Structure matters: Feature selection for transformer-based news and opinion classification

Laura Nixon
Final Paper
CS224U: Natural Language Understanding
Stanford Online
laurajnixon@gmail.com

Abstract

The ability to detect whether a news article is a straight news story or an opinion piece is a valuable task for media research. News articles are structured pieces of writing, but past approaches to news vs. opinion classification have largely ignored that structure. I hypothesize that features that incorporate the parts of articles most likely to aid in news vs. opinion differentiation will improve classifier performance. To explore this question, I use labeled data drawn from a large, proprietary dataset of English-language news articles. I engineer contextual representation features that draw on different parts of a news article (start, end and metadata) and use them to fine tune a BERT model. I find that features that incorporate article metadata and the end of the article result in better performance than features that rely only on the start of the article.

1 Introduction

News articles generally fall into two categories: 1) straight news articles or 2) opinion pieces. Straight news articles are intended to impart information and report on factual events, while opinion pieces (including editorials, letters to the editor, columns and op-eds) present an opinion, and aim to persuade readers to adopt a particular stance.

Straight news articles are typically written by reporters. On the other hand, opinion pieces are written by specialized staff (i.e. columnists and editorial writers in the case of columns and editorials), or by members of the community (in the case of op-eds and letters to the editor), and tend to differ from straight news articles in style, length, and format.

Past work on classifying news articles into news vs. opinion has focused on creating features based on the entire full text of the article, or the beginning of the article text. However, news articles and opinion pieces are structured pieces of writing, generally containing a headline, byline (i.e. author), introduction, body, and closing. As a result, past approaches like extracting contextual representations from the first paragraph or so (e.g. Alhindi et al. (2020)) or counting linguistic features present anywhere in the article (e.g. Krüger et al. (2017)) may exclude potentially valuable information for this task.

Furthermore, most news datasets also contain the name of the news outlet that published the article. Particularly for multi-publisher datasets, this metadata may also provide valuable clues to aid a classifier. For example, newswires such as the Associated Press or Reuters may be less likely to publish opinion pieces.

I hypothesize that, for transformer-based models using contextual representations as features:

Incorporating metadata will improve performance: Models that utilize article metadata (i.e. headline, byline and publisher) and the article body for feature creation will perform better than models that only use the article body, even if the sequence length is constant.

Metadata alone will not be sufficient: Models that only use article metadata in feature creation will not perform as well as models that only use the article body, or that use both the article body and metadata.

Incorporating the end of the article will improve performance: Models that incorporate information from the beginning and end of the article in feature creation will perform better than models that only use the beginning of the article, holding sequence length constant.

2 Related Work

2.1 News vs. Opinion Classification

Hand-built features: One strand of earlier work on news vs. opinion classification focused on hand-built, frequency-based features. Krüger et al. (2017) used part-of-speech tagging and and other text parsing techniques to create individual, hand-built features (for example, the frequency with which first person pronouns appear) for English-language news classification. The general approach introduced in Krüger et al. (2017) has been replicated for news in Italian (Totis and Stede) and French (Escouflaire, 2022).

Transformer-based features: With the advent of transformer-based large language models, several researchers have attempted to leverage these models for the news vs. opinion classification task. In particular, Alhindi et al. (2020) built on the work of Krüger et al. (2017) by reimplementing their linguistic features approach, and comparing it to the performance of transformer-based models.

For this, they used two types of features: static contextual representations, and what they call 'argumentation' features. For their baseline model, they fine-tuned a 'bert-base-uncased' model on their data collections, used it to obtain contextualized embeddings for the first 512 tokens of each article, and then used these embeddings to fine-tune a BERT classifier (Devlin et al., 2019). For their argumentation features, they used a sentence-level argumentation corpus (Al Khatib et al., 2016) and used these to fine-tune a BERT model. They used this model to predict an argumentation type for each sentence in their corpus, and then used the predictions to generate features for support vector machine (SVM) and recurrent neural network (RNN) models.

They found that transformer-based models outperformed the re-implementation of Krüger et al. (2017)'s linguistic features model, and that the approaches they tried to combine the hand-built linguistic features with transformer-based features did not seem to improve performance.

Bouter (2022) also experimented with BERT-based models for the news vs. opinion classification task, but for Dutch-language news. Bouter compared Dutch BERT models (BERTje, De Vries et al. (2019)) using contextual representation features to models using frequency-based bag-of-words features. They experimented with various sequence lengths for their input data and training set sizes, and found that reducing the sequence length hurt the performance of the model more than reducing the number of examples the model was trained on. They found that even with limited training set sizes and/or sequence lengths, BERTje models achieved results that were comparable to the best results obtained by bag-of-words models.

2.2 Feature Selection.

Article body: Previous approaches have used two main methods for selecting text inputs for features. For hand-built features such as the ones used by Krüger et al. (2017), researchers have typically searched across the entire full text of the article. For transformer-based models (i.e. Alhindi et al. (2020) and Bouter (2022)), researchers typically use the beginning of the full text, truncated according to the maximum length of their input sequence. The maximum input sequence length for a BERT model is 512 WordPiece tokens, which corresponds to roughly the first couple of paragraphs of the article.

Researchers working on related news classification tasks such as topic classification have experimented with using different types of input text for BERT-based models. Working on hyper-partisan news detection, for example, Naredla and Adedoyin (2022) experimented with different input sequence lengths for their BERT model. They found that shorter sequence lengths (150-400 tokens) actually performed slightly better than 512-token sequences. They also noted that previous researchers had experimented with inputting different parts of the news article into BERT.

Among these previous researchers was Sun et al. (2019). They tried out a number of different text input schemes for BERT fine-tuning. They experimented with several different datasets, including a Chinese-language news article dataset for topic classification. For that task, they found that using the beginning and end of the article worked better than head-only or tail-only truncation. That method also outperformed hierarchical methods involving dividing up the articles into chunks, obtaining representations for each chunk, and then pooling them in some way.

This may be because the ends of news and opinion articles tend to follow distinct patterns. Opinion texts often close by stating or restating an opinion, where news stories usually close with either a quotation or statement of fact. As a result, the end of a news article may be quite valuable for news vs. opinion predictions, an area that previous research has not explored.

Metadata: Previous approaches have also not made use of the metadata that typically accompanies news articles (i.e. headline, byline and publisher). This metadata often contains clues about whether an article is news or opinion. On the contrary, researchers have often attempted to exclude article metadata from their training and test data. Bouter (2022), for example, attempted to remove any mentions of authors or publishers from their dataset, and Alhindi et al. (2020) removed words like 'editorial' that would be strong clues about the genre of the article.

Understanding the predictive value of contextual representations of metadata for news vs. opinion classification may allow for performance gains when metadata and full text data are combined. Furthermore, in some media datasets, only the metadata is available. If metadata alone is sufficiently predictive, it may allow for automated news vs. opinion classification on these datasets.

3 Data

For this analysis, I use a proprietary dataset of 308,400 news articles, published in U.S. news outlets from 2010-2022. After performing deduplication and removing articles that are missing news vs. opinion labels or the full text field, there are 238,885 articles (220,000 straight news stories and 95,000 opinion pieces).

The articles were obtained from the Factiva and Lexis Nexis databases, and, for certain outlets, through web scraping. Each record contains the full text of the article, metadata provided by the publisher (including the headline, the author of the piece, and the name of the publisher), and hand-annotated metadata, including whether the article is a straight news article or opinion piece. The annotation was performed by professional annotators.

I randomly selected 5,000 straight news articles and 5,000 opinion articles to create a training set of 10,000 articles, balanced across the two classes. I then created a randomly selected, balanced test set of 2,000 articles (1,000 news articles and 1,000 opinion articles). This is slightly larger than the training and test sets used by Alhindi et al. (2020) (a balanced training set of 6,386 articles, and two balanced test sets of 706 articles and 386 articles).

4 Models

The baseline model for my experiments is described in Alhindi et al. (2020): a *'bert-base-cased'* model (Devlin et al., 2019) implemented from the HuggingFace transformers library (Wolf et al., 2019), trained for 3 epochs, with a maximum sequence length of 512, learning rate of 2e-5, and batch size of 16. Its features are BERT-based contextualized representations of the first 512 tokens of each article. My experiments focus on how changing the features of the model impacts performance, rather than on changing the model architecture itself.

5 Experiments

Article start: For my first experiment, I attempt to replicate the BERT model that Alhindi et al. (2020) used as their baseline, with parameters as described in the previous section. For this model, I used the default right-side truncation parameter, with padding, extracting the first 512 tokens from the start of the body of the article. Alhindi et al. (2020) reported macro-average F1 scores of 0.93 and 0.89 on their two test sets for this set-up. On my test set, the macro-average F1 score (hereafter 'F1 score') was 0.90.

Article ending: I then reversed the truncation scheme to left-sided truncation, in order to extract the last 512 tokens in the article body. Surprisingly, this model performed slightly better on the test set, with an F1 score of 0.91.

Metadata: To create the metadata input text, I concatenated the headline, author name, and publication name fields, each separated by the following string: '*****' and two newline characters. Using BERT embeddings of the resulting string as a feature, the model achieved an F1 score of 0.89.

I also tried using only contextual representations of the headline field as a feature for the model. This reduced the performance of the model, resulting in an F1 score of 0.85 on the test set.

Combinations: The best results were obtained using combinations of the above input features. Combining the metadata with the start or end of the article both achieved F1 scores of 0.91.

I also tried combining the start and end of the article, while keeping the input sequence length constant at 512 tokens. To do this, I made a conservative estimate of the number of characters in-

Model input	Max sequence length	Macro-average F1 score
Body (Start)	512 tokens	0.903
Body (End)	512 tokens	0.908
Metadata	128 tokens	0.885
Headline	64 tokens	0.854
Body (Start) + Metadata	512 tokens	0.913
Body (End) + Metadata	512 tokens	0.914
Body (Start + End)	512 tokens	0.917
Body (Start + End) + Metadata	512 tokens	**0.926**

Table 1: Experiment results

cluded in a typical sequence of 512 WordPiece tokens (2,000 characters). For articles with fewer than 2,000 characters, I used the entire article as the feature input. For articles with 2,000 characters or more, I extracted the first 1,000 characters and last 1,000 characters of the articles, and concatenated them, with a separator string ('*****' plus two newline characters).

Combining the start and end of the article in this way resulted in an F1 score of 0.92. Finally, I combined the metadata, start and end of the article, which resulted in the highest F1 score of all the experiments, at 0.93.

6 Analysis

The experiments above provide support for the hypothesis that adding metadata to feature inputs improves performance on the news vs. opinion classification task, even with the maximum sequence length held constant. This suggests that the model gains more information from the metadata than it loses from the loss of the sentence or two of the full text that is truncated to make room for the metadata in the input sequence.

However, as I hypothesized, using metadata alone as a basis for contextual representation features seems to result in decreased performance as compared with using the start of the article body as a feature input. This is particularly true when the metadata made available to the model is limited to the headline.

Finally, the above experiments also provide support for the hypothesis that incorporating the end of articles into input features improves news vs. opinion classification performance. I was surprised to find that training the model on the end of the article body actually resulted in slightly better performance than training on the beginning of the article body. Furthermore, the two inputs seem to contribute uniquely valuable information to the model, since performance improves slightly when both are incorporated into the feature input.

7 Conclusion

In this paper, I studied how using text from different parts of a news article as the basis for contextual representation features affected the performance of a news versus opinion classifier. While past approaches have relied solely on the beginning of the article body, I showed that adding text from the end of the article body and/or the article metadata improved the performance of the classifier.

Overall, this suggests that close attention to the structure of news articles, and how that structure applies to the task at hand, can be quite valuable in designing features for news classification. Future research could explore other methods for leveraging the unique structure of news articles, such as using sentence-level embeddings, or paraphrasing parts of the article and using the resulting summaries as feature inputs.

In addition, the broader idea that different parts of news articles are valuable for different types of machine learning tasks could have implications for a wider range of news classification issues, such as topic classification, the identification of misinformation, or the identification of hyper-partisan news articles.

Acknowledgments

I would like to thank Raul Padua for his guidance and feedback in developing this paper.

Authorship Statement

This paper is the result of the individual work of the author, Laura Nixon, for the Natural Language Understanding course (XCS224U) running from

January to April 2023, as part of the Stanford Online Artificial Intelligence Professional Program.

References

Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Mattias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, page 3433–3443.

Tariq Alhindi, Smaranda Muresan, and Daniel Preoţiuc-Pietro. 2020. fact vs. opinion: The role of argumentation features in news classification. In *Proceedings of the 28th international conference on computational linguistics*, page 6139–6149.

David Bouter. 2022. *Knowing the difference between news and opinion: An explorative research project into classifying news and opinion*. Leiden Institute of Advanced Computer Science, Bachelor Thesis.

Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. ArXiv preprint arXiv:1912.09582.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 4171–4186.

Louis Escouflaire. 2022. Identification des indicateurs linguistiques de la subjectivité les plus efficaces pour la classification d'articles de presse en français.(identifying the most efficient linguistic features of subjectivity for french-speaking press articles classification. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2: 24e Rencontres Etudiants Chercheurs en Informatique pour le TAL (RECITAL*, page 69–82.

Katarina Krüger, Anna Lukowiak, Jonathan Sonntag, Saskia Warzecha, and Manfred Stede. 2017. Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering*, 23(5):687–707.

Navakanth Reddy Naredla and Festus Fatai Adedoyin. 2022. Detection of hyperpartisan news articles using natural language processing techniques. *International Journal of Information Management Data Insights*, 2(1):100064.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019*, volume Proceedings 18, page 194–206, Kunming, China. Springer International Publishing.

P. Totis and M. Stede. Classifying italian newspaper text: news or editorial? *Computational Linguistics CLiC-it*, page 372.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, R.'emi Louf Tim Rault, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*.