Opinions and Facts Recognition in News Articles

Sergei Loptev
Faculty of Computer Science
National Research University Higher School of Economics
Moscow, Russia
seloptev@edu.hse.ru

***Abstract*—With the expansion of controversial topics covered in social media, it is increasingly difficult for readers to retrieve factual information from news articles containing many opinionated statements. This problem is also commonly known as the subjectivity detection task — distinguishing opinionated or subjective sentences from unopinionated or objective ones. This paper provides an end-to-end tool for subjectivity detection in news articles. First, we collect a modern dataset of news articles annotated for subjectivity by crowd-sourced workers. Then, we fine-tune a BERT model, gaining accuracy close to the state-of-the-art model on the subjectivity detection task. We also use advanced fine-tuning techniques, such as one cycle policy. Finally, we create a convenient command-line interface (CLI) for the automated annotation of news articles with the fine-tuned model. The implications of this work extend beyond the realm of news media, as accurate recognition of opinions and facts has important implications for fields such as information retrieval and sentiment analysis.**

***Index Terms*—natural language process ing, subjectivity detection, crowdsourcing, human-annotated data collection, news processing**

I. Introduction

This paper covers the problem of distinguishing facts and opinions from news articles. The task is also known as the subjectivity detection problem. It is an essential sub-task of sentiment analysis since sentiment analysis models perform better when operating on only opinionated corpora.

Nowadays, many active news agencies, including BBC, The Guardian and others, publish tons of materials daily. The number of controversial topics is also rising today: the recent ones include ChatGPT and Bing AI, COVID-19, the search engine race between Microsoft and Google, and others. With this amount of disputed news, it is difficult for an ordinary reader to successfully extract facts to have an unbiased view.

This paper covers the creation of an end-to-end tool for distinguishing factual and opinionated sentences from news articles. Effectively, the tool allows a user to download the text of a news article and annotate each sentence with a fact or opinion label.

We are going to achieve this in several steps as follows. First, we will create a dataset of sentences from modern news articles. This dataset will be annotated for subjectivity by humans. Wiebe, J. et al. created a similar dataset called MPQA in [1]. Their study is the main base study for creating the modern news dataset. The sentence-wise annotations for this dataset were created in 1999 in [2] and are redundant now. Second, we will fine-tune a BERT model for subjectivity detection. Pang, B. et al. examined various fine-tuning techniques for BERT in [3]. We are going to re-use and verify the findings of their study. They achieved an accuracy of around 90%, so we will strive to achieve similar results. Third, we will create a brand new tool that given a link to a news article, will scrape its contents, annotate the text for subjectivity, and publish the annotated article. There are currently no competitor tools; the closest one is Grammarly, which can help detect too opinionated sentences and make a text look solid.

We expect our new dataset to be as large as MPQA. Our fine-tuned model should be as precise as the one in [3], i.e. we expect the accuracy to be around 90%. We also anticipate the tool to be convenient to use and precise enough.

The structure of this paper is as follows. Section 2 covers papers related to our study, i.e. the work covering the creation of both datasets and models. In section 3, we describe our methodology for all task steps. In section 4, we describe the means of evaluation and anticipated results in detail. Section 5 contains the conclusions of the paper.

II. Related work

Reasonable amount of research was made in the area of subjectivity detection. There are several different datasets designed for this task in several domains. In [4], the SUBJ dataset was introduced. It consists of 5000 subjective and 5000 objective sentences, sourced from 1346 human-annotated web-pages. The sentences are labeled with respect to subjectivity (subjective/objective) and polarity (positive/negative). In [5], a dataset containing POV-tagged statements from Wikipedia is created. POV-tagged statements are the statements that are reported to be violating the NPOV (neutral point of view) principle of Wikipedia. These statements are then annotated by humans, and eventually the dataset consists of 1843 biased statements, 3109 neutral ones, 1843 neutral ones from featured articles in Wikipedia and 1843 neutral ones from featured articles equipped with same type-balanced distribution in biased statements. In [1], the MPQA dataset was introduced. It contains sentences from 535 Spanish news articles, translated to English. In total, there are around 9700 sentences, and around 55% of them are subjective. All these dataset are relevant to our task, but only MPQA satisfies our domain, which is news articles. However, the MPQA dataset was created around 20 years ago. We believe that in 20 years the writing style of news agencies has changed; moreover, new topics like COVID-19 have been introduced. Some new agencies were created, and some old ones went bankrupt. Due to all these reasons we have decided to create a more modern dataset that will generally be similar to MPQA, but will consist of modern data.

Apart from creating datasets, a large amount of work was dedicated to training models for subjectivity detection. In [6], the long progress of such models is shown. The first methods were using hand-crafted features, including keyword spotting. Then researchers moved on to ontology models, i.e. defining a set of ontologies that define relationships between different classes of words and projects them into a vector space. Next were statistical methods, which were similar to classical machine learning — models trained on an annotated dataset. One example of such a model is a Passive Aggressive classifier, mentioned in [7]. The authors of this study claim that their method achieved the F1-score of around 85% on cross-validation. This shows that early approaches to the task were already quite successful. Next models were Latent Dirichlet Models, which used word frequency to compute posterior distribution for the task. All of the above methods were syntactic, i.e. they tried to use syntax as features rather than using words' meanings. The next methods, on the other hand, do use the semantic meanings. To start with, there are semantic sentence models, that try to model the probability of a word given the words before it. Then, there are parse trees, the methods that try to model sentences' probabilities by decomposing the representations into matrix multiplications and using recurrent neural networks (RNNs) on top of it. Next, there are convolutional models (CNNs) for subjectivity detection, where CNNs serve as a backbone, and then the answer is retrieved by an RNN or other methods. One of them is described in [8] — they use a bayesian network-based extreme learning machine (BNELM) on top of a CNN to achieve an accuracy of 89% on TASS 2015 — a dataset of Spanish tweets for subjectivity detection. Finally, the most recent models in subjectivity detection are the ones that are based on the transformer architecture, like BERT. One of these studies is [3], where the authors achieve accuracies from 84% to 95% on different datasets. They also try different fine-tuning techniques to enhance the quality, and find out that one cycle policy and multi-task learning are the most helpful ones. We are going to repeat this experiment in our work.

As mentioned before, there are currently no competitors for subjectivity detection tools. There are writing aids for news, like Headline Analyzer, Ahrefs and Grammarly, but there are no tools for automated analysis of the news.

III. Methodology

In this section, we will describe our methodology in detail.

A. Data collection

We aim to create a dataset containing sentences from modern newspapers annotated by humans. We will mine the data using the Python library called newspaper3k[1]. This library provides a convenient way to download raw data from news agency websites. Next, we will filter the dataset, removing or unifying the sentences that are too short and dealing with citations. For example, a typical case would be a whole sentence being inside a citation — in this case, it can be more challenging to annotate the sample correctly. Then we will use the well-known crowdsourcing service called Yandex.Toloka. This service is widely used throughout the world for the task of annotating datasets. Our annotation schema will likely resemble the one in [2]. There the schema is as follows. Annotators need to rate the subjectivity of the sentence on a scale from 0 to 4, provided the following instruction: ”If the primary intention of a sentence is objective presentation of material that is factual to the reporter, the sentence is objective. Otherwise, the sentence is subjective”. After the annotation, modern statistical tools will be used to extract the ground truth.

B. Model training

We will broadly consult with [3] in this part of our work. We will take a pre-trained BERT model, e.g. from HuggingFace[2], and fine-tune it for our specific downstream task. As per fine-tuning techniques, we will repeat the ones from [3], verifying that they also work on the news articles domain. These techniques are one-cycle policy and multi-task learning. One cycle policy is a way to schedule learning rate fluctuations: it includes a warm-up (i.e., rising learning rate) for the initial part of the training and then a cool-down (i.e., lowering learning rate) for the final part. Multi-task learning involves jointly fine-tuning the model for different tasks, and it is believed to improve the models' robustness and performance. Examples of such tasks are text similarity and pair-wise relevance ranking tasks.

C. Creation of a CLI tool

In this part of our work, we will create a convenient CLI tool allowing users to parse news articles online and automatically annotate them for subjectivity. Since it fits our case nicely, the newspaper3k library will power the part of downloading and parsing the articles. Then, we will split the article into sentences, just like for the model training. After that, we will sequentially apply our fine-tuned model to each sentence to predict whether it is subjective or objective. Finally, we will mark the objective and subjective pieces with the respective styles (e.g., text colours) and publish the annotated paper.

IV. Evaluation

We will evaluate all three parts of our work separately. There are not many means of evaluating datasets; however, we aim to gain a dataset of not less than 10000 samples in total; since both objective and subjective sentences are present in large quantities, we will aim to get a close to equal proportion of these samples.

We will also evaluate the fine-tuned model. For this, we will measure the accuracy on the MPQA corpus and our newly gathered dataset. We anticipate achieving the accuracies of not

[1] https://newspaper.readthedocs.io/en/latest/

[2] https://huggingface.co/bert-base-cased

less than 90%, which would be consistent with the general accuracy of modern subjectivity detection models.

As per the CLI tool, it is harder to measure its quality because convenience cannot be measured quantitatively. Instead, we will evaluate it by conducting a user experience survey. We will gain feedback on the tool from several users and adjust it according to their feedback.

V. Conclusion

This paper described the project proposal for creating a new end-to-end solution for opinions and facts recognition in news articles. So far, all of the preparation work has been done. We have analysed the research area thoroughly and outlined techniques that can be helpful to us. The most helpful ones are annotation schemas from the MPQA dataset and the fine-tuning techniques from [3]. We also properly planned the future work, covering all three parts: data collection, model training and creation of the CLI tool. Generally, we will reuse a lot of already achieved findings but add our research. Then, we planned the means of evaluation and described the desired results. The results that we anticipate are adequate: they are not overly optimistic, but they are still consistent with the current performance in the area. The following steps are precisely to perform all of the actions that we planned. We are confident this project will significantly impact news analysis, and the tool will be convenient enough to share publicly.

References

[1] J. Wiebe and E. Riloff, "Creating subjective and objective sentence classifiers from unannotated text," in *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, 2005, pp. 486–497.

[2] J. Wiebe, R. Bruce, and T. O'Hara, "Development and use of a gold standard data set for subjectivity classifications," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, 1999.

[3] H. Huo and M. Iwaihara, "Utilizing bert pretrained models with various fine-tune methods for subjectivity detection," in *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer, 2020, pp. 270–284.

[4] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of ACL 2004*, 2004.

[5] C. Hube and B. Fetahu, "Neural based statement classification for biased language." in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 2019, pp. 195–203.

[6] I. Chaturvedi, E. Cambria, R. Welsch, and F. Herrera, "Distinguishing between facts and opinions for sentiment analysis: survey and challenges," *Information Fusion*, vol. 44, pp. 65–77, 2018.

[7] A. Stepinski and V. Mittal, "A fact/opinion classifier for news articles," in *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR)*. ACM, 2007, pp. 807–808.

[8] I. Chaturvedi, E. Ragusa, P. Gastaldo, R. Zunino, and E. Cambria, "Bayesian network based extreme learning machine for subjectivity detection," *Journal of The Franklin Institute*, vol. 355, pp. 1780–1797, 2018.

Word Count: 1874