NLP Project Report

Elisa Ronga

Purpose

The purpose of this project will be to classify pieces of text into ‘opinion’ or ‘fact’.

Context

Especially during the pandemic and through the more public bipartisan outcry, there has been a slew of climate, vaccine, political, economic and demographic misinformation spreading throughout the world. The reason why this type of information ‘sticks’ is due to its provocative nature. The dangers of misinformation may result in mental or physical harm (Nelson et al., 2020), governmental coups/distrust (Anti-defamation league, 2021) and impact citizens lives and decisions. One way to combat this is by checking and reviewing information, as well as evaluating the trust of the sources. A preliminary approach is to evaluate if the information is a fact or an opinion.

Dataset

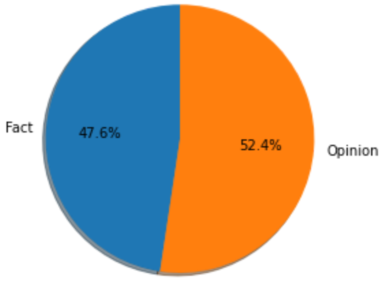
There are two datasets that I will be using, the first one is a dataset of movie plots and movie reviews. They are cleaned so that only the fact and opinion of the dataset are seen, respectively. Below are experts from both sets of data:

```
Led by Woody, Andy's toys live happily in his room until Andy's birthday brings Buzz
Lightyear onto the scene. Afraid of losing his place in Andy's heart, Woody plots against
Buzz. ...

One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be
hooked. They are right, as this is exactly what happened with me. ...
```

Given this data, the datasets are spread out as can be seen below:

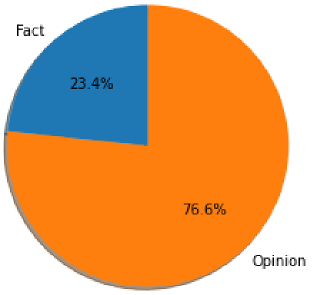
Frequency of Dataset Pre-Preprocessing



With `Count of facts:  45466 and Count of opinions: 50000.`

Doing some preprocessing where the sentences are separated, the following graph reflects the actual sentences:

Frequency of Dataset per Sentence



```
Count of facts:  164896 and Count of opinions: 541161.
```

This is important to know in order to normalise for the opinions and train on an even amount of sentences.

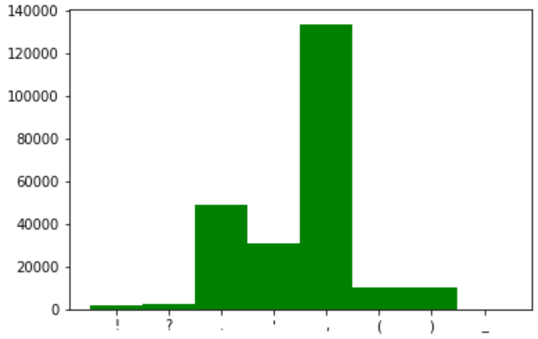
Preprocessing Analysis

Before inputting the dataset into the models, a brief analysis of punctuation frequency, words frequency per sentence and key word frequency is performed to check if there is any

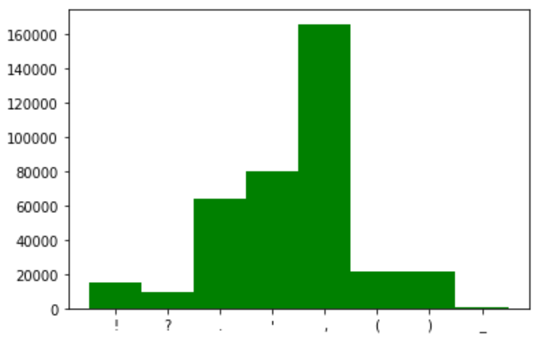
Punctuation Frequency

The punctuation frequency, normalised by the ratio of fact sentences vs opinion dataset is as follows:

Punctuation Frequency of Facts Dataset



Punctuation Frequency of Opinion Dataset



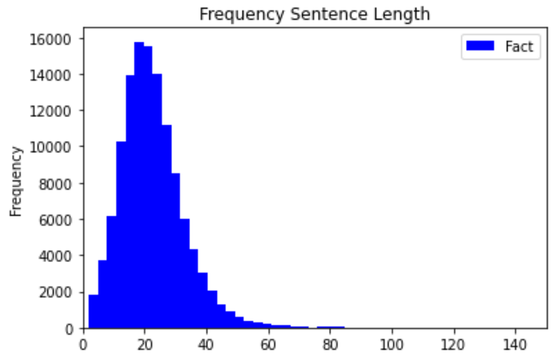
	Fact	Opinion	By how much?
!	1546	14981	9.690168
?	2765	9854	3.563834
.	49119	63793	1.298744
`	31173	80370	2.578193
,	133694	166010	1.241716
(	10036	21203	2.112694

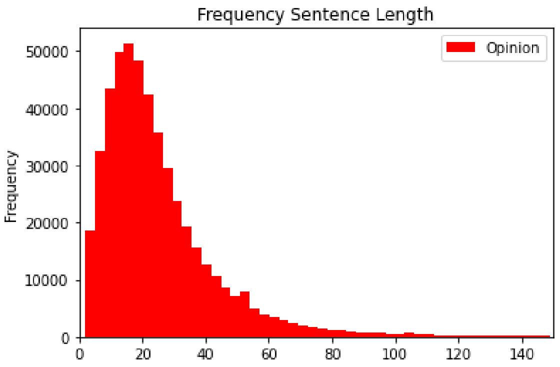
)	10060	21716	2.158648
_	19	304	16

As can be seen by the graph and table, punctuation was more used in the opinion dataset, and especially the underscore, exclamation point and question mark, even though all of the punctuation had an impact on the classification of fact vs opinion.

Word Frequency per Sentence

The word frequency in each sentence doesn’t necessarily have to be normalised per dataset since the average was calculated, and an understanding of the frequency of each in a histogram is shown, which isn’t affected by the amount of data (if one only wants to look at the histogram shape).





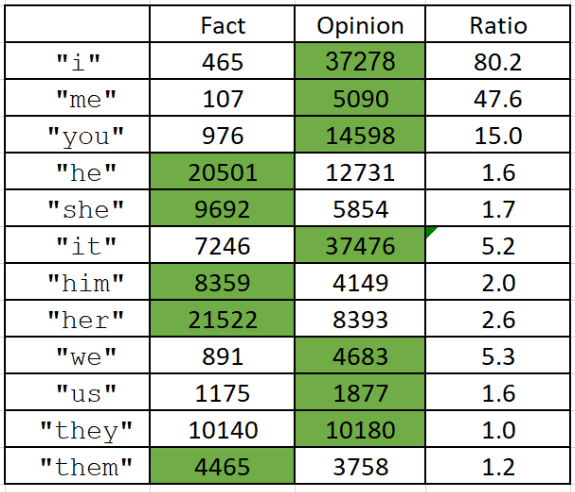
The average word frequency are 22 and 26 for fact and opinion sentence respectively. Additionally, from the graphs above, opinion sentences have a wider range of word frequencies. This is concurrent with the idea that opinions are more free-form, with less structure to the delivery as a fact would be.

Key Word Frequency

Hypothesised kew words frequency seemed also important, given that opinions tend to have a higher frequency of personal pronouns. The pronouns checked were:

```
["i", "me", "you", "he", "she", "it", "him", "her", "it", "we", "us",
"they", "them"]
```

The data was normalised based on the amount of sentences that are in each dataset, to get a comparable number.



	Fact	Opinion	Ratio
"i"	465	37278	80.2
"me"	107	5090	47.6
"you"	976	14598	15.0
"he"	20501	12731	1.6
"she"	9692	5854	1.7
"it"	7246	37476	5.2
"him"	8359	4149	2.0
"her"	21522	8393	2.6
"we"	891	4683	5.3
"us"	1175	1877	1.6
"they"	10140	10180	1.0
"them"	4465	3758	1.2

As hypothesised, personal pronouns like 'me' and 'I' are much more abundant in the opinion dataset, even after normalisation. Third person pronouns are more common in the fact dataset, but not by as much.
Given the important that the frequency of words have on classifying fact and opinion dataset, bag of words word embedding will be used for the models.

Word Embedding

Bag of words

Bag of words will depend on the frequency of each word, which is useful given that in the previous section there were key words that seemed to weigh a lot in the classification between opinion and fact.

tf-idf

Difference to bag of words, there might be rarer words that are more vital to differentiating between the classifications. This is why tf-idf is also used as a comparison, to see if this is the case with the dataset.

Model Implementation + Results

Naive Bayes

Naive Bayes is used for classification, especially when wanting to check the probability of a sentence belonging to a specific class. After training, accuracy for training data is 90% and for testing data is 87% with the BOW embedding.

After training, the accuracy of the training data is 90% and the testing data 88% using td-idf word embedding.

Logistic Regression

Due to the binary nature of the classification, logistic regression (still trained with bag of words embedding and tf-idf) was performed.
Bag of words accuracy is 93% on training data after training and 87% on testing data after training.
For tf-idf word embeddings, accuracy after training for training data was 91% and for testing data is 88%.

Random Forest Classifier

Random forest classifier is another model that can be used for classification, looking at decision trees but making sure that the data doesn’t overfit by balancing the classification. Due to the higher work time, the dataset was cut to only 40000 sentences, instead of the 200000 used previously.
Bag of words accuracy is 100% on training data after training and 87% on testing data after training.
For tf-idf word embeddings, accuracy after training for training data was 100% and for testing data is 87%.

Support Vector Machines

Support Vector Machines is another classification methods that uses the support vectors (which are the data points) and has a hyperplane to classify between these support vectors. The radial basis kernel function (rbf) is the default one.
Bag of words accuracy is 95% on training data after training and 87% on testing data after training.
For tf-idf word embeddings, accuracy after training for training data was 98% and for testing data is 88%.

Conclusion

word embedding	model	Dataset	accuracy	precision	recall	f1
BOW	naïve bayes	240768	87%	86%	88%	87%
tf-idf	naïve bayes	240768	87%	86%	87%	87%
BOW	logistic regression	240768	87%	88%	87%	87%
tf-idf	logistic regression	240768	88%	89%	86%	88%
BOW	random forest classifier	40000	87%	86%	87%	87%
tf-idf	random forest classifier	40000	82%	76%	92%	84%

BOW	SVC (rbf)	20000	87%	89%	85%	87%
tf-idf	SVC (rbf)	20000	51%	82%	2%	4%
BOW	SVC (poly)	20000	74%	93%	53%	67%
tf-idf	SVC (poly)	20000	76%	69%	92%	79%

As can be seen above, none of the models with word embeddings combination reached over 90% accuracy, but the best one was the logistic regression using ID-IDF word embedding. Hyper Parameters were changed but none performed as well as the default parameters.

Future experiments/Work

Next steps include two parts:

1. Try different models and hyperparameters, specifically CNNs and LSTM to understand if the statistical results can be improved. This wasn't done because the embedding would've been different
2. Test with news dataset, to see if it can be applied to newspapers. This wasn't done because the dataset doesn't exist for news that is opinionated.

Citations

Nelson, T., Kagan, N., Critchlow, C., Hillard, A. and Hsu, A. (2020). The Danger of Misinformation in the COVID-19 Crisis. Missouri medicine, [online] 117(6), pp.510–512. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7721433/ [Accessed 8 Apr. 2022].

Anti-Defamation League. (2021). *The Dangers of Disinformation*. [online] Available at: https://www.adl.org/education/resources/tools-and-strategies/the-dangers-of-disinformation [Accessed 8 Apr. 2022].