”-Ernie, how do I look? -With your eyes, BERT” Sarcasm classification with BERT

Blazej Dolicki
Maastricht University
i6155906

1 Introduction

Sentiment analysis is a very common task in NLP. Contemporary models obtain brilliant result on this task on various datasets obtained from the internet such as IMDB movie reviews, Yelp reviews etc. However, especially in texts in the internet, people often use sarcasm, which by definition flips the polarity of the text. Therefore, if not detected, it can strongly decrease the quality of sentiment analysis. Sarcasm is very subtle and context-dependent. It is often difficult even for people to detect sarcasm especially in text forms, where ability to display emotions is limited. All these factors render sarcasm detection as both important and sophisticated task.

2 Prior literature

2.1 Sarcasm detection

In early years, attempts of sarcasm identification were based on linguistic analysis of this phenomenon. Sarcasm can be expressed in many diverse forms, for example one of the ”common forms of sarcasm on Twitter consists of a positive sentiment contrasted with a negative situation. Many sarcastic tweets include a positive sentiment, such as love or enjoy, followed by an expression that describes an undesirable activity or state (e.g., taking exams or being ignored).”(Riloff et al., 2013). Joshi et al., 2016 showed that word embeddings-based features also improve sarcasm detection.

A major direction of contemporary research is the context of the target text (Joshi et al., 2017). Even humans sometimes are not able to detect sarcasm without reading a post that a sarcastic comment refers to. This is an intuitively correct reasoning that I also would like to examine (Wallace et al., 2014, Bamman and Smith, 2015). Lastly, sarcasm can be easier predicted if it is known how sarcastic the author of the comment tends to be on average, so user-based approach was also investigated (Hazarika et al., 2018, Bamman and Smith, 2015).

2.2 Pretrained language models

In the last few years, using pretrained language models became a strong trend in NLP obtaining state-of-the-art results on multiple tasks. This approach adapts the idea of transfer learning NLP similar to Computer Vision. A great strength of these models is that they can be trained once and then fine-tuned to a wide variety of tasks. One of the earliest such models were OpenAI’s GPT (Radford et al., 2018) and ULMFiT (Howard and Ruder, 2018). These models make use of unidirectional architecture. Another important contribution was ELMo (Peters et al., 2017) which introduced deep contextualized word representations - the same word can have different vectors depending on its context. To show how important this is let us consider two sentences: ”Would you like a glass of wine?” and ”This building has a glass floor.”. In case of using standard word2vec (Mikolov et al., 2013) vectors (as oppose to ELMo), the word ”glass” in both sentences has the same vector. However, it is obvious that meaning of ”glass” in these two sentences is very different. Finally, at the end of 2018 Google’s BERT model was published. It introduced two novel significant enhancements: bidirectional Transformer jointly conditioned on left and right context and ”next sentence prediction task”(NSP) (Devlin et al., 2018). Thanks to the latter, BERT handles really well tasks were we want to predict something based on two related sentences/documents. Example of such task is Quora Question Pairs were the goal is to predict if two sentences are semantically similar. In Ablation studies section Devlin et al., 2018 prove that

NSP improves the accuracy of the model in all "sentence pair classification" tasks.

3 Data

Although there has been some work done in the field of sarcasm detection, it is not yet profoundly researched. To my best information there is no large, benchmark dataset. Majority of the datasets in the existing literature are not publicly available what limits the possibility of results comparison. There seem to be a disagreement in the field whether it is better to use self-annotated datasets (the author of a comment/review states himself if he is sarcastic with a tag or label) or corpora annotated by independent annotators. However, probably because human annotations require a lot of resources especially for large corpora, self-annotated datasets are commonly used. Only for Twitter datasets, Joshi et al., 2017 in their survey mention 17 papers using self-annotated data and only 4 papers experimenting with corpora with manual annotations.

Davidov et al., 2010 mention that in the case of Twitter, many sarcastic comments are not marked with "sarcasm" hashtag what makes the dataset noisy. Moreover, they claim that self-annotated comments are examples of exceptionally subtle form of sarcasm which are almost undetectable without context. However, nowadays most of sarcastic comments appear as a response to some other post and, as aforementioned, contemporary research in the field vastly focuses on context-dependent sarcasm detection. On the contrary, Khodak et al., 2017 thoroughly describe in their work what steps they made to reduce the described problems. Also they claim that Reddit is a much more reliable source than Twitter in this case - e.g. proportion of Reddit users adding the tags is much larger than proportion of Twitter users.

On the other hand, "prior work on sarcasm detection on Twitter found low agreement rates between annotators" (Bamman and Smith, 2015).

I have chosen to use SARC (Self-annotated Reddit Corpus) which has a massive total number of 533M comments (Khodak et al., 2017). However, for my experiments I would like to use a smaller subset of that data - only comments from the Politics subreddit. Another advantage of this dataset is that every comment is mapped to its ancestor, so context can be incorporated in the model.

3.1 Benchmarks

This work will be compared to 3 other papers that also used SARC. The first one is Khodak et al., 2017, the paper that introduced SARC, which performed only simple baseline models. Second is CASCADE model (Hazarika et al., 2018) which combined both context and user embeddings with the comments. For encapsulating user information, first stylometric representations are extracted using an unsupervised method ParagraphVectors (Le and Mikolov, 2014). Then a pretrained CNN is used to retrieve personality representations. Eventually, stylometric and personality features are combined using Canonical Correlation Analysis which results in user embeddings. Then discourse embeddings (which denote topics - e.g. sarcasm is more common in threads about politics than natural disasters) are created again using ParagraphVector. Then a text representation of the comment is computed which "captures both synactic and semantic information" (Hazarika et al., 2018). Eventually, the concatenation of the user, discourse and comment embeddings is the final vector on which predictions will be performed. Despite this architecture achieved better results than SARC (Khodak et al., 2017), it seems overcomplicated as mentioned by Kolchinski et al., 2018. This paper is the last benchmark, which only uses text representations from comments and user embeddings much simpler than those proposed by Hazarika et al., 2018. The authors propose two types of embeddings - first is crude Bayesan approach and second is created by training user embeddings using comment embeddings similar to creating user embeddings in collaborative filtering (although in this step, only user embeddings are trainable). Kolchinski and Potts, 2018 concludes that Bayesian prior is better when we have small numbers of comment per user, because in those cases the second representations strongly overfit, however, given more data, they are superior. In the overlapping datasets, Kolchinski and Potts, 2018 obtains the same or better performance thatn CASCADE.

4 Models

The following models have been evaluated on the SARC dataset:

- BERT as single sentence task without ancestors

	all balanced	politics balanced
Kolchinski et al. (2018) - Bayes	74.0	77.6
Kolchinski et al. (2018) - embeddings	75.3	75.1
CASCADE	77.0	75.0
Khodak et al. (2017)	75.8	76.5
Human (Average)	81.6	83.0
Human (Majority)	92.0	85.0
BERT without context	-	79.28
BERT with context - NSP	-	80.01

Table 1: Mean macro-averaged F1 scores comparison with benchmarks.

	BERT without context	BERT with context - NSP
Accuracy	79.28 $\pm$ .0063	80.01 $\pm$ .0092
Macro f1	79.28 $\pm$ .0063	80.01 $\pm$ .0092
Precision	0.7998 $\pm$ .007	0.81 $\pm$ .0101
Recall	0.7924 $\pm$ .0052	0.7947 $\pm$ .0093

Table 2: Different metrics for this work, the results are means of 5 runs.

- BERT as sentence pair task

My hypothesis is that the model with ancestor and response comment input as a pair of sentences should outperform previous work and the model containing only response sentence All models are trained with parameters proposed by Devlin et al., 2018 that is:

- batch size: 32
- epochs: 3
- learning rate: 2e-5
- warmup proportion: 0.1

The model without context was trained for approx. 20 minutes and the model with context was trained for approx. 40 minutes. For computations I used free GPU on Google Colab.

5 Results

The model with incorporated context via NSP indeed significantly outperforms BERT without context and best previous work by 2.4% (Figure 1). Detailed metrics are presented in Figure 2.After manual inspection of misclassified examples I noticed two types of errors: the text of at least on of the two comments is very short (3 words or less) or both comments seem sarcastic.

6 Future work

Our model achieved better results than prior work without any user representations, therefore an obvious direction is to incorporate also that information. Moreover, it is important to find a solution for classifying short comments which are currently problematic.

References

David Bamman and Noah A Smith. 2015. Contextualized sarcasm detection on twitter. In *Ninth International AAAI Conference on Web and Social Media*.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. Cascade: Contextual sarcasm detection in online discussion forums. *arXiv preprint arXiv:1805.06413*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):73.

Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark James Carman. 2016. Are word embedding-based features useful for sarcasm detection? *CoRR*, abs/1610.00883.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2017. A large self-annotated corpus for sarcasm. *arXiv preprint arXiv:1704.05579*.

Y Alex Kolchinski and Christopher Potts. 2018. Representing social media users for sarcasm detection. *arXiv preprint arXiv:1808.08470*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714.

Byron C Wallace, Laura Kertz, Eugene Charniak, et al. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 512–516.