Fake News Challenge
COMP9417 Assignment 2

Ambrose Hill z3470165
Yuvraj Singh z5059978

Semester 1 2018

Contents

1 Introduction

We have chosen to tackle the Fake News Challenge, topic 1.5 in the list. In recent years, more and more focus has been placed on social media and new outlets to provide a thorough analysis on what is true in modern media. The reason for this is the spread of misleading or 'fake' news has become increasingly popular and is starting to affect how our society at large consume media and come to meaningful conclusions. We are now starting to realise the true potential of power that social media holds and how persuasive the spreading of misleading articles can be. In response to this there is a lot of research going into detecting fake news and how to avoid it. There are even countries, such as Taiwan, that have implemented new curriculums for children in school that teach them how to identify and combat fake news. Since this is such a popular topic currently, we decided to try and tackle a simplified version of this problem, namely, given a headline and body article, determine/predict whether the body is related or unrelated to the headline. This would allow media devoted entities to split data based on a probability of it being unrelated to the article. This could be looked at as a first step in detecting fake news in a world filled with distracting and misleading pieces. As this would be a text classification problem we decided on using the Naive Bayes Algorithm.

2 Implementation

This challenge fundamentally is a text classification problem. We are only given the articles headline and body and need to predict the author's stance. This stance can be one of two categories to begin with, unrelated or related. If the article is determined to be related, it requires an additional piece of information stating the strength of the relation, whether the article agrees, disagrees or discusses the given headline.

Since this was a competition challenge that has already concluded there was no need to collect data as they made the competition datasets available. This included a training and a test set as well as a python script to score your results on the test data. As this is a text classification problem with a large dataset we decided to use The Naive Bayes algorithm. Although it is considered a 'naive' approach it has been proven to work accurately on large datasets and especially in relation to text classification. The major advantage of using such a simple algorithm is that it is quicker to get a operating solution implemented. From there it is a matter of optimising and tweaking the manipulation of the data to increase accuracy and build a better model. Since this was part of a larger challenge we were supplied with both training and test datasets.

To begin with we researched into how the average headline is written, exploring the underlying structure typically present. As the problem lies with whether or not an article's body is related to its headline, understanding the structure of typical headlines is essential. We were presented with a simple structure when it comes to headlines, using the below example of a headline:

Coast Guard (S)	seizes (V)	ship (O)	in drug raid (PP)

This structure will generally be:

The Subject (S) of the headline
A Verb (V) describing what the subject is doing to
The Object (O) of the topic

A Prepositional Phrase (PP), which is further broken into a Preposition and Object.

Relating this to the problem, we can determine if an article is discussing the problem at hand from this. We began by loading all the data into a class that contained, body, headline and a stance, not yet identified. We then organised that data in an attempt to determine if the body was related to the headline of the article, observing how many words from the headline appeared in the body. To predict whether or not the article was related we broke the headline into its individual words and attempted to identify the subjects and objects and their presence in the body. If the subjects and objects of a headline are present in the body of the article, it stands to reason that the article would be discussing them. As the number of words used to represent subjects, objects as well as verbs and prepositions, or any other words present in headlines that don't adhere to the base structure. As a result, we instead used a proportion of how many words in the headline appear in the article.

To maintain this data we implemented two main data structures. A dictionary that keep track of all seen words and a word count which kept track of how many times a word was seen in an article and sorted by what stance it was. To load the data we created a FakeNewsClass consisting of a headline, body and stance. We then loaded each headline and corresponding body into an object and compiled them into a list, which we then iterated through to populate the data structures mentioned previously. The equation we use to determine whether an article is related or not is:

$$unrelated = \sum(x) > \sum(y) \times 60\%$$

Where x is a word in the headline that does not appear in the body and y is a word in the headline that does appear in the body. Unrelated is true when less than 60% of the terms in the headline are present in the body.

If the prediction was unrelated our classifiers job was complete. However if the article was determined to be related the next step was to determine how strongly to article was related. To do this we need to establish a word count of each word in the body and headline and relate them. This is done through our classifier, which reads all inputs words presented, head and body,

and adding them into a hash table for frequency, and dictionary to know what words have been covered. The frequency hash table in particular is represented in two dimensions, the first key being the stance and the second being the word as key. This lets us determine how often a word will appear in each different stance.

Once all the training data has been read, we begin working on predictions for all articles which have been categorised as related. We iterate over each word, comparing it against the three possible stances (agree, disagree, discuss), using Bayes conditional probability rule to calculate the posterior probability of each stance. The stance with the highest probability becomes the classification for the article. The equation we use to calculate the prior probability of each stance is:

$$\sum \log\left(\frac{x}{totalWords}\right)$$

Where x is the number of times that word appeared in the article. The highest of these scores becomes the prediction.

The implementation itself is one that is quite costly, needing to iterate over every word and calculate its probability of appearing given each individual stance. In addition, we take the log of this probability, to translate the value into a larger value (to avoid issues from python having too small a value, and treating it as zero). As our goal is to find the largest probability of the three stances, and the log function being a monotonically increasing one, this allows us to take the log without hindering the results.

3 Experimentation & Results

We used the datasets provided by the fake news challenge for our experimentation. Due to the time cost of running our program on large sets, we chose to use subsets of the testing data instead. To begin with, we needed to test the accuracy of identifying related and unrelated articles based on the headline and body. Using a script to treat the stances 'discuss', 'agree' and disagree' as the single stance, we chose to test on sets of size 100, 1000 and 5000 to get our results, we found approximately:

Set of size 100, correctly identified

$$\frac{92}{100}$$

Sets of size 1000 correctly identified

$$\frac{844}{1000}$$

Sets of size 5000 correctly identified

$$\frac{4312}{5000}$$

Overall having an average accuracy of approximately 87.55%

From here, we then needed to proceed with attempting to use the Naive Bayes classifier on the related articles, as well as a few modified cases. We attempted three modified forms of the classifier, the first using only terms from the headline, the second using some words in a list of terms to ignore, and finally using the body alone during classification. The averaged results of correct classifications for these different forms are as shown below:

No. Articles	**Unmodified Naive Bayes**	**Headline Terms**	**Ignore Word List**	**Body Only**
100	84	80	84	86
1000	828	815	826	801
5000	4110	4036	4098	4088

Table 1: Averaged Results from Different tweaks on Naive Bayes

The first modification was done in an attempt to note the presence of the headline specific words, in a suggestion that focussing on the headline and its terms can better focus on its stance. This however ended up having the worst overall results. This is likely a result of a majority of the headline being situational, meaning that the frequency of the subject and object of the headline not being an indication of its stances.

The next modification was to create a list of words to ignore. This was fundamentally common pronouns, "he, she, him, her, them, they", as well as articles, "a, an, the, it", which would have no major stake in determining the stance while being prominent in articles. This ultimately led to similar results, seeming to have no significant (or any in this case) evidence of improvements over the Naive Bayes function alone.

The choice to use the body was done simply to observe how the headline might impact the result. The results again were similar, though still showed no evidence of possible improvements. This suggests that accounting for the headline in Naive Bayes places some stronger focus for the terms, and allows the stance to better be determined.

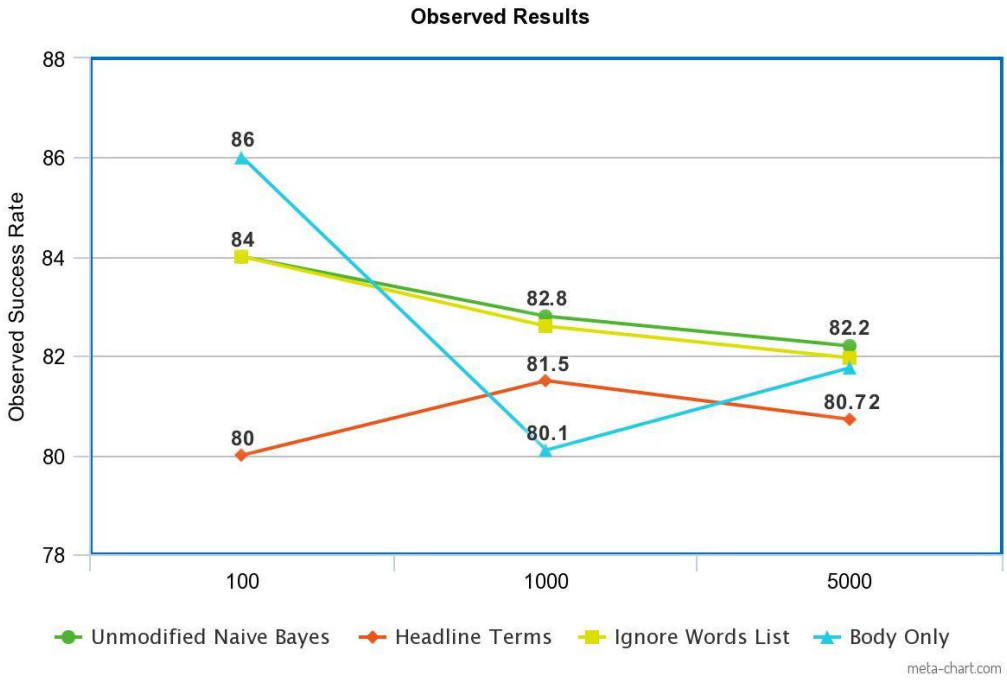
4 Verification

To verify that the results were not situational, we also chose to perform a 10-fold cross verification over our data. The training set provided consists of 49972 articles to train with. We split this into 10 subsets of approximately 4997 articles. Each set was then run as test data, with the other 9 sets acting as a training set. The results of these are as shown below:

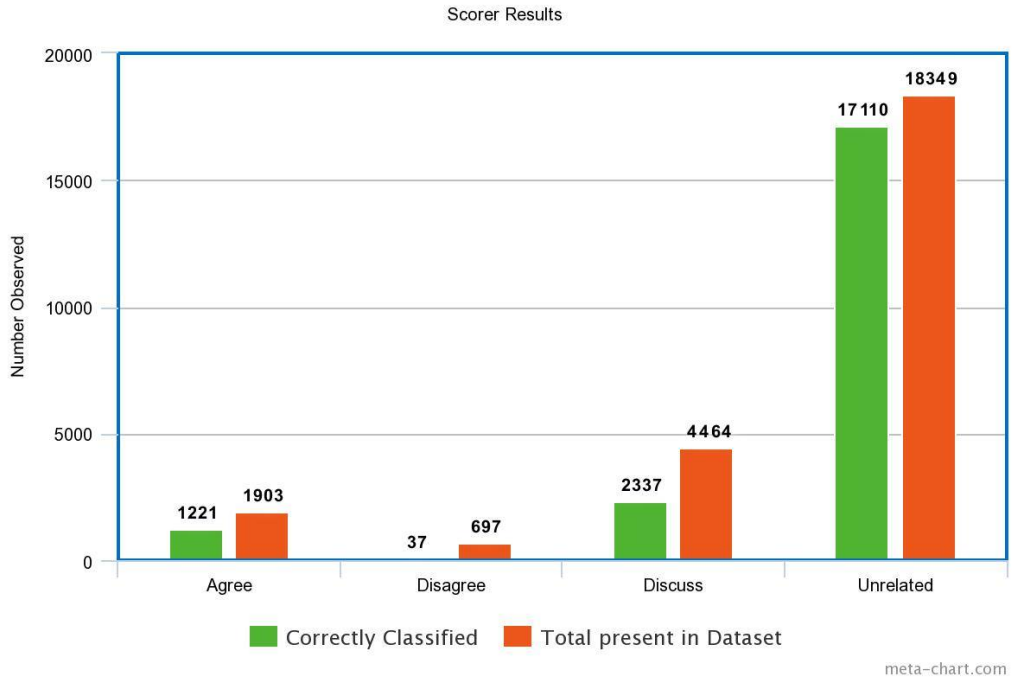
Set (Size)	1(4997)	2(4997)	3(4997)	4(4997)	5(4998)
Correct	4224	4251	4261	4201	4181
Incorrect	773	746	736	796	817
Success Rate(%)	84.50	85.07	85.27	84.07	83.65
Set (Size)	6(4997)	7(4997)	8(4997)	9(4997)	10(4998)
Correct	4228	4254	4218	4250	4241
Incorrect	769	743	779	747	756
Success Rate(%)	84.61	85.13	84.41	85.05	84.87

Table 2: 10-fold Verification Success Rates

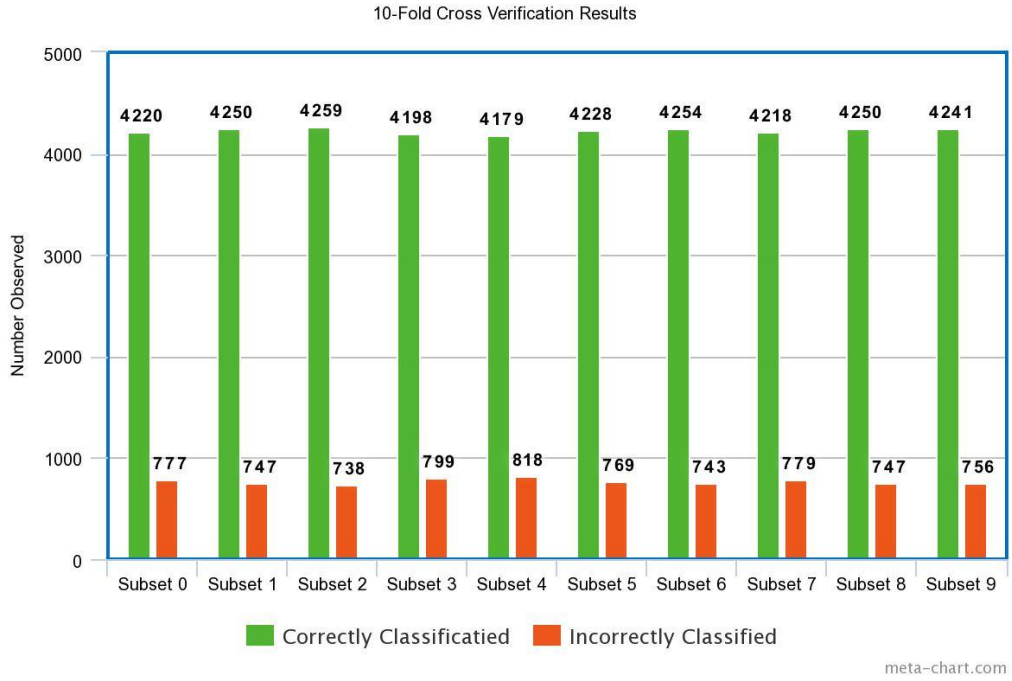
4.1 Result Graphs



Different Algorithms and the effects on results



Results of model tested with the Fake News Challenge test set



Result of the 10-fold verification

5 Conclusion

Our above results led to us retaining a general Naive Bayes function over the words in the headline and body. We attempted to explore different variations in hopes of improving on the function, however all attempts did not have any perceivable improvements from our experiments. Use of the Naive Bayes algorithm with our distinguishing method of related and unrelated articles gave us an approximate success rate of 83%. This is similar to the accuracy presented from our training data too, when running a 10-fold cross verification, suggesting this isn't just chance from the competition data sets. We considered possible improvements that could not be implemented for future goals. Notably, the ability to indicate the type of words in an articles headline and body, noting the subject/object/verbs etc, can allow for more accurate identifications of news articles by being able to focus on the terms that are relevant. This can also allow for potential recognition of stances from like terms, noting terms which appear together rather than the 'naive' choice to simply observe their frequency individually. These however require their own research and implementations that were beyond the scope of this project, but do allow potential for improvements in future.

References

[1] Fake News Challenge
https://github.com/FakeNewsChallenge/fnc-1

[2] Naive Bayes Tutorial
https://pythonmachinelearning.pro/text-classification-tutorial-with-naive-bayes

[3] Mike Bain
COMP9417 Lecture Content UNSW, 2018

[4] Structure of Headlines
https://dinfos.blackboard.com/bbcswebdav/library/Library%20Content/Public%20Affairs%20-%20PAD/Newswriting/News%20Headlines.pdf

[5] Taiwan Fights Fake News
http://time.com/4730440/taiwan-fake-news-education/