Introduction

The purpose of this work was to create a subject specific sentiment prediction model, test it against data from a different domain and compare the results with those gained from a generic sentiment analysis tool.

The subject specific classifier was expected to provide the best accuracy of any of the models when predicting for its own subject, but the worst accuracy when predicting for a subject from a different domain. The generic model was expected to provide an intermediate accuracy level between the two extremes of prediction of the subject specific model.

Tweet data was collected for the search term ‘Thunberg’, referencing climate activist Greta Thunberg. The tweet texts were cleaned, a few additional features were extracted, and positive/neutral/negative sentiment labelling was completed manually. Following limited feature engineering, an SVM classifier model was built and tested on a test set of the data. The sentiment analyser, VADER, was then also used to predict sentiments for the test data set. Finally, the classifier was also tested on tweets collected under the search term ‘Vegemite.’

Overall, the subject specific SVM classifier performed as expected. VADER results were not as high as expected, mainly because the tweets it predicted for had already been cleaned and did not contain the emoticons and other features that help it achieve better performance.

Data

Tweet data were collected from Twitter using the Twitter basic search API and tweepy (see *app.py*) 400 tweets were collected for the search term Thunberg, referencing Greta Thunberg, and after cleaning, duplicate removal (see *tweets.py*) and manual sentiment labelling (see *labelling.py*), 382 Thunberg tweets were available. This process was repeated for the search term 'Vegemite' to provide a dataset for testing against a different domain, and 389 Vegemite-related tweets were kept.

The data retained about each tweet (features) are the following:

- **Full tweet text**
- Author (Users) account data
    - User-defined location
    - **Number of followers**
    - **Number of users the account is following (friends)**
    - Following (a now deprecated attribute, all entries are Null)
- Geographic location of the tweet (as reported by user)
- Place tweet is associated with (if present)
- **Number of times tweet has been retweeted**
- **Number of times the tweet has been liked (favorites)**

Unfortunately, many features are null in most cases (due to errors in the data retrieval process and users choosing to not add personal details). The features highlighted in the list above contain accurate and useful data and will be used in this project.

Each tweet text was manually labelled with a sentiment: positive, neutral or negative, based on the sentiment it contained towards the subject (Thunberg or Vegemite). The sentiment labelling was carried out by the author after the tweets had been cleaned (for ease of understanding) and the labels were added as a final feature in the dataset (which was stored in semicolon-delimited txt files). The labels are subjective and may contain errors.

The proportion of tweets associated with each sentiment in the Thunberg dataset is:

negative 0.446809

positive 0.340426

neutral 0.212766

Unfortunately, due to the timing of the project and the data collection (26/03/2020), the content of the tweets collected has been heavily influenced by the Covid-19 pandemic. The spread of negative and positive tweets, as well as the accuracy of the classifier and VADER may be different if the data was collected at another time.

SVM Classifier

Following a very brief exploration with different classifier models, an SVM model was chosen as the classifier for this project as it gave slightly higher accuracy rates. The following manipulations attempt to improve the accuracy of the model further):

- Further text cleaning to remove punctuation and convert all text to lower case (see mainline.py)
- Addition of text length feature and number of words feature (see mainline.py)
- Stratifying the test-train split (as the dataset is not balanced and the test size is small, see mainline.py)
- Adding subject specific vocabulary to the stopwords (see pipeline.py)
- Creating a pipeline with feature unions to include all the features in building of the classifier (see pipeline.py)

All together these manipulations added roughly 5 % extra accuracy to the model.

The SVM classifier model gives the following results when predicting on test data:

- THUNBERG TWEETS -

Results from SVM Classifier

Accuracy: 0.6052631578947368

Confusion matrix:

	0	1	2
0	28	0	6
1	10	2	4
2	9	1	16

(The confusion matrix labels: 0, 1, 2 refer to negative, neutral and positive sentiment respectively.)

The confusion matrix shows that the classifier is over predicting by the greatest margin with the negatives, but also overpredicting on the positives. Given the context, it is useful that it does not overpredict the neutrals, as little is gained from falsely predicting neutral.

Comparisons and Testing

Thunberg data with Vader

The NLTK VADER Sentiment Intensity Analyzer gives polarity scores for the sentiment of each text it analyses. For this project, the 'compound' score has been used as it is a sum of the normalised ratings, and so effectively gives an overall score. It is returned as a number between -1 (full negative) and 1(full positive). This score has been converted into a label based on the proportion of each label in the original dataset. For example, as 45% of the tweets in the dataset are negative, score in the lowest 45% of the -1 to +1 range (i.e. -1 to -0.1) have been converted to a negative label.

As the NLTK VADER Sentiment Intensity Analyzer is a pre-prepared model, the full dataset is not used to train the model. Therefore, to simplify comparison with the SVM classifier model, only the test data was run through it. It gave the following results:

- THUNBERG TWEETS -
Results from VADER Sentiment Prediction
Accuracy score:
0.35526315789473684
Confusion matrix:

	0	1	2
0	14	13	7
1	6	7	3
2	10	10	6

(The confusion matrix labels: 0, 1, 2 refer to negative, neutral and positive sentiment respectively.)

When the whole dataset is run through VADER, the accuracy score increases slightly to 38%. Here the model has most difficulty in correctly identifying positive sentiment; only 6 of the 26 positive tweets were correctly labelled.

The VADER model, developed from a non-subject specific dataset, has an accuracy rate at just over half the accuracy of the subject specific SVM model (36% compared to 61%). The subject-specific model was expected to give better results than the 'generic' model but this difference is larger than expected due to the lower-than-

expected VADER results. This is primarily because VADER is good at labelling features such as emoticons and emojis which have been removed from this dataset. Additionally, the comparison is not direct because the SVM model incorporates more features than the VADER model, such as number of followers and number of retweets, so the classifier has access to more data on which to base its predictions.

Vegemite data with the SVM model

The SVM classifier model was then tested with data from a different domain, on the subject of Vegemite. A test dataset was extracted from the full Vegemite dataset in the same way as for the Thunberg data and the model was tested against this test set, so that the results are comparable. Below are the results:

- VEGEMITE TWEETS -
Results from Thunberg-trained SVM Classifier
Accuracy: 0.3076923076923077
Confusion matrix:

	0	1	2
0	16	0	1
1	13	6	1
2	38	1	2

(The confusion matrix labels: 0, 1, 2 refer to negative, neutral and positive sentiment respectively.)

When the model is tested against the full dataset, a very similar accuracy of 28% is given.

When predicting for a different topic, the subject specific SVM has an accuracy rate of half that when predicting for its own subject (31% compared to 61%). Therefore, a subject specific classifier can be effective where either the subject of the input can be guaranteed, or where errors can be accepted if the subject is misaligned.

The results given for the Vegemite tweets from the subject specific SVM Classifier are close to those given for the Thunberg tweets from the generic VADER model (at 31% and 36% respectively). Although VADER may have given better results if the full tweet text (including emoticons and hashtags etc) had been used, this similarity suggests

that any sentiment prediction model, whether subject specific or generic performs similarly on an independent dataset. More research and analysis would be needed to explore this further, as there is also a performance difference between the SVM model and the lexicon analysis-based VADER model.

Vegemite data with VADER

As a final step in the comparison, VADER sentiment analyser was also used to predict for the Vegemite test set. The results are as follows:

- VEGEMITE TWEETS -
Results from VADER Sentiment Prediction
Accuracy score:
0.4230769230769231
Confusion matrix:

	0	1	2
0	7	4	6
1	3	8	9
2	7	16	18

(The confusion matrix labels: 0, 1, 2 refer to negative, neutral and positive sentiment respectively.)

VADER gives a better score for the Vegemite data than it did for the Thunberg data. This may be due to more complex opinions and language relating to Thunberg as it deals with a political topic, compared with Vegemite which generally draws out a like it or loathe it opinion.

Conclusions

A range of manipulations were carried out on the dataset and included in the classifier model pipeline, in an attempt to improve the accuracy of the model. The increase in accuracy achieved was small, but further manipulations may increase it further. For example, additional features could include

- Average length of non-stopwords
- Retweets per tweet and/or favorites per tweet

Following the limited feature engineering, an SVM classifier model was created which gave 61% accuracy when predicting for test data from the original dataset.

This model was also tested on a dataset from a different domain, namely ‘Vegemite’, giving an accuracy of 30%. This shows that the SVM Classifier is very domain sensitive, giving much better results when predicting for tweets on the training topic, as expected.

Sentiments were then also predicted for both datasets using the off-the-shelf sentiment analyser VADER, giving results of 36% and 42% respectively. This was weaker than expected mainly because VADER performs better when emoticons and other microblogging elements are included. The tweets used here had already been cleaned. A future improvement would be to collect uncleaned tweets and carry out the cleaning as part of the modelling pipeline, rather than cleaning at collection.

Finally, the VADER model gave better results when predicting sentiment for the Vegemite tweets than for the Thunberg tweets and this may be due to the complexity of the subject. This complexity of the political opinions associated with Thunberg may have also weakened the accuracy of the SVM model when predicting both for the same subject and for a different subject. To analyse this further, the Vegemite data could be used to build the SVM model and tested against the Thunberg data.

References

1. Baghern, D. A Deep Dive into Sklearn Pipelines. Available at: https://www.kaggle.com/baghern/a-deep-dive-into-sklearn-pipelines/data [Accessed 17th April 2020]