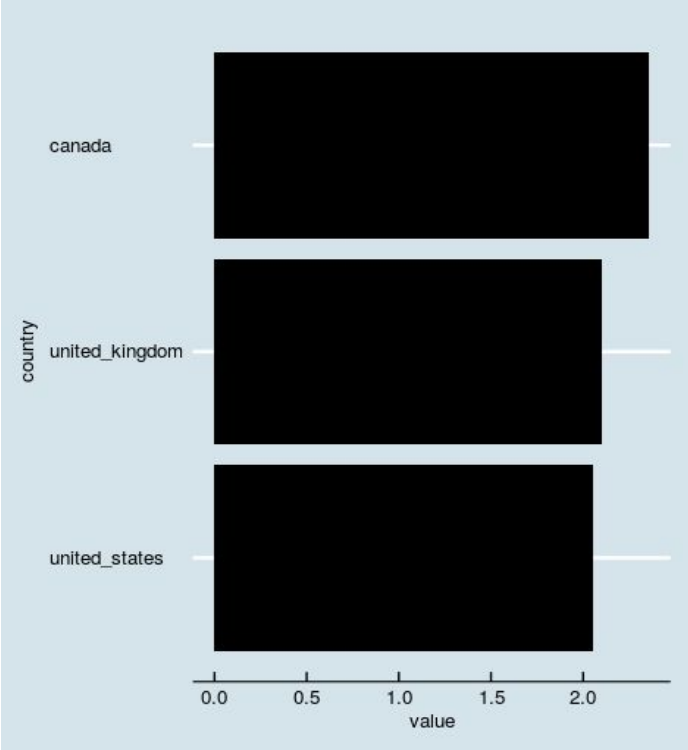
Variable Importance Plot for Run 5.



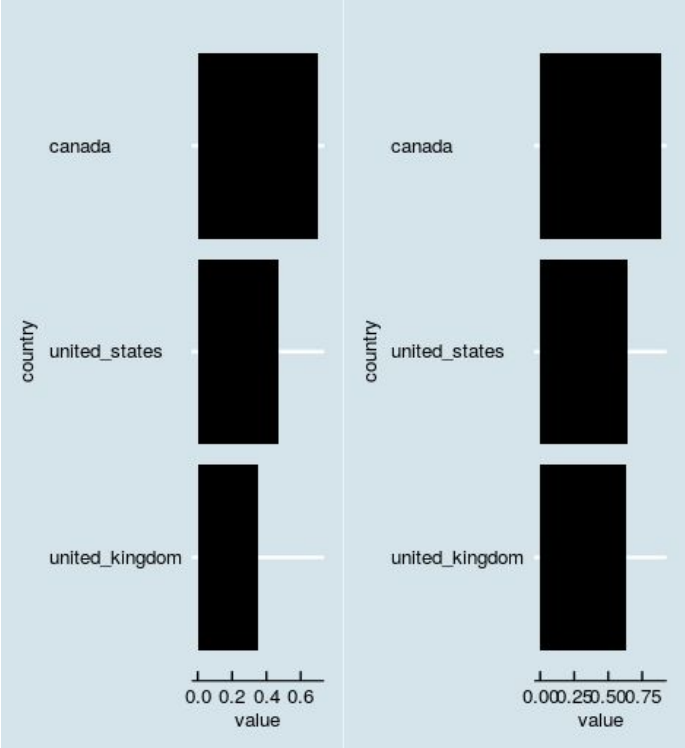
Summary of Principal Components Analysis of Run 4.

Sentiment	First Principal Component	Second Principal Component
Neutral	.982	.184
Positive	.0969	-.0868
Anticipation	.0846	-.102
Trust	.0739	-.120
Negative	.0731	-.149
Joy	.0564	-.199
Fear	.0416	-.215
Sadness	.0401	-.364
Anger	.0340	-.400
Surprise	.0326	-.479
Disgust	.0205	-.551

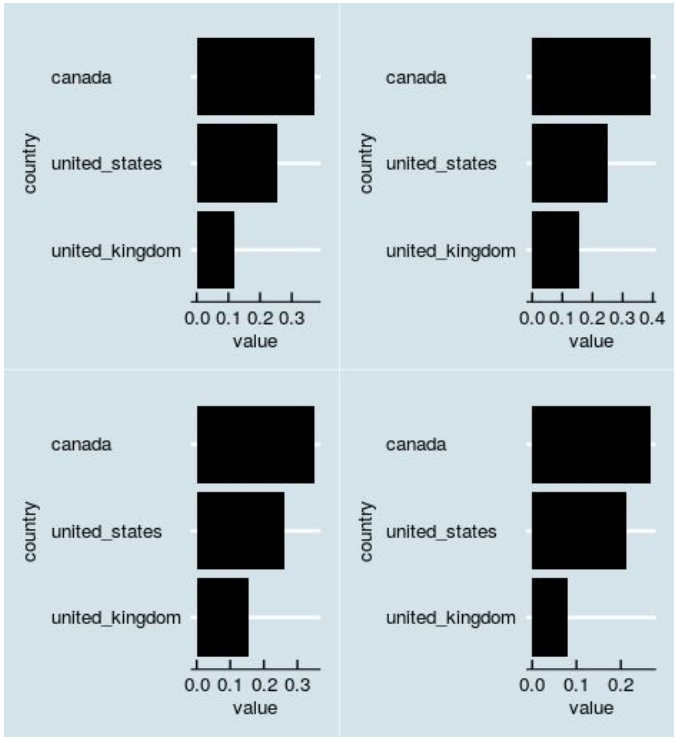
Plot of Neutral Sentiment by Country



Plots of Positive (left) and Negative (right) Sentiments by Country



Plots of Anger (top left), Sadness (top right), Fear (bottom left), and Disgust (bottom right) Sentiments by Country



Plots of Anticipation (top left), Joy (top right), Surprise (bottom left), and Trust (bottom right) Sentiments by Country