Automatic pre-selection of potential misinformation for later expert evaluation

Author: Jakub Fedorko

Supervisor: doc. Ing. Jakub Šimko, PhD.

Thesis goals

- find, modify if need be and run available check-worthiness solutions to act as baseline methods
- design and implement an original solution to outperform said baseline methods.

Introduction

On social media, false information spreads faster and to more users than legitimate information. Combined with the fact that more than 60% of people acquire their news on social media, this creates a genuine threat to human society. Organizations like FactCheck.org, PolitiFact or Demagog try to mitigate the potential damage of false information by manually and comprehensively fact-checking societally relevant claims present in public space. However, the lengthy and demanding process of manual fact-checking cannot keep up with the easy unrestricted creation and spread of false information.

Figure 1: Standard automated fact-checking pipeline

Therefore, the need for an automatic fact-checking system arises, although it proves to be just as complex a problem as its manual equivalent. An automatic fact-checking pipeline usually consists of three main steps: detecting check-worthy parts of the text (i.e., sentences containing the claim that should be fact-checked), collecting information relevant to the check-worthy claim, and finally, using this information to infer the veracity of the check-worthy claim. This thesis focuses on the first part of the pipeline, that is choosing check-worthy sentences or as it is usually called check-worthiness task.

Our approach

Our solution consists of two independent pipelines, each based on a different approach to check-worthiness task. The first model is centred around the bidirectional LSTM layer, which uses tensors of BERT-based word embeddings as its sentence representation. The second model relies on a BERT-based sentence transformer as its embedding model, followed by a simple neural network.

We also extract text features such as sentence length, less and more granular POS tags and syntactic dependencies to help our sentence representations convey meaning. The features are encoded either using one-hot or sum encodings.

To acquire baseline methods for comparison with our models, we download and modify two check-worthiness solutions, which made their source code publicly available (Cheema et al. & Martinez et al.) to run with the data we use. We evaluate our models with average precision and f1-score on the positive class to conduct a fair comparison as baseline methods optimized the former metric and our models the latter.

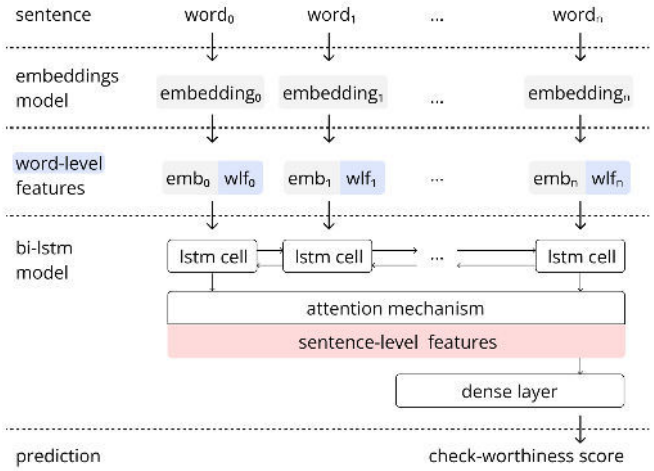

Figure 2: Bi-LSTM-based model (bi-lstm)

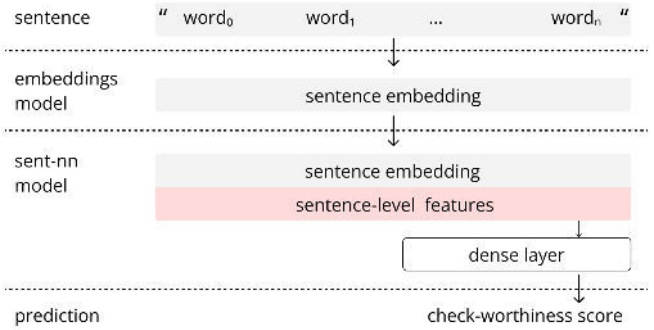

Figure 3: Sentence transformer-based mode (sent-nn)

Results & Conclusion

In this thesis, we tackled the problem of identifying the statements that potentially contain societally relevant and therefore check-worthy claims. We thoroughly analysed the existing solutions, taking inspiration from them in the process, to learn about the effectiveness of the various previously applied techniques and models. Two of the analysed works had their corresponding code publicly available, which allowed us to use them as our baseline methods. Inspired by the analysed works, we designed and implemented two different check-worthiness models, bi-lstm and sent-nn.

We tested their base performance and the effect of extracted text features on it. We found that our best performing model is the featureless sent-nn model. The model also outperformed both selected baseline methods. It gained more than 2.5% in average precision over Martinez et al.'s solution and more than 4.2% in f1-score on positive class over Cheema et al.'s. The best version of our bi-lstm was the one utilizing word-level text features. However, the model generally underperformed. The inclusion of the sentence-level features did not yield any performance improvement. Although both of our models could be improved upon, we consider the sent-nn model a contribution to the field as it considerably outperformed baseline methods using state of the art technologies.