Chinese Fake News Detection : confirmation or invalidation ? Machine Learning for Natural Language Processing 2020

Quentin Mascart
ENSAE
quentin.mascart@ensae.fr

Raphaëlle Villers
ENSAE
raphaelle.villers@ensae.fr

Abstract

Social networks and Internet forums have become a source of information for many users. Given potential instrumentation of those media channels, automatic detection of fake news through NLP processing and modelling could offer scale-efficient regulation. Based on a Kaggle dataset, we perform fake news detection, with a few subtleties. Considering an incoming news B, we classify it into one of the following categories:

- agreed: B upholds the fake news in A
- disagreed: B refutes the fake news in A
- unrelated: B is unrelated to A

We use a sample of the available train dataset, given computation power constraints. We compare the classification outcomes for different approaches : classification with a linear SVM, neural network with a LSTM network and word-embedding with BERT. We will then draw conclusions and possible future work.

1 Problem Framing

The data consists of

- Three id columns (respectively **id, tid1, tid2** for each news pair, for fake news title 1 and fake news title 2)
- **title1_en** (resp. **title2_en**) - the fake news title 1 (resp. 2) in English.
- **label** - indicates the relation between the news pair: agreed/disagreed/unrelated.

Following data exploration, we perform task-specific modelling for Sequence classification, with different approaches. Our main metrics of focus for quantitative evaluation are the F1-Score and Balanced accuracy, upholded with qualitative evaluation.

2 Experiments Protocol

2.1 Pre-processing

The dataset is quite clean. Our pre-processing thus consists of removing stop words and tokenizing. Multi-word expression detection with Gensim allows us to perform faster processing[1].

Additional pre-processing for the linear SVM consists of vectorizing the corpus with *Count Vectorizer*. We keep 10,000 news pairs and split the data into 60% train, 20 % validation and 20% test samples.

For BERT, we keep 50,000 news pairs, 10% of which are assigned to our test sample.

2.2 Techniques and models and training

Our initial Linear SVM fails to classify news into the ”disagree” category. This is due to the data being imbalanced.

Table 1: Labels out of 10,000 subsample

Unrelated	6882
Disagreed	240
Agreed	2878

We adjust the model by over-weighting errors which wrongly classify ”disagreed” in the new cost function and thus reach better results.

We compare classification outcomes with those obtained with BERT as embedding technique.

2.3 Evaluation

For qualitative evaluation, we randomly select news titles X and Y from the dataset, pair them and look whether the model correctly classifies the combination. We do the same, this time taking both news from an identical randomly chosen pair X. On several testes examples, classification

[1] Google Colab Part II

is consistent with human deduction. For quantitative evaluation, our main metrics of interest are : F1-score (and ROC curves), Balanced accuracy. Accuracy wouldn't be an adequate evaluation metric as classes are imbalanced.

3 Results

The Linear SVM achieves the best performance among tested models, although the score for classification as ”Disagree” remains disappointing, despite adjustment, with F1-score falling to 0.19 for this specific class (Table 2). Note, however that this is to compare with a 0.10 F1-score for the Disagree class before adjustment of the cost function [2]. The adjustement does not imply a trade-off as it also yield an increase in other classes' F1-Score (from 0.77 to 0.84 and from 0.46 to 0.60 for Unrelated and Agree, respectively).

We note that our BERT model should be strengthened (or replaced with another embedding technique ?) as we only reach a à.57 balanced accuracy.

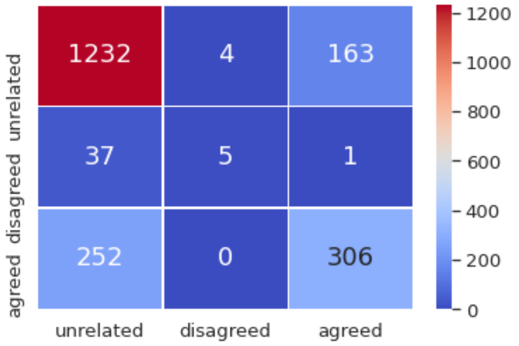

Figure 1: Classification outcomes with linear SVM

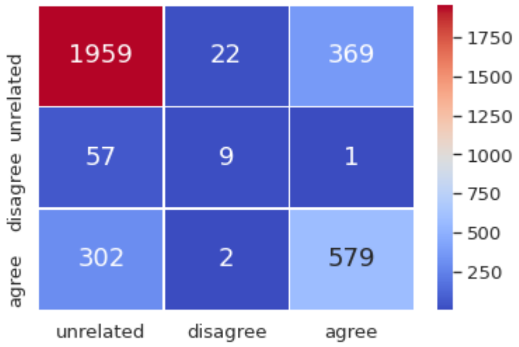

Figure 2: Classification outcomes with BERT

Table 2: Quantitative evaluation Linear SVM

	Precision	Recall	F1-score	Support
Unrelated	0.81	0.88	0.84	1399
Disagree	0.56	0.12	0.19	43
Agree	0.65	0.55	0.60	558
Macro avg	0.67	0.52	0.54	2000
Weighted avg	0.76	0.77	0.76	2000

Table 3: Quantitative evaluation BERT

	Precision	Recall	F1-score	Support
Unrelated	0.81	0.90	0.85	3471
Disagree	0.27	0.28	0.27	119
Agree	0.71	0.52	0.60	1410
Macro avg	0.60	0.57	0.58	5000
Weighted avg	0.77	0.78	0.77	5000

4 Discussion/Conclusion

In the data set, texts are available in Chinese as well as English. Note that the English titles are machine translated, so it is recommended to use the Chinese version titles to perform the analysis. As non-Chinese speakers, We would have not been able to perform qualitative evaluation, but it could be envisioned to strengthen our results by modelling over the original Chinese news articles.

We can also suppose we could achieve better performance with more computation power, allowing us to train over the whole available data set, instead of restricting to a subsample as we did.

We could prolong current work by comparing performance of the implemented SVM with a non-linear kernel SVM.

Finally, a further work axis would be to correct our LSTM model [3]. The network classifies *every* news into the ”unrelated” category. This probably has to do with the imbalance in our classes (the model learns to always return the majority class). Two equivalent options consist in either specifying class weights for the loss function or using class over-sampling (accomplished by WeightedRandomSampler in PyTorch).

[2] Google Colab Part III.4.a) vs IV.5.d)

[3] Google Colab IV