Natural Language Processing Assignment Report

wcrr51

1 Introduction

In recent years, there has been increased political and academic discussion and research around the issue of fake news, from its influence on elections to unjust outrage directed towards people, companies, or governments. One of the main forms of fake news is *click-bait*, this is where the content of an article does not match up with its title, or the title as reported from outside the article [1].

There are many data sets aimed at providing researchers a benchmark for comparing their detection techniques. For example, Potthast et al. [2] provides a data set of around 30,000 tweets. The data set that will be used in this report is the Fake News Challenge (FNC-1) data set by Misback and Pfeifer [3].

The FNC-1 data set was originally intended as part of the competition to perform the first step in the fake news detection pipeline, however it still serves as a good benchmark to test new language and classification models on.

2 Problem Definition

The problem of detecting first-stage detection of fake news is defined as follows. Given a headline and potentially corresponding body text, classify them into one of the four following categories:

- Agrees: The headline is reflective of the body.
- Disagrees: The headline is not reflective of the body.
- Discusses: The headline is reflective of the body, but does not take a position.
- Unrelated: The headline has a different topic to the body.

3 Proposed Solutions

3.1 Feature Extraction

3.1.1 Data Reprocessing

The FNC-1 training and testing data sets each come split into two tables: one for headline, body ID, and stance labels, and the other for bodies. For all models in the solution, a standard procedure is followed for pre-processing the data. Firstly, the title and body in their respective tables are cleaned by converting them to ASCII (removing non-English characters), removing newline characters, converting to lower case, and removing punctuation. Further classifier or model-specific data preprocessing is performed and discussed in their respective subsections.

3.1.2 TF-IDF

Term frequency-inverse document frequency (TF-IDF) is a statistic for measuring or ranking the importance of words within a document belonging to a larger corpus, making it a popular method for feature extraction. For each word not in a blacklist of semantically unimportant words, its value is calculated, the words are sorted by their value, and those meeting a threshold are extracted as features.

For TF-IDF, English *stop words* provided by `spacy` are used to reduce the number of low-entropy output features as these would hinder. Similarly, lemmatisation is used in pre-processing to convert common variations of words into their most basic form, again reducing the number of output latent features. Finally, a minimum data frequency value of 2 is used to remove single-use words which may include one-off typographical errors.

3.1.3 BERT

The next ML technique applied for feature extraction is bi-direction encoder representations from transformers (BERT) [4], which stacks many transformer encoders [5] together forming a language model capable of deep understanding and representation of input text. For feature extraction, BERT outputs one 768-sized vector per input token.

BERT is powerful because it comes with a pre-trained English language model, `bert-base-uncased`, trained from large data sets such as BookCorpus [6] and Wikipedia [7]. Using the pre-trained model, BERT can then be specialised, which, in this case, will include feature extraction and classification. As the transformer encoder stacks are naturally bi-directional (enabling good parallelisation to enhance training speeds), BERT is additionally able to learn information and context from both left to right and right to left.

Some disadvantages of BERT are that the model is very large due to its large corpus and training structure. Training it and its derivative task-specific extensions is slow as it has many weights and biases to update. Additionally, it may require some further fine tuning to ensure its applied downstream tasks work correctly. Furthermore, documents are limited to 512 tokens, which some of the concatenated headline and article bodies in the FNC-1 data set exceed, meaning they must be truncated, potentially leading to loss of context.

3.2 Related/Unrelated Classification

Firstly, models are trained to identify whether headlines are related to article bodies (agree, disagree, or discuss stance labels), or not (unrelated stance label).

3.2.1 Support Vector Machine Classifier

To gauge how well traditional machine learning (ML) approaches perform on the broader two-state problem, the features and embeddings provided by TF-IDF and BERT were first used to train their own support vector machine classifiers. The parameters used to fit the SVM classifications were $c = 1$, with a radial basis function (RBF) kernel, and a degree of 3. The models are evaluated using the competition testing set.

3.2.2 Deep Classifier

A deep classifier is trained using the hybrid features of the top 500 TF-IDF tokens and the standard 768-element BERT sentence embeddings (concatenated together). This attempts to give the trainable parameters the best of both worlds, enabling them to make decisions based on local knowledge within a sentence (BERT), or global knowledge about the most important words as determined by TF-IDF, which is important for the task of deciding whether text content is related. This model is evaluated during training using a split of the training set, and evaluated after training using the competition testing set.

3.3 Agree/Disagree/Discuss Classification

Models are now trained to identify, of those documents who are related, which of the heading-article pairs agree, disagree, or simply discuss with no bias. Unlike with related/unrelated classification, only a deep learning model is considered here. For this, a custom-implemented stacked gated recurrent unit (GRU) network is trained on the same articles, this time pre-processed to remove documents with a stance label of *unrelated*.

3.4 Combined Classification

Finally, a classification model to determine which of the full four stances a headline-body combination belongs to is used. This simply calls to the deep classifier used in Section 3.2.2, and if it deems the headline to be related to the body, calls to the classifier in Section 3.3 to predict the bias in the article.

4 Analysis and Discussion of Results

4.1 Preprocessing

ASCII and lowercase are used to reduce the total number of unnecessary word features for those with practically the same meaning. Similarly, punctuation is also removed to avoid having to tokenise punctuation separately. It could however be argued that both uppercase and punctuation may convey valuable meaning, particularly for attention-grabbing headlines. In fact, Table 1 depicts how headlines containing no fully-capitalised words have a similar stance label distribution to all headlines, yet a fairly different distribution to those who do contain fully-capitalised words. Interestingly, unrelated, agreeing and disagreeing headlines are less likely to contain fully-capitalised words, whereas discussion headlines were more likely to contain fully-capitalised words.

Table 1: Normalised distribution of stance for all headlines, headlines containing no fully-capitalised words, and headlines containing fully-capitalised words.

Stance Label	All headlines	Headlines containing fully capitalised words	Headlines containing no fully capitalised words
unrelated	0.731310	0.726582	0.732232
discuss	0.178280	0.203899	0.173283
agree	0.073601	0.057013	0.076837
disagree	0.016809	0.012506	0.017649

4.2 Related/Unrelated Classification

4.2.1 Support Vector Machine Classifier

The confusion matrix for the SVM classifiers evaluated on the competition testing set using the TF-IDF and BERT features respectively are shown in Figures 1 and 2. These results demonstrate that, with the SVM, the BERT embeddings only out-perform the TF-IDF features by a small margin. Even then, their performance is about what would be expected from random guessing, as about 70% of the documents are unrelated (from Table 1), with the SVM models only correctly predicting unrelated about 70% of the time.

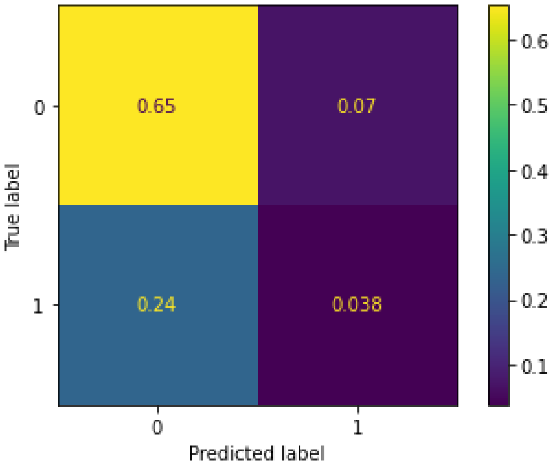

Figure 1: Confusion matrix for support vector machine using TF-IDF features.

4.2.2 Deep Classifier

Figures 3 and 4 show the confusion matrix and training/validation loss across the epochs. The overall score shows an impressive 97.96% accuracy when predicting whether headlines and bodies are related or unrelated. As it was evaluated on the unseen competition data, it is more likely that this is a valid result, and not from over-fitting. However, depending on the bases and source of the competition testing data, there may still be some over-fitting.

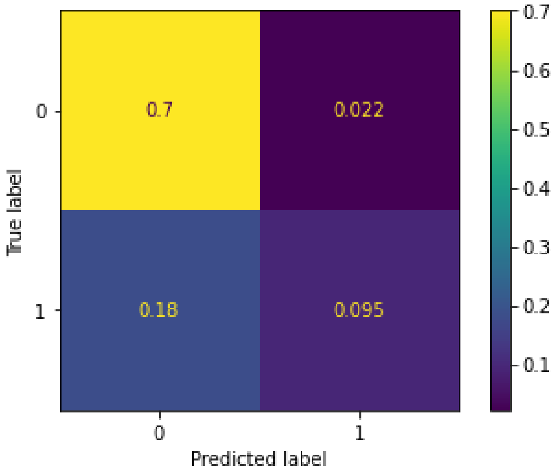

Figure 2: Confusion matrix for support vector machine using BERT sentence embeddings.

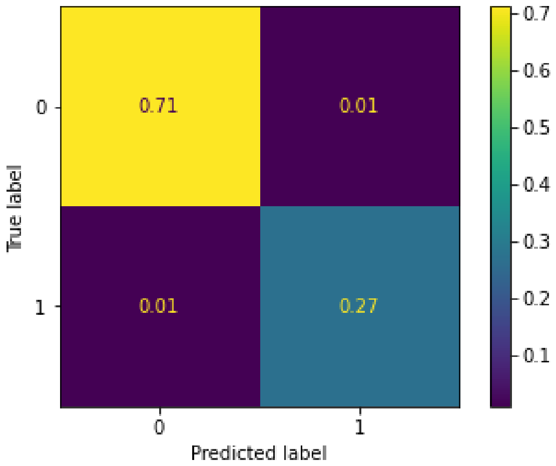

Figure 3: Confusion matrix for deep classifier using BERT sentence embeddings and TF-IDF features.

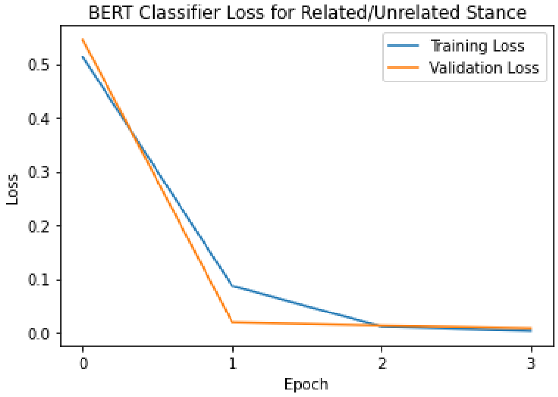

Figure 4: Training and validation loss for deep classifier using BERT sentence embeddings and TF-IDF features.

4.3 Agree/Disagree/Discuss Classification and Combined Classification

Unfortunately, due to time (and GPU memory) constraints, the implementation for the agree/disagree/discuss classifiers could not be completed and hence there are no results.

4.4 Potential Improvements

A potential improvement to the model would be to implement locality sensitive hashing (hashing similar tokens into the same buckets with high probability), thereby improving the performance on simpler models by giving tokens more semantic against each other when compared to simple one-hot encoded vectors.

Finally, due to the large imbalance in the data set, with less than 2% disagree, 7% agree and 18% discuss leaving 73% unrelated. The reasoning behind having lots of unrelated work is sound as its quite easy to build new samples into the dataset through randomly matching headlines and bodies, however, the imbalance in this case leads to potential over-fitting.

5 Ethical Implications

Word embeddings have been frequently shown to lead to bias in articles towards marginalised groups [8]. This is particularly worth being aware of for this application, since articles from different countries will naturally have some bias when talking about other countries, and disinformation has been known to target such groups in the past.

In much of ML and DL-based research, classification models can end up playing the role of a discriminator in a generative adversarial network (GAN). The models trained to detect fake news can assist those who deliberately spread disinformation in making it more believable, and potentially even implementing their own GAN using work like this as a discriminator to generate false headlines.

As BERT is trained on such large texts, including both fiction and non-fiction, it identifies patterns which lead to harmful stereotypes (such as sexism). On the brighter side, large open pre-trained models such as BERT benefit from being freely available for scrutiny, and their cross-application re-use helps offset the large amount of energy required to train such models.

6 Conclusion

In conclusion, modern transformers such as BERT encoders combined with conventional classification decoders provide a robust and effective way of classifying whether headlines are related to articles, but most models still struggle, achieving just better than random. Additionally, the deep learning architectures used out-performed the standard machine learning methods such as SVMs using TF-IDF.

References

[1] S. Zannettou, M. Sirivianos, J. Blackburn, and N. Kourtellis, “The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans,” *Journal of Data and Information Quality (JDIQ)*, vol. 11, no. 3, pp. 1–37, 2019.

[2] M. Potthast, T. Gollub, K. Komlossy, S. Schuster, M. Wiegmann, E. P. G. Fernandez, M. Hagen, and B. Stein, “Crowdsourcing a large corpus of clickbait on twitter,” in *Proceedings of the 27th international conference on computational linguistics*, pp. 1498–1507, 2018.

[3] E. Misback and C. Pfeifer, “Stance detection dataset for fnc-1.” `https://github.com/FakeNewsChallenge/fnc-1`, 2017.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[6] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books,” in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[7] W. Foundation, “Wikimedia downloads.”

[8] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, “Word embeddings quantify 100 years of gender and ethnic stereotypes,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 16, pp. E3635–E3644, 2018.