Evaluating Existing Models for Sentiments Analysis of News Corpora

Yunzhe Sun
sy2825@nyu.edu

Haohai Pang
hp1397@nyu.edu

Mingxuan Wu
mxw210@nyu.edu

Abstract

When dealing with domain-specific tasks, using domain-specific data in pre-training has proven to be useful. Improved BERT models for finance, chemistry and biology already exist, and they significantly outperform base BERT in these fields. However, a model specialized in understanding news is lacking. To achieve SOTA in news corpora, the first step is to find the base model best suited for it. This work evaluates three base models for sentiment analysis of sentence-level news and document-level news. Results show that the fine-tuned DeBERTa and Big Bird model achieved the highest F1 score on sentence-level and document-level respectively. We further explore the underlying reasons and provide empirical advice on which features tend to be more useful than others in understanding news. The code is available at: https://github.com/Richard1007/sentiment_news

1 Introduction

Sentiment Analysis (SA) is one of the most commonly applied downstream tasks in NLP. With the concept of Transformer proposed by Google in 2017 (Vaswani et al., 2017), many models with different Transformer architectures, such as BERT (Devlin et al., 2019) and GPT (Brown et al., 2020), claimed to have achieved SOTA in sentiment analysis. Although the statistics in papers showed that these models can achieve high accuracy in the field of general SA, it is desirable to have a model that is particularly well-performed when dealing with SA problems in a certain field.

To better handle domain-specific tasks, researchers modified existing models by conducting fine-tuning processes with domain-specific data. Compared to the original BERT model, BioBERT (Lee et al., 2019) and FinBERT (Araci, 2019) are more effective and efficient when dealing with biomedical and financial corpora, respectively. However, a model specifically designed to better understand news (e.g. NewsBERT) is lacking. This can cause difficulties for those trying to apply NLP-related techniques to journalism. In order to achieve SOTA in understanding news, we first need to decide which base model to use.

Since it is impossible to test every existing base model, we carefully chose three models as candidates. Based on the evaluation loss, accuracy, and F1 scores, we find the best model in understanding sentence-level news and document-level news. Considering the different Transformer architectures, different mechanisms, and different pre-training tasks between the candidates, we further explore reasons behind it and derive which features are more significant for understanding news. We hope this work can be served as an empirical guidance for future models specifically designed or modified to understand news corpora.

2 Related Works

This section presents the details of the three models we evaluate and explains why they can be considered as representatives of the three methods. According to our experimental results, we will later analyze which method performs the best in understanding news.

2.1 Deep Bidirectional Method

DeBERTa v3 (He et al., 2021) uses a similar structure as BERT, but it made two vital improvements. First, it replaces the original MLM pre-training task with Replaced Token Detection (RTD), which aims to detect whether a token was replaced in a given sequence. This task greatly helps the model to distinguish between synonyms by augmenting differences between similar words. Secondly, it keeps feeding relative positionings instead of absolute positionings in the Transformer, which is more reasonable considering the features before/after any token compared to a specific position of the sequence.

We expect this model to understand news well, as RTD may be very useful for understanding subtle synonyms in document-level news stories. Also, relative positionings may be instrumental for censoring sentence-level news because it is crucial to get features next to each word in short sentences. Since (He et al., 2021) points out that DeBERTa v3 outperformed BERT and RoBERTa on SST-2 binary sentiment classification task. We believe it is safe to regard it as a representative that uses deep bidirectional pre-training tasks to understand text.

2.2 Longer Sequence Method

Big Bird attempts to reconcile the disadvantages of limited input length and high computational complexity caused by the "full attention mechanism" by proposing a "sparse attention mechanism" (Zaheer et al., 2020). It theoretically reduces the computational complexity of achieving full tokenization from $O(n^2)$ to $O(n)$ by combining three approaches: Random attention, Window attention, and Global attention (Zaheer et al., 2020). Specifically, it connects some random tokens with other random tokens, connects each token with its neighbors, and connects some significant tokens (e.g. [CLS], [SEP]) with all others. Its unique structure guarantees that each token can be reached by any other in at most two steps (e.g. A–[CLS]–B), so we will not lose much information compared to the "full attention mechanism".

Based on the fact that some news articles may exceed five hundred words, we are interested in exploring whether the ability to process longer sequence is more important in news corpora. Since Big Bird greatly increases the Maximum sequence length from 512 to 4096 while keeping the running time tolerable, we may regard it as a model specialized in longer sequence method.

2.3 Tokenization-free Method

CANINE claims to be the first model that gets rid of the explicit word-tokenization step. Alternatively, inputs to this model are sequences of Unicode characters converted from letters (Clark et al., 2022). CANINE then uses a hashing strategy to represent these Unicode characters. Its structure consists of three parts: down-sampling functions, a primary encoder and up-sampling functions. The core encoder executes directly on Unicode characters after a number of hash functions. One of the great advantages of Tokenization-free methods is that they can alleviate the potential mismatches between tokenizations and the vocabulary (Clark et al., 2022). Moreover, CANINE's character-level approach makes it " have access to a far larger amount of (unsupervised) data to learn word composition" (Clark et al., 2022).

Although it might be meaningless to split words into letters and represent letters by Unicode characters, CANINE surprisingly surpasses BERT in some downstream tasks. More importantly, it provides an opportunity to directly compare traditional Transformers with tokenizers to character-level Tokenization-free methods.

3 Dataset

To evaluate three methods for sentiment analysis of news corpora, we choose Dow Jones Factiva as our dataset [1].

Dow Jones Factiva is a news database covering over a thousand sources in various fields. It contains pairs of news headlines and body texts with pre-labeled sentiments. We extract 1000 negative reports and 1000 positive reports. Since news headlines are good summaries, we treat headlines as **sentence-level news** and body texts as **document-level news**. After preprocessing, they are divided into training, validation, and test groups according to the ratio of $0.6 : 0.2 : 0.2$.

Dataset	Word count (Avg.)	Longer than 512
Headline	11.7795	0%
Body Text	755.53275	47.85%

Table 1: Distribution of length of headlines (sentence-level) and body texts (document level).

Table 1 reports the average word count of body text in each pieces of news as well as the percentage of news that longer than 512. This feature is important because the maximum sequence length of the BERT family models during tokenization is 512 due to limited memory and computational complexity (Devlin et al., 2019). It shows that 47.85% of the document-level news exceeds the sequence length of 512, so direct truncation may not be a good method for sentiment classification of body texts because of the probability of information loss.

4 Methodology

The code base is implemented adopting from Hugging Face Library (Lhoest et al., 2021). To con-

[1] https://www.dowjones.com/professional/factiva/

duct experiments on three methods, we import the pre-trained models of "deberta-v3-base", "bigbird-roberta-base", and "canine-s". A hyperparameter search is performed by running the three models on headlines and main texts, separately. In detail, the learning rate are fine-tuned with the Transformer Trainer method, randomly sampling from the range of ($1e^{-5}$, $5e^{-5}$). The evaluation loss is selected to be minimized, we record the accuracy and F1 score of the best run under each model. All computation is done using a single V100 GPU in HPC.

Previous work shows that the DeBERTa V3 and Big Bird achieve high accuracy (96.5% and 94.6%) on the Stanford Sentiment Treebank (SST-2) binary classification dataset (Socher et al.). SST-2 corpus is parsed and annotated by human judges that can allow for elaborate analysis on compositional effects of sentiment in language; however, it is composed of single sentences extracted from film reviews (Socher et al.), which is similar to sentence-level news but inconsistent with the conditions of document-level news (longer texts and might be harder to determine a single sentiment from mixed sentences). Therefore, we expect to see similar result on sentence-level sentiment analysis (i.e. headline): DeBERTa V3 may outperform Big Bird. Next, we aim to verify if the performance of Big Bird can increase significantly in the document-level sentiment analysis, given that the architecture of Big Bird is designed to understand longer sequences well. Finally, we want to test the performance of CANINE, namely, to explore the feasibility of applying the character-level tokenization-free method in news sentiment analysis.

5 Results and Analysis

Model	Class	Accuracy	F1
DeBERTa V3	Headline	0.9575	0.94117
	Content	0.9675	0.96821
Big Bird	Headline	0.9175	0.91473
	Content	0.97	0.96984
CANINE	Headline	0.855	0.84408
	Content	0.94	0.93034

Table 2: Accuracy and F1 scores of the three models on headlines and contents

5.1 Experimental Results

To find the best model/method on sentiment analysis of news corpora, we determine the baseline results as the measurements of **DeBERTa v3** on sentence-level and document-level news. Table2 shows the accuracy and F1 score of the best models after hyperparameter search.

5.2 Interpretation and Analysis

DeBERTa performs well at both levels. DeBERTa V3 achieves the highest accuracy and F1 score in sentence-level news sentiment analysis. It conforms to our first envision because DeBERTa outperforms Big Bird on SST-2, which can also be considered as a sentence-level SA problem. Since DeBERTa v3 consistently achieves over 95% accuracy on SST-2, sentence-level news, and document-level news, we can conclude that deep bidirectional methods are good at understanding both short and long news. Moreover, DeBERTa v3 outperforms the other two by more than 4% on headlines (sentence-level news). We believe the method of applying deep bidirectional Transformer with pre-training tasks is more suitable for understanding short news than the other two.

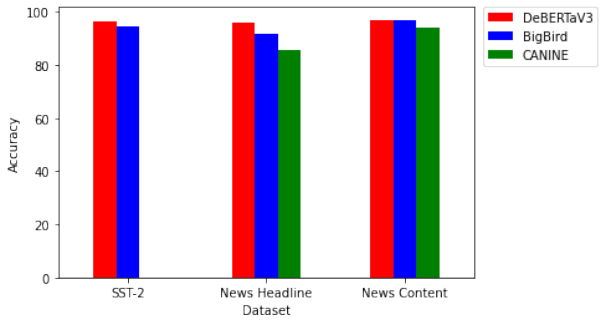

Figure 1: Comparison on accuracy among three models

Big Bird improves significantly on document-level news. We noticed that Big Bird performed much better in document-level news. This proves that Big Bird is indeed specialized in understanding longer sequences, as highlighted in its documentation. Specifically, it was only 91.75% accurate on SA of news headlines, but 97% accurate on the body texts. This result even exceeds the accuracy of DeBERTa v3 on ducument-level news and is the highest among the three models. Hence, we suggest using transformers designed to process longer sequences to handle ducument-level news

Tokenization-free methods are not competitive for understanding news. Getting rid of the word tokenization approach sounds promising as it may avoid mismatches between words and tokens. However, the accuracy on CANINE are lower than traditional tokenization methods at both levels. Considering that CANINE directly encodes the letters splitted from each word, we believe this method actually leads to more mismatches. The results are reasonable because splitting words into letters easily loses a lot of information about each word. Although we are not sure whether Deep Bidirectional method or Longer Sequence method can achieve SOTA in news corpora, we can certainly tell that future works should avoid splitting words into letters in understanding news corpora.

Document-level news are more comprehensible for machines. The results show that the accuracy of document-level news is higher than that of sentence-level news under each model. Although one might intuitively think that it is difficult to extract a single sentiment in longer texts because positive and negative sentences can be mixed up, our experimental results refute this idea. We believe this is because most news reports are carefully worded, in which case sentiment inconsistencies between sentences may not be common in document-level journalism. Besides, previous work claims that a sentence is just a short document and document is a collection of sentences, therefore, there is no fundamental differences between sentence-level and document-level sentiment analysis (Liu, 2012). However, our results suggest that the deeper underlying mechanism is worth investigating. For example, some preliminary research have already focused on longer texts by increasing weights of the important sentences at the document-level (Choi et al., 2021).

6 Github

All code files for this project can be found here: https://github.com/Richard1007/sentiment_news

7 Conclusion

Based on our results, we can safely conclude that deep bidirectional methods consistently perform well in understanding sentence-level and document-level news. For the main news text, the method for processing longer sequences achieves the highest performance among the three methods. However, token-free approaches are not competitive on either level.

Respectively, although DeBERTa v3 and Big Bird achieve the highest accuracy for understanding headlines and body texts, it cannot be concluded that these models will continue to be the best after pre-training with news specific data. Since DeBERTa v3 is very close to Big Bird in understanding document-level news, it is reasonable to suspect that DeBERTa v3 might outperform Big Bird after pre-training on news corpora.

Furthermore, the three methods introduced in this work are not comprehensive. There might exist other methods that we have not covered and also perform well in understanding news. Therefore, the purpose of this work is not to propose a SOTA model in understanding news, but only attempts to provide empirical advice on which features tend to be more useful than others.

Ethical Consideration

Nowadays, news plays a huge role in guiding public thought. A specialized model towards news corpora can tell readers the sentiment of the news in advance, so as to protect readers from believe the reports on trust. For instance, in a country where multiple parties take turns governing, news articles published by some media are more inclined to publicize the benefits of one party and belittle rival parties. If we can inform readers the overall sentiment of the news in advance, readers can view the article from a more objective point of view. Moreover, research shows that some news reports deliberately exaggerate the sentiment of news headlines to attract readers (Reis et al., 2015). If there exist a news-specific model, readers may see the pre-labeled sentiments of news headlines and body texts. If they are different, people can identify in advance which headlines might be biased.

On the other hand, such a model can be detrimental. If it can tell the sentiment of a news report in advance, news editors may only include those articles that are beneficial to their own interests. For example, an U.S. news organization may only choose to promote news that is unfavorable to Serbia. Such one-sided reporting may leave readers with incorrect information if they are not aware of.

Collaboration Statement

Haohai is responsible for understanding the details of three models and writing the background information for them and part of the introduction and results analysis. Yunzhe Sun is responsible for code implementation, running the models, the data processing and analysis, and providing the introduction and part of the background. Mingxuan collects the positive and negative data from the online dataset. We completed the reference together and finished the revision and editing.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *CoRR*, abs/1908.10063.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Gihyeon Choi, Shinhyeok Oh, and Harksoo Kim. 2021. Improving document-level sentiment classification using importance of sentences.

Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Sasko, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clement Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. *CoRR*, abs/2109.02846.

Bing Liu. 2012. Sentiment analysis and opinion mining.

Julio Reis, Fabrıcio Benevenuto, Pedro O. S. Vaz de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. 2015. Breaking the news: First impressions matter on online news.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. *Advances in neural information processing systems*, 30.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.