**Quora: Is it Sincere?**

***Introduction***

******

Quora was formed in June 2009 by two former Facebook employees who had a vision for the public sharing of knowledge. In what was perhaps a nod to the popular nineties aphorism about the internet existing as a “vast information superhighway,” Adam D’Angelo and Charlie Cheever had a vision for curating the vast amounts of valuable knowledge stored in the minds of millions of internet users for the benefit of everyone. The website became available to the public in June 2010 and was received with wide acclaim. With the goal of having users “gather around a question” and presenting singular, apolitical versions of each one, Quora sought to bring together people of various backgrounds and paradigms in solidarity to share knowledge and nothing more. While their goals were admirable and noble, anyone who has spent even marginal time on the internet understands an open forum invites opinions, misinformation, and insincerity.

The internet is unique in that it has made possible real-time communication behind an insular shield of anonymity. While the viciousness afforded by the invisibility the internet provides is not new or unique, it has thrived in the seemingly unbounded space between users. Social cues provide a method of evaluation, prohibition, and retaliation in face-to-face communication as strongholds against repugnant speech and behavior. These defenses do not exist in the online world, leaving a void where unsatisfactory behavior can flourish. According to Gizmodo’s online article “The Birth of the Internet Troll,” there has been an evolution in internet troublemakers from causing annoyance based on beliefs, to “simply believing in causing annoyance.” This is the type of user Quora wishes to avoid: one who annoys because annoyance is its own reward.

To that end, Quora has initiated a contest via the online community Kaggle, which describes itself as an “AirBnB for Data Scientists.” The contest seeks to find the most accurately tuned algorithm for identifying insincere questions – questions phrased to express an opinion, express sarcasm, or where no real answer is being sought. Quora defines the contest goal as detecting content that is toxic in order to improve online conversations. The dataset provided by Quora is heavily skewed, so a machine learning accuracy score is not a realistic measure of its success. Instead, the contest entries will be evaluated on F1 score between the predicted and observed classifications. While Quora has already been working to flag insincere content using both machine learning and manual methods, the company wishes to employ more scalable ways of identifying the kind of content they wish to mitigate.

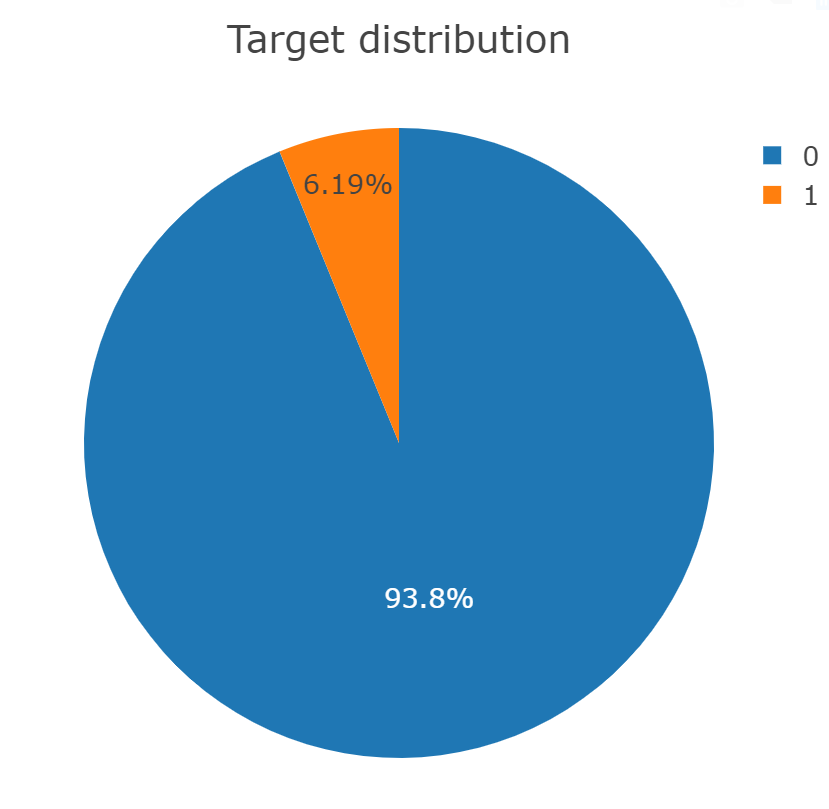
***The Data***

Quora provided training and testing datasets to contest participants. The training data is skewed profoundly to the side of sincere questions, with 1,225,312 of the examples classified as sincere, and 80,810 of them labeled insincere. The term ‘insincere’ is used as a general characterization in this report to describe what Quora refers to as toxic, misleading, divisive, and insincere content interchangeably. The data are labeled with integers: 0 for sincere questions and 1 for insincere.

The data was preprocessed by several methods, each of which is listed below.

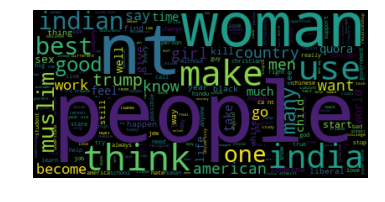
* All words changed to lowercase
* Regex used to remove punctuation and other unwanted characters
* Tokenized, stop words removed
* Non English words removed
* Porter stemmer applied
* Part of speech tagging was used to lemmatize the data
* Tf-idf and Boolean vectors were created from the cleaned data
* Two vectors using Bigrams and Trigrams were also created

As previously mentioned, the data is heavily skewed to the side of sincere questions, as seen in the graph below. With more than 93% of the dataset being made up of sincere questions, decisions needed to be made about how to analyze the data. Understanding that the skewed data could have an impact on results, Quora’s method of analyzing the results of the contest entries was to judge them according to the F-score, which will be explained later in the report. For this report, two methods were used: one with a more balanced split between the categories, and another with a smaller, but equally balanced, set of random samples.



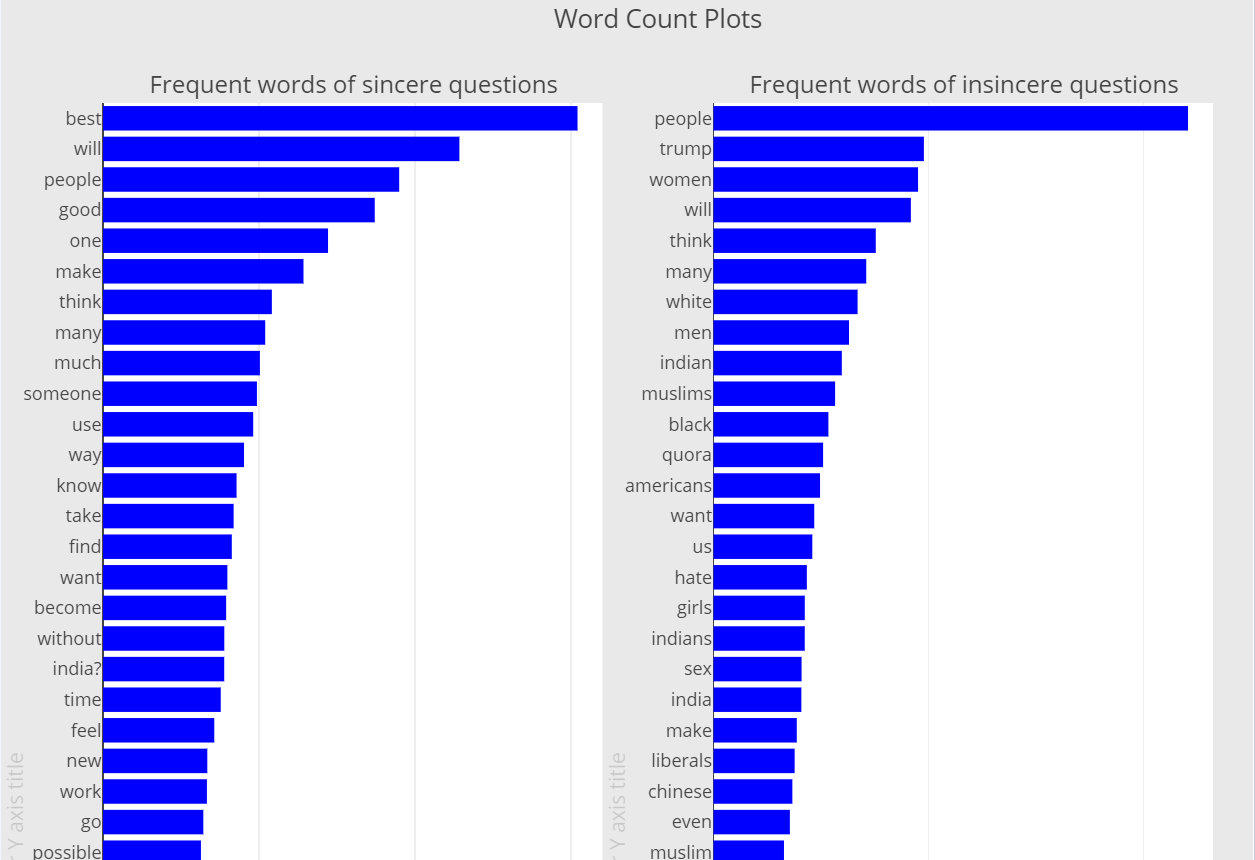
***Exploratory Analysis of the Data***

After the initial sampling, which consisted of 6,000 sincere and 4,000 insincere questions, a word cloud was created to get a sense of the most frequent words.

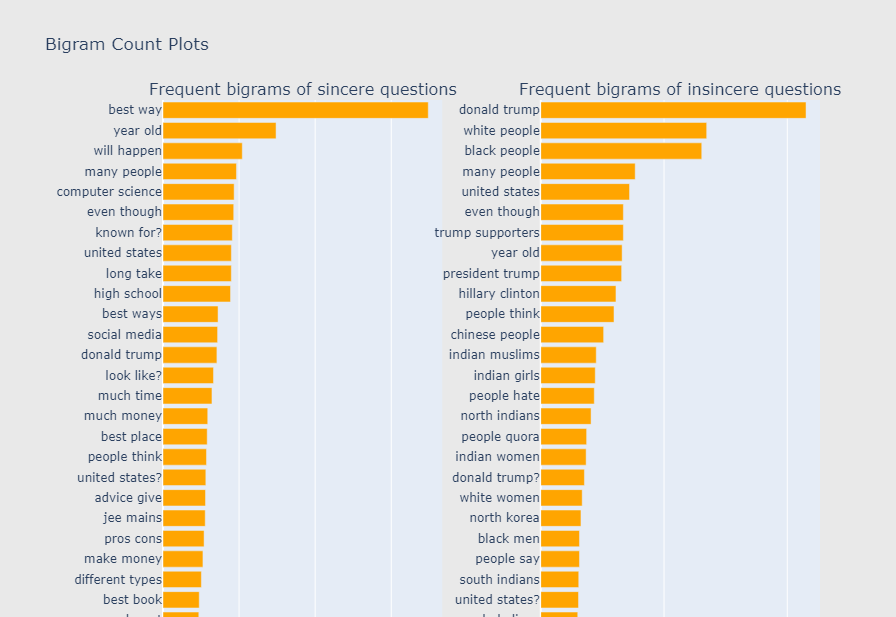
****

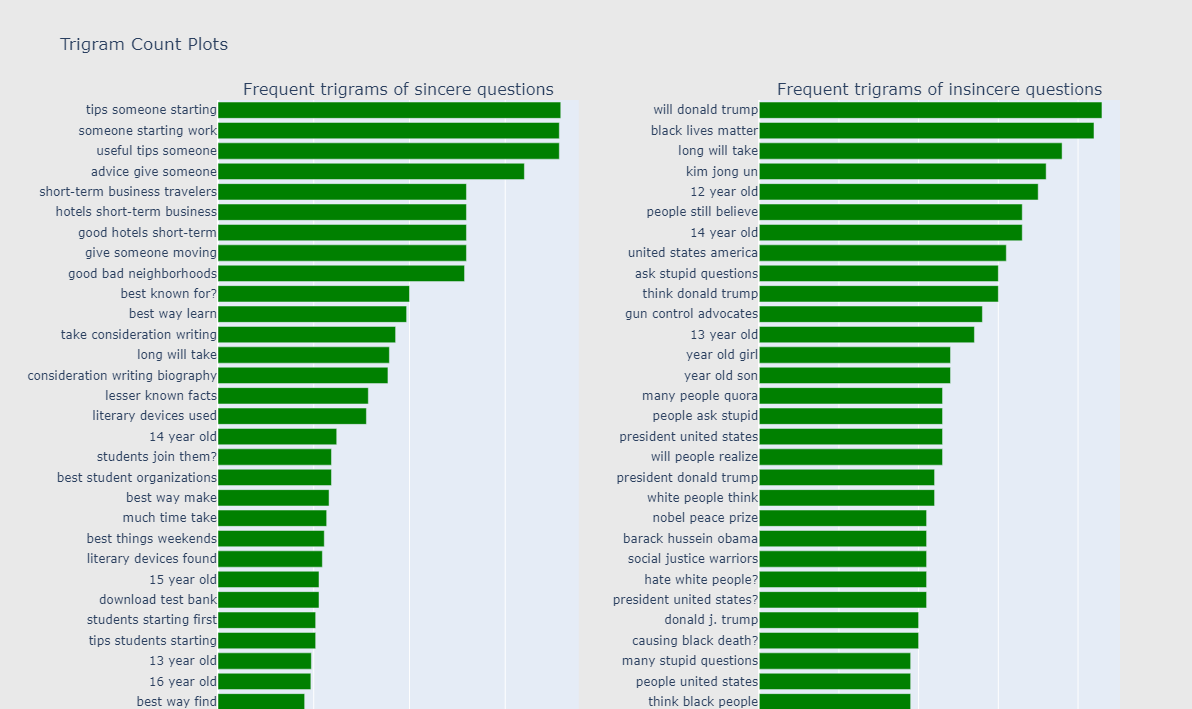
“People” is a very popular word in the data. This is evident in both sincere and insincere questions, which include many samples such as “why do American people…” or “why do people from America…” and similarly structured questions. Other similar questions were structured specifically to confuse or entertain, as in the example below.

Exploration also included looking at word frequencies as unigrams (single words), bigrams (pairs), and trigrams (groupings of 3) according to the classification. Unigrams included the kinds of insincere words expected: people, Trump, Muslims, hate to name a few. The sincere category was much more general and not as easy to classify. 



Bigrams showed similar results where insincere questions were concerned. Once again, topics such as politics, race, and nationality dominate the conversation. With the sincere portion of the data, it remains difficult to solidify groupings of any kind. The questions seem to be across a large number of topics, which is to be expected.



Trigram frequencies began to give a clearer picture of not only the differences between sincere and insincere questions, but also their purpose. Once again, topics such as race, nationality, and politics seem to make up much of the insincere category, suggesting a desire to express an opinion. The sincere questions still seem to be much more general, but have begun to show signs of users seeking knowledge about business travel, advice, and tips. This suggests the purpose of these questions is seeking knowledge, which is the purpose Quora wants to advance.

***Prediction Models***

**Naive Bayes**

The Bayes Theorem assumes that every variable of a feature is conditionally independent. The theorem calculates probability P(c|x), where c is the class of possible outcomes and x is the given instance classified. This method is frequently used for natural language classifications. For this project, the multinomial (mNB) and Bernoulli methods were employed. Multinomial quite simply means more than two classifiers were available. While that was not the case in this experiment, the use of mNB is appropriate when data is unbalanced as word frequencies and orders can affect the output of binomial models. In the Bernoulli model, an indicator is generated for each for each term of the vocabulary, either a 1 indicating the presence of the term, or a 0 indicating its absence. mNB and Bernoulli are equally complex models.

**Multinomial Naive Bayes**

To train the mNB model, a tf-idf (term frequency-inverse document frequency) vector was created from the preprocessed data. The default values for this vector were used, including the use of unigrams. The vector was split into two separate data frames, one containing the insincere questions and the other with the sincere questions. Each of these data frames were then split into a test and train set with the test data being 30%. This resulted in four data frames, two test data frames and two train data frames. A concatenation was done to combine both test sets into one data frame and likewise for the train sets. This was done to ensure proper representation of both insincere, as well as sincere questions in test and train data. The target column (prediction of either sincere or insincere) was then removed from the data frames and stored in two new data frames: train labels and test labels. The train labels data frame was used along with the train data frame to train the multinomial naive bayes classifier. Predictions were then done on the test data set without the target column. A confusion matrix was created to compare the prediction results to the true target values in the test labels data frame. This confusion matrix can be seen below.

The mNB process was then repeated twice more using bigrams and trigrams. For these models, the count vectorizer was used. The two parameters that were set for the count vectorizer were the range of n\_grams (as either (1,2) or (1,3) for bigrams and trigrams respectively) and the min\_df = 5. Setting the min\_df to 5 meant that any terms with a document frequency lower than 5 were ignored. As a result, these infrequent words were not used when training the models and predicting the target. The confusion matrices can be found below.

**Bernoulli**

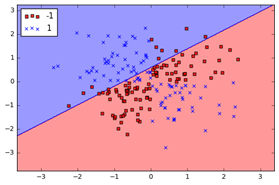
The Bernoulli model was trained in a similar fashion to the mNB model. However, the one difference between the models was that Bernoulli used a boolean vector. This was accomplished by using CountVectorizer with the binary parameter set to true. A confusion matrix for the Bernoulli predictions can be seen below.

**SVC**

A Support Vector Classifier (SVC), is a predictive modeling technique in which a formula is algorithmically created to “separate” out the varying classifications. This separation can be geometrically described by a line, curve, or hyperplane. In the case of a two dimensional vectorization (i.e. one in which there are 2 components), this can be easily visualized by a line optimally separating two groups of points, as shown below (the blue points can represent one categorical value, and the red, another):

The mode of separation can also be varied, so as to increase the accuracy of the classifier. Using different kernel values can alter the way in which the algorithm looks to separate the points.

In the context of this analysis, the resultant separation will be in the form of a “hyperplane”, since each body of text is vectorized from its word frequency, and weighted usage (for tfidf). This means it cannot be visualized, as in the above example, since each vector carries with it too many components to map in a geometric space. The algorithm used to achieve maximum separation involves calculating the “centroid” for all of the data points – that is, the point that is located at the average of all of the components. From this point, the hyperplane is pivoted and adjusted about each successive data point, such that the region is split in such a manner that all sincere questions are located in one region, and all insincere in another, localized, region.



Varying the Kernel and Gamma value:

To vary the approach within this classification method, different kernels can be used. This kernel determines the type of separation between regions sought after by the algorithm. A linear kernel, for example, will try to find a linear path or plane to separate out the points. A polynomial kernel identifies a plane with polynomial function level-sets, and a radial basis function (rbf) kernel uses the squared Euclidian distance to weave the plane through the space between the points.

The gamma value is another feature of Support Vector Classification, and it specifically relates to the non-linear options mentioned above. The higher the gamma value, the more the algorithm attempts to exactly fit the training data set. A gamma value that is too high can result in overfitting to the training dataset, which will mean decreased accuracy for the model overall.

In this circumstance, over-fitting is a legitimate concern, as it is heavily skewed. Since the number of insincere questions, by proportion, was quite small – attempting to weave out a separation between sincere and insincere questions with a high gamma value could easily render a result that only works for the training, and does achieve optimal results within the test set.

**Logistic Regression**

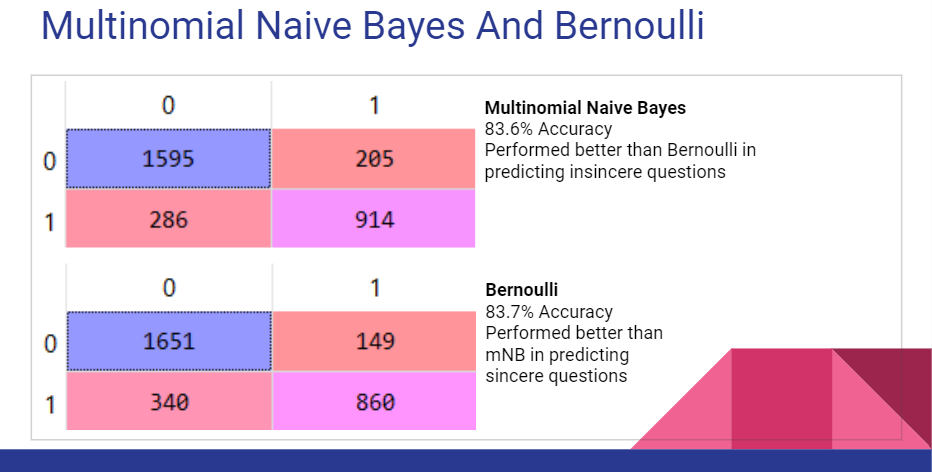
To further vary the approach in binary classification, a logistic regression is appropriate. Similar to the SVC method, the logistic regression attempts to minimize the “loss” in wrongly classified points. The difference is the means by which the algorithm accomplishes this. While SVC pivots, or “hinges” the hyperplane to separate the points, Logistic regression attempts to effectively fit a curve to minimize the sum of squared error (SSE) between the predicted class, and actual class. The SVC output is effectively Boolean, as the tested data point is either on one side of the hyperplane, or another – and this determines its classification. The logistic model, on the other hand, gives a *probability* that the point, or in this case, question, is either sincere or insincere. If the model shows that it is more probable that the question is insincere based on its coefficients, then it predicts it to be so.

**Random Forest**

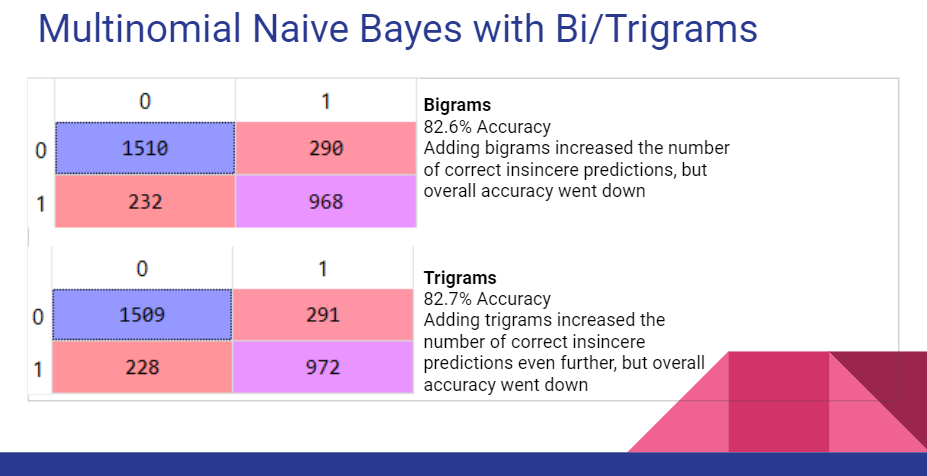
A random forest is a collection of randomly generated decision trees. When creating a single decision tree, the tree is created using all of the data and variables and then pruned to generate the best tree. In decision tree analysis, each tree is created using a randomly selected group of data and variables. The output from these trees is combined to create the final output of the model. Each tree in the forest gets a “vote” to classify the data point. The label with the highest percentage of votes will classify the data point.

***Model Results***

Since the data for the mNB and Bernoulli models contained 60% sincere questions and 40% insincere questions, the baseline accuracy that needed to be beat was 60%. In short, if the model were to guess sincere for every test case, it would be correct 60% of the time. The accuracy for the unigram mNB model and Bernoulli model were calculated by dividing the total number of correct predictions by the total number of predictions. The mNB model resulted in an accuracy of 83.6%, while the accuracy for Bernoulli was 83.7%. Although mNB had a slightly lower accuracy, it did better in predicting the correct number of insincere questions with 914 correct predictions as opposed to 860 correct predictions for Bernoulli. While the accuracies are important, Quora’s goal is to find a model that will help them identify insincere questions. Accordingly, judging models both by accuracy and their ability to predict insincerity is important to the goal of this project.

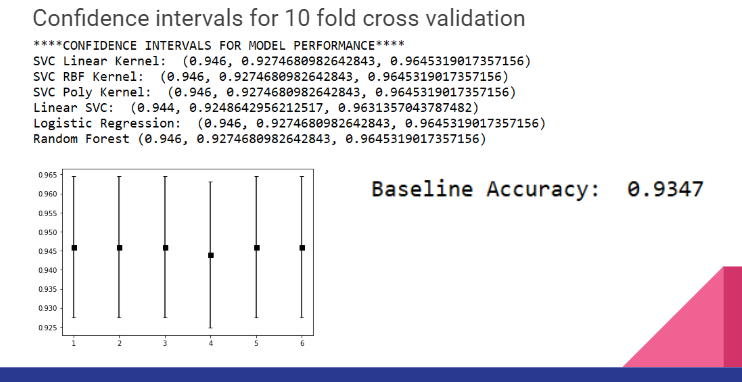


The mNB models trained with bigrams and trigrams also outperformed the baseline accuracy of 60% with accuracies of 82.6% and 82.7%, respectively. Although the accuracy was lower than the previous models, the number of correct insincere predictions increased. The correct insincere predictions for the bigram model came to 968, the second highest of all four models. The model with the highest number of correct insincere predictions was the trigram model, with 972 correct predictions. The trigram model only increased slightly over the bigram model, but the increase over the unigram mNB and Bernoulli models was significant.



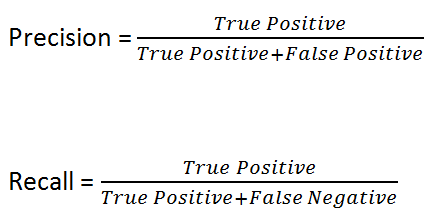
***Confidence Intervals for Cross Validation***

In order to establish the sustainability of the models, it is important to replicate the process of training and testing across multiple subsets. One such method of doing so is known as cross validation, and involves breaking up the dataset into “folds”, and having different combinations of these folds constitute the training and testing split. In this circumstance, 10 such folds were used, and a 60/40 training/testing ratio was used. For each of the 10 iterations, 6 folds were assigned to the training set, and 4 to the testing. Each model described above was then trained, and its accuracy tested. Each resulting accuracy was appended to a list that was subjected to analysis after completion of all iterations. In order to assess which model was most effective, it is necessary to look at not only the model accuracy, but also the margin of error in its accuracy. In this circumstance, all models – SVC with variation in kernel and gamma, logistic regression, and random forest – returned very similar accuracies and margins of error. The comparison was made to a baseline “majority rule” accuracy, that was computed every iteration. The baseline accuracy was used since the dataset exhibited major skew, and simply randomly guessing “sincere” on every question, would render an accuracy of 93.7% on its own. The plot below demonstrates that all models beat the baseline accuracy by a slight margin, and offer a nice contrast to the improved accuracy generated from the MNB and Bernoulli NB models that were executed on an assembled data set with a 60/40 split in sincere/insincere questions. Models did reach accuracies of as high as 96%, but at times, also dropped below the baseline.



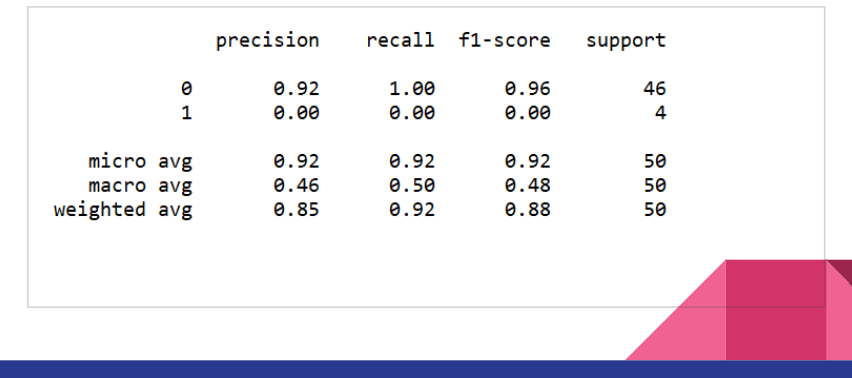
***F Scores, Precision, and Recall***

Quora sought F1 scores as the preferred results for this analysis. F1 is a measure of the accuracy of a test. It takes into account both the precision and recall of the test, measures which are visualized below. It is calculated by dividing the product of precision and recall by the sum of the same, and multiplying the dividend by 2. The result is a measure of the balance between the test’s precision and recall, which is a way to measure (and mitigate the consequences of) false positives and negatives.



The F1 result for this analysis was 0.96 for the positive classifier, which is not surprising since the training data was 96% positive. The aggregated micro and weighted averages produced promising results with good support.

The results also rendered a 100% recall rate on smaller testing subsets. This measured the ability of the model to extract all insincere questions. This meant that the only error in the model was its generation of false positives. The sample output below is for one such run of the model on a testing subset of size 100,000.

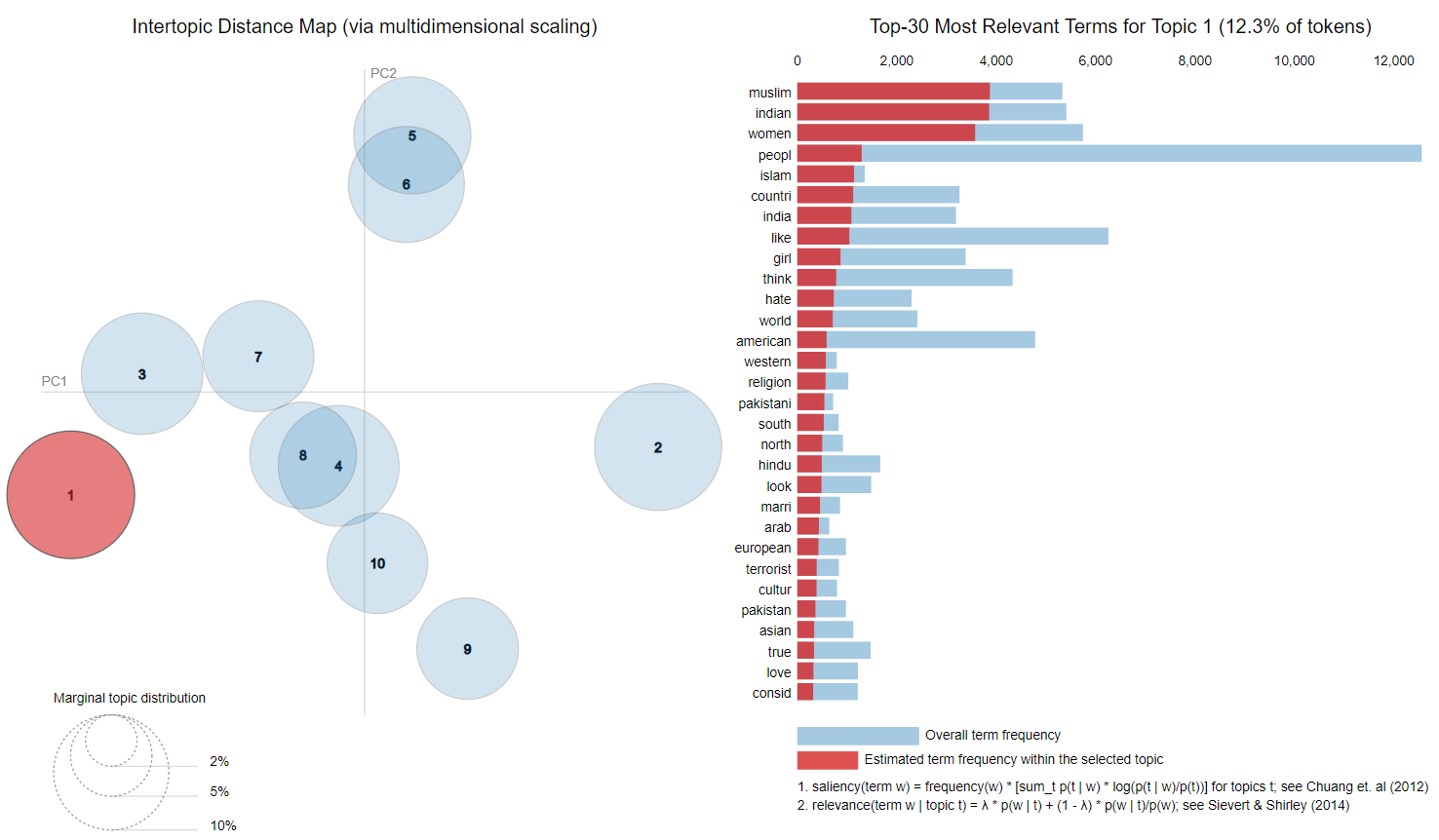


***Topic Modeling***

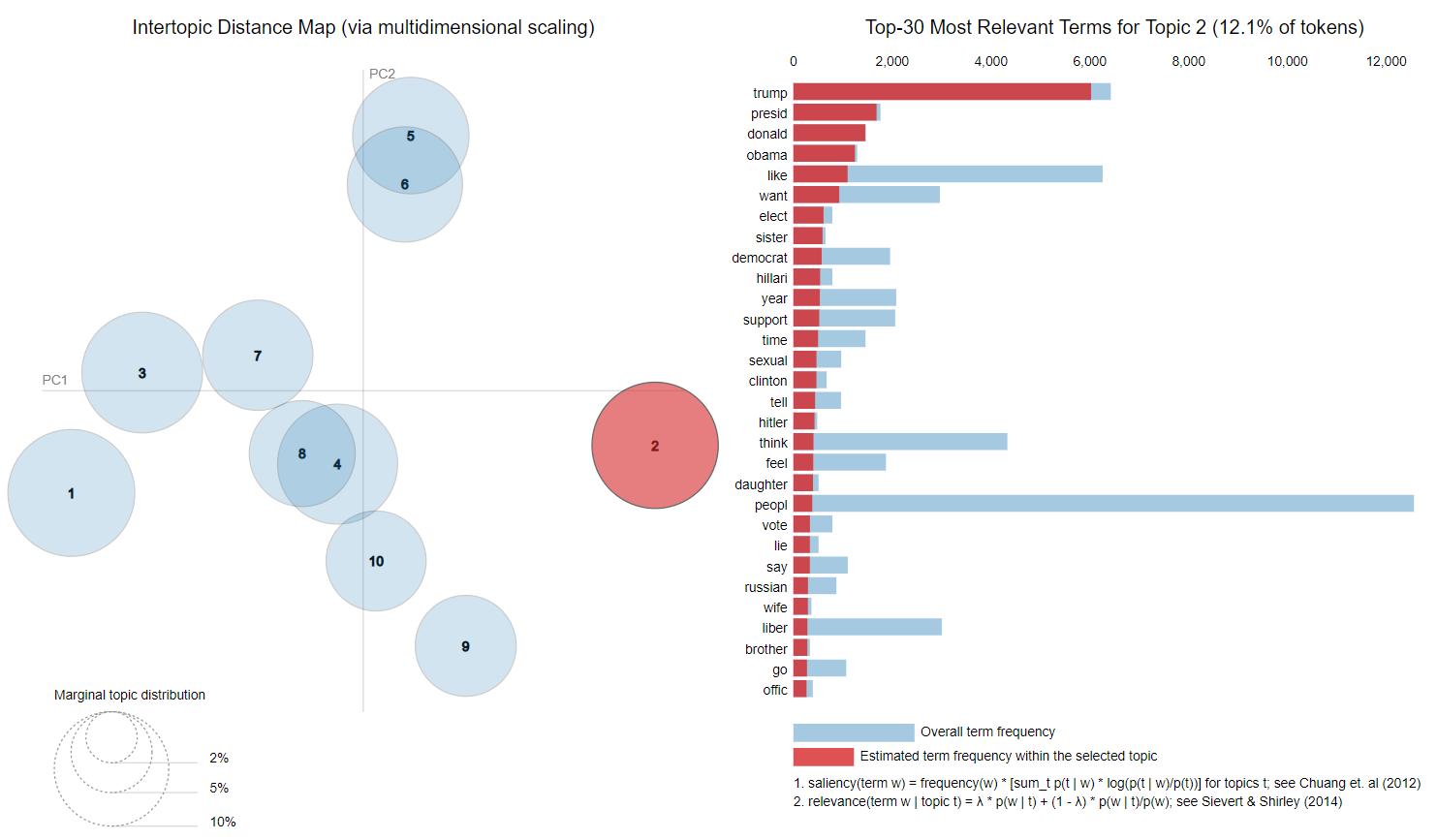
Topic modeling is an important tool in recommendation engines, where the books a person has read in the past, for instance, can be modeled to gain insight into what that person would want to read in the future. Newswire services may use it to cluster articles together or to recommend stories. For the topic modeling portion of this report, the pyLDAvis package was used. LDA topic modeling imagines a fixed set of topics, each of which represents a set of words. The task of topic modeling is to identify topics that best describe a set of words, and each of these topics will emerge during the modeling process. As such, this process is described as a latent process. The LDA in pyLDAvis stands for *Latent Dirichlet Allocation.* Again the name is indicative of the fact that the topics emerge during the process of modeling, as opposed to words being fit into existing topics. The pyLDAvis package outputs an interactive visual that allows the user to see the values of each word in the topic, and the frequency with which the words appeared in the corpus and other topics.

**Examples of Topics**

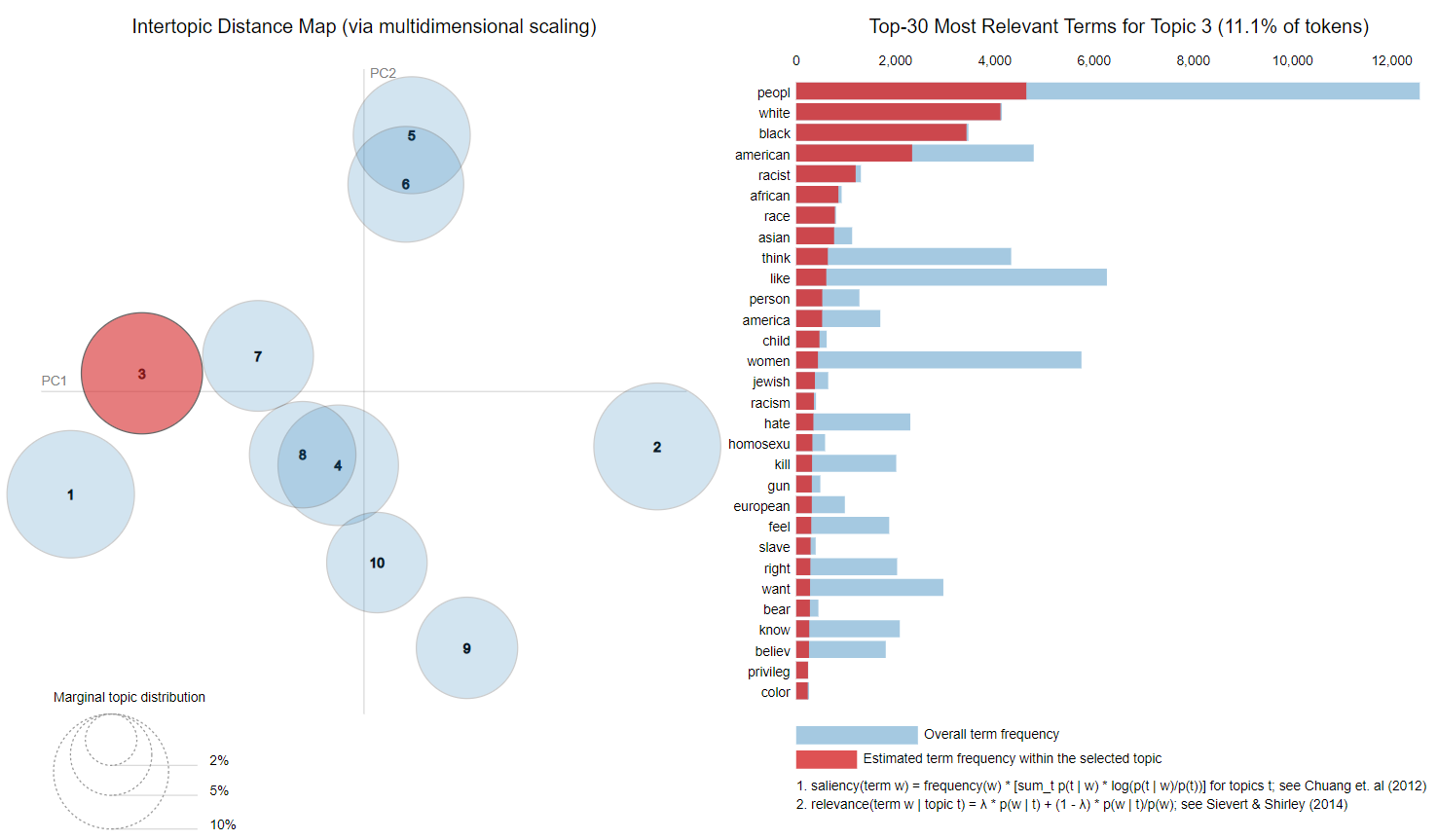
Topic 1 is visualized below. It seems to be a collection of negative sentiments about Islam, especially where the religion intersects with the Asian subcontinent. This is a good example of the kind of content Quora wishes to mitigate.



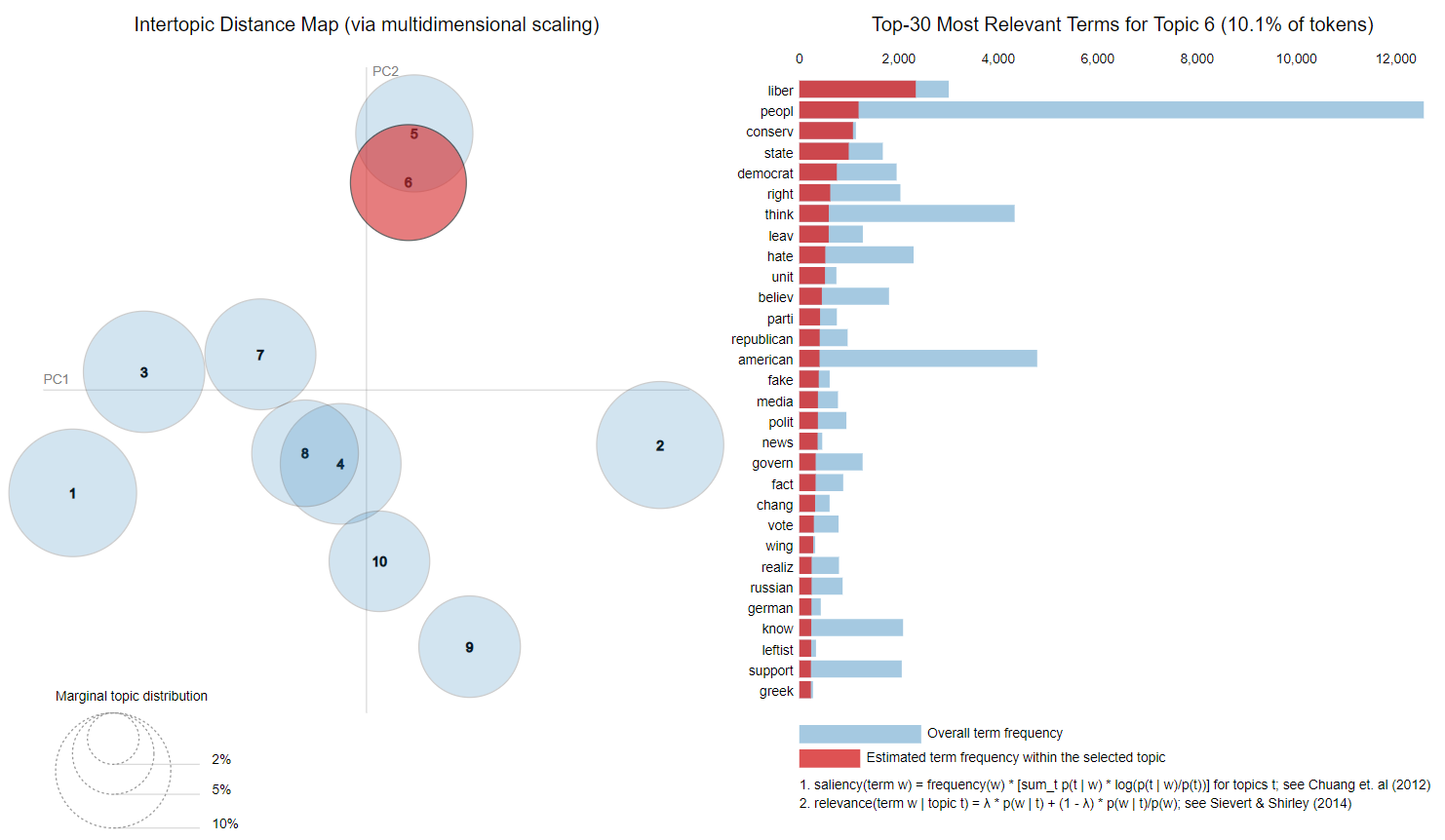
Topic 2 below seems to be about comparisons of presidents Barack Obama and Donald Trump. Again, this is a representation of content that is not in keeping with the vision Quora has for its user-curated library of knowledge. Such subject matter likely consists mostly of opinions rather than the sharing of widely-accepted knowledge.



The LDA visualization for topic 3 can be seen below. The two words nearly exclusive to this topic are ‘black’ and ‘white.’ ‘Racist,’ ‘racism,’ ‘African,’ and ‘race’ are also almost exclusive to this topic, which can be seen in the red lines of the graph. An important part of Quora’s mission is to provide knowledge free from toxic attitudes, and racism is toxic by nature.



In topic 6 below, the general theme is politics. While topic 2 was specific to the current and most recent former American presidents, topic 6 seems to be more generalized to the political spectra the two most recent administrations occupy. While political discussions can and often are civilized and valid forms of debate, internet forums are not typically places where philosophies can be deliberated reasonably. An open forum is much more likely to be seen as a safe place to loudly espouse opinions which may or may not be rooted in careful consideration and critical thought. As such, political content can quickly become pernicious and detrimental to Quora’s mission.



***Conclusion***

While the results of this analysis were promising, many of the contest entries (visible to the public on Kaggle) were better. The analysis was challenging in many ways, the most significant being the imbalance in the data. It was also challenging in the sense that insincerity can be difficult to judge. While the topic modeling found that politics, religion, and race were frequent subjects in the insincere questions, it is not necessarily true that all questions around these matters are insincere. Quora’s users may sometimes want an earnest explanation of the differences between liberal and conservative politics, or there may be users who are genuinely interested in comparing and contrasting religious tenets.

The results of the experiments with balanced data were promising, with accuracy more than 20% above the baseline. These experiments yielded good results in terms of insincere prediction, which suggests further experiments could possibly be tuned to produce the kinds of results Quora is seeking. Experimenting with the data in different configurations of balance could prove fruitful.

The models trained on the skewed data did not render results that performed well above the “random guess/majority rule” baseline, although they were able to improve the accuracy by a small margin. This could be explained by insufficient cleaning of the data, or lack of granularity within the vectorization. Limited processing power rendered it difficult to eliminate non-English words, as checking every word within the massive dataset against the NLTK corpus was not tenable. The presence of these words may have contributed to the increase in false positives within the model.

Additional vectorization approaches may also have benefited the accuracy of the models. A Word2Vec approach may have provided a good alternative to the tfidf vectorization approach of giving weight to distinguishing features within the text. Because of the topical associations within the LDA among insincere questions, further analysis within those topical subsets could also provide more insight into the issue.

While modifying the kernel in the SVC did not provide a variation in results, as the accuracy did not change over the course of a 10-fold cross validation, an examination of over-fitting may explain this shortcoming. Although several gamma values were used to combat this, it is likely that the combinations used thus far were insufficient to offset the skew in classification of the dataset. Combining this approach with the control over the number of insincere questions within the training and testing subsets, as implemented with the NB models, could yield more informative results.

Of all the experiments, topic modeling provided perhaps the most easily interpretable results. Since Quora’s goal is to predict insincerity, topic modeling can provide some clarity about what general topics are found in the questions identified as insincere. Unfortunately, this requires that the classification be done prior to the modeling, but it can provide extra insight into what the prediction models need to look for. At the very least, topic modeling revealed the potential toxicity of conversations around politics, religion, and race, especially in online forums where a veil of anonymity can shield users from the social cues that normally mitigate anti-social behavior.

***References***

1. <https://www.kaggle.com/c/quora-insincere-questions-classification/data>
2. <https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158>
3. <https://www.kaggle.com/aashita/word-clouds-of-various-shapes>
4. <https://www.kaggle.com/tunguz/just-some-simple-eda>
5. <https://www.kaggle.com/lopuhin/eli5-for-mercari>
6. <https://www.kaggle.com/sudalairajkumar/simple-exploration-notebook-qiqc>
7. <https://gizmodo.com/the-first-internet-troll-1652485292>
8. <https://www.kaggle.com/getting-started/44916>
9. <https://www.cs.waikato.ac.nz/~eibe/pubs/FrankAndBouckaertPKDD06new.pdf>
10. <https://nlp.stanford.edu/IR-book/html/htmledition/the-bernoulli-model-1.html>
11. <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>
12. <https://towardsdatascience.com/support-vector-machine-vs-logistic-regression-94cc2975433f>