Exploring the effects of different n-gram models and embedding techniques in automatic fake news detection

Natural Language Processing - Project Group 16

Alessandro Cortese
University of Twente
a.cortese@student.utwente.nl

Gianmarco Lodi
University of Twente
g.lodi@student.utwente.nl

Introduction

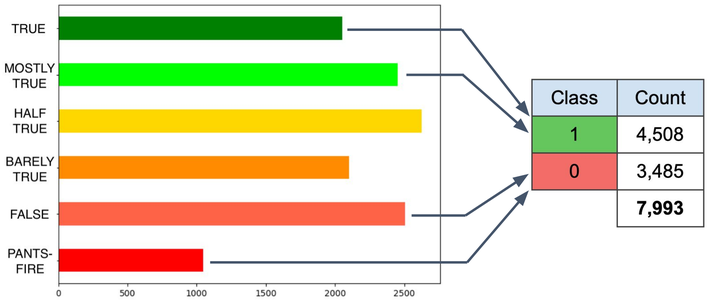
The aims of our project are:

- investigating how different NLP models and techniques would affect the accuracy scores of SVM and Random Forest models in automatic fake news detection for political statements;
- understanding whether textual features are enough to determine the truthfulness of a sentence.

After finding the best models we focused on exploring how the classification would improve by embedding other context information for each statement.

Data

The analysis was carried on the **LIAR dataset**, which contains 12,792 short statements manually labeled by the fact-checking website PolitiFact. In order to get a binary classification problem we dropped the two intermediate classes and relabeled the remaining with values 1 and 0.



Class	Count
1	4,508
0	3,485
	7,993

Each statement also comes with a series of *metadata*:

- the ***state*** where the sentence was pronounced.
- the ***topics*** of the sentence.
- the particular ***context*** in which it was said.
- the speaker's ***name***.
- the speaker's ***job***.
- the speaker's ***party affiliation***.
- the speaker's ***truth credit history***.

For each sentence we applied:

- case folding;
- stopwords and punctuation removal;
- tokenization and Porter stemming.

Methods

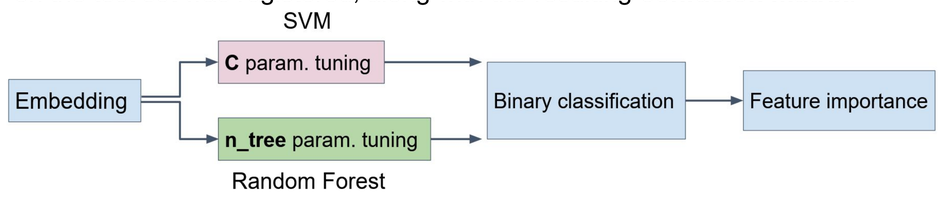
We used three techniques to turn the statements into feature vectors:

1. A **bag-of-words** model with ***tf-idf*** weighting.
2. A (weighted) average ***Word2Vec*** model.
3. A ***Doc2Vec*** model.

Different configurations are reported in Table 1, 2 and 3.
These embeddings were used as input for two binary classifiers:

- Support Vector Machine
- Random Forest

After performing **hyperparameter tuning** on the validation set, their **accuracy** on the test set was registered, along with the resulting **confusion matrix**.



The above procedure was then repeated by adding some of the sentences' **metadata** to the models with the highest test accuracy.

Results

- Overall, the best models were the bag-of words with mixed unigrams and bigrams and the Word2Vec trained on Google news.
- All the models **perform similarly**, i.e. their accuracy is around 0.59-0.62.
- The classifiers also perform rather similarly.
- Adding *subject, state_info, party_affiliation* and *context* brings a gain of **~2.5-4.5%** in accuracy (Table 4).

	SVM	RF
Doc2Vec	0.592	0.596

Table 3: Accuracy of Doc2Vec embeddings

TF-IDF	SVM	RF
Unigrams	0.612	0.616
Bigrams	0.604	0.587
Trigrams	0.593	0.585
Mixed Uni-Bi	**0.624**	**0.628**

Table 1: Accuracy of *tf-idf* n-grams

Word2Vec	SVM	RF
Google News	**0.631**	**0.602**
Simple Avg	0.585	0.590
Weighted Avg	0.575	0.575
Phrases Bigrams	0.578	0.576

Table 2: Accuracy of Word2Vec embeddings

- **Work in progress**: gaining insight using feature importance.

Adding Metadata	SVM	RF
Word2Vec	0.676	0.633
TFIDF	0.651	0.622

Table 4: Accuracy with metadata

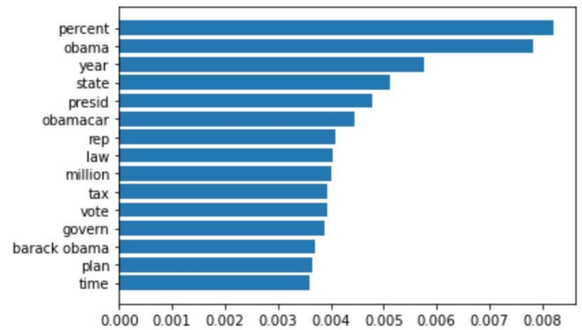


Table 5: RF feature importances for tf-idf mixed model

Conclusions

- Simpler bag-of-words models may perform even better than word embeddings.
- Textual features alone are not sufficient to obtain good results in our classification task.
- Adding metadata slightly improves accuracy.
- Fake news aren't always *pants on fire* false, but their deceit is often hidden in details.
- This makes fake news detection a difficult task, as human themselves have a hard time discerning them.
- More research is needed to develop AI systems able to better solve this task (e.g. probabilistic models, deep neural networks, using visual features, …)

Relevant literature

- Shu, Kai, et al. "Fake news detection on social media: A data mining perspective." *ACM SIGKDD explorations newsletter* 19.1 (2017): 22-36.
- Reddy, Prannay, et al. "A Study on Fake News Detection Using Naïve Bayes, SVM." *Neural Networks and LSTM. J Adv Res Dyn Control Syst* 1 (2019): 942-947.
- Pérez-Rosas, Verónica, et al. "Automatic detection of fake news." *arXiv preprint arXiv:1708.07104* (2017).
- Oshikawa, Ray, Jing Qian, and William Yang Wang. "A survey on natural language processing for fake news detection." *arXiv preprint arXiv:1811.00770* (2018).
- https://politifact.com

UNIVERSITY OF TWENTE.