Z-Index at BLP-2023 Task 2: A Comparative Study on Sentiment Analysis

Prerona Tarannum[1]†, **Md. Arid Hasan**[2]†, **Krishno Dey**[2], **Sheak Rashed Haider Noori**[1]
[1]Daffodil International University, Dhaka, Bangladesh
[2]SE+AI Research Lab, University of New Brunswick, Fredericton, Canada
prerona15-14134@diu.edu.bd, arid.hasan@unb.ca,
krishno.dey@unb.ca, drnoori@daffodilvarsity.edu.bd

Abstract

In this study, we report our participation in Task 2 of the BLP-2023 shared task. The main objective of this task is to determine the sentiment (Positive, Neutral, or Negative) of a given text. We first removed the URLs, hashtags, and other noises and then applied traditional and pretrained language models. We submitted multiple systems in the leaderboard and BanglaBERT with tokenized data provided the best result and we ranked 5th position in the competition with an F1-micro score of 71.64. Our study also reports that the importance of tokenization is lessening in the realm of pretrained language models while the base models outperform the large models. In further experiments, our evaluation shows that BanglaBERT outperforms, and predicting the neutral class is still challenging for all the models.

1 Introduction

Sentiment Analysis is one of the most modern and sophisticated Natural Language Processing (NLP) applications. It is used for analyzing how people feel about the words they write in publicly accessible spaces like social media in the form of posts or comments. Social networking sites and other ways to use digital technology are commonly used to post a lot of information about feelings, ideas, and actions. Access to such a great amount of data provides the researchers the advantage to analyze the contents in order to help make decisions to process and understand the sentiment of a product and system, or views on social, international, cultural, and political agendas Hasan et al. (2020a).

The majority of current research is limited to resource-rich languages due to the availability of resources. The interest in low-resource languages is growing over time in sentiment analysis(Batanović et al., 2016; Nabil et al., 2015; Muhammad et al., 2023). Unlike other languages, a limited number of study has been done to develop resources for Bangla sentiment analysis (Hasan et al., 2020a; Alam et al., 2021; Islam et al., 2021; Hasan et al., 2023b; Islam et al., 2023). From the perspective of modeling, there have been studied both classical (i.e., SVM, RF, Naive Bayes) and deep learning (i.e., CNN, LSTM) models. Pretrained language models (i.e., BERT, XLM-RoBERTa, DistilBERT) have also been studied in recent years (Hasan et al., 2020a; Alam et al., 2021) for sentiment classification. Due to the availability of public data and inadequate information on annotation agreements (Alam et al., 2021), it is challenging for the researchers to focus on this area. This shared task provides a dataset by combining perfect and moderate agreement to shed light on sentiment analysis.

In this study, we participated in the Sentiment Analysis Shared Task at BLP-2023 and worked on a multiclass dataset where the class labels are Positive, Negative, and Neutral. We utilize both classical and transformer-based pretrained language models. For the classical model, we choose Support Vector Machine (SVM) and Random Forest (RF). We fine-tuned BERT multilingual, BanglaBERT base, and BanglaBERT large pretrained language models to train and evaluate models. Our findings from the study conclude as:

(i) The importance of tokenization before feeding into the models is diminishing in the presence of pretrained language models. There is little to no difference in performances between tokenized and non-tokenized data.

(ii) All the models are struggling to classify the neutral class.

(iii) Fine-tuned monolingual pretrained models outperform multilingual models.

(iv) Base model outperforms the large model.

The rest of the structure of this paper is as follows. We provide a brief overview of the literature in section 2. We discussed the data and approaches that we used for our experiments in section 3. Fol-

†The authors contributed equally to this work

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 324–330

lowing this in Section 4, we report results and discuss our findings. Finally, we conclude our work in Section 5.

2 Literature Review

Researchers are increasingly interested in investigating sentiment analysis utilizing social media data as a result of the rise of social media. The development of sentiment analysis began in the early 2000s (Pang et al., 2002). Early research includes rule-based and classical methodologies whereas recent studies include deep learning-based and pretrained language models. Researchers have been trying to develop resources over time and as a result, manual and semi-supervised approaches (Chowdhury and Chowdhury, 2014; Alam et al., 2021; Islam et al., 2021, 2023; Kabir et al., 2023) have been adopted in developing sentiment classification datasets. Chowdhury and Chowdhury (2014) used a semi-supervised technique to annotate data and train classical models. The study by Islam et al. (2021) constructs a dataset using manual annotations done by the annotators and presents 15,000 data in 13 domains. Rahman and Kumar Dey (2018) in their work, used the ABSA dataset consisting of human-annotated user comments on cricket and customer reviews of restaurants where SVM offered the maximum precision rate for both datasets.

Islam et al. (2016) developed a sentiment classification system utilizing SVM and Naive Bayes for textual movie reviews in Bangla and provided comparative results. Additionally, Naive Bayes with rules has been studied by Islam et al. (2016) for Bangla Facebook statuses sentiment classification. Hassan et al. (2016) worked on 10,000 post-processed text samples in both Bangla and Romanization of Bangla and by experimenting with LSTM, the authors achieved the maximum accuracy score of 55%. Hasan et al. (2020a) conducted comparison experiments using various datasets that existed in the literature to understand model performances, training difficulties, and consequences for real-world deployment. In this study, deep learning-based models outperform traditional models.

Furthermore, Alam et al. (2021) used the most sophisticated techniques currently available to compare datasets and conclude that XLM-RoBERTa exhibits the best performance over other deep learning approaches. Classifying the tweets of positive, negative, and neutral polarity was the major goal of SAIL-2015 Patra et al. (2015). Various well-known supervised classification methods have been studied in this study. Tripto and Ali (2018) used LSTM for identifying sentiment and emotions in Bangla writings achieving an accuracy of 65.97 and 54.24 for three and five classes respectively. Chowdhury et al. (2019) providing a method for conducting sentiment analysis on Bangla-language movie reviews that can automatically analyze viewer responses to a certain film or television program was the main work and the authors used social media websites' publicly accessible comments and posts serving as the source of the dataset that was manually compiled and labeled for this experiment.

Focusing on the largest publicly available dataset MUBASE (Hasan et al., 2023b) consolidated from social media data consisting of 33,605 tweets and Facebook comments about Bangla news and carried out experiments that went beyond traditional approaches and smaller transformer-based models. The authors focused on the efficiency of sophisticated algorithms in zero- and few-shot conditions, including Flan-T5, GPT-4, and Bloomz. The findings show that while LLMs are an interesting study area, smaller variations of precise pre-trained models perform better. In the context of sentiment analysis Cambria et al. (2022) provides a commonsense-based neurosymbolic framework that seeks to address these problems. They evaluated SenticNet 7 and concluded that of all 20 lexica, SenticNet 7 was the most effective. Ye et al. (2022) worked with the manually produced and labeled datasets that were obtained from social media. The accuracy achievement at the end of 140 epoch with the best performance using the NADAM optimizer.

3 Methodology

3.1 Data

Class	Train	Dev	Dev-Test	Test
Positive	12,364	1,388	1,126	2,092
Neutral	7,135	793	600	1,277
Negative	15,767	1,753	1,700	3,338
Total	**35,266**	**3,934**	**3,426**	**6,707**

Table 1: Class label distribution of the shared task dataset for each data split.

We utilized the dataset provided by the organizers of the BLP-2023 for task 2: Sentiment Analysis (Hasan et al., 2023a) . The goal is to iden-

Input text: **নতুন সেনাপ্রধান লে. জে. এস এম শফিউদ্দিন আহমেদ **২৪ জুন দায়িত্ব নেবেন এস এম শফিউদ্দিন আহমেদ

Tokenized Text: ['*', '*', 'নতুন', 'সেনাপ্রধান', 'লে', '.', 'জে', '.', 'এস', 'এম', 'শফি', '##উদ্দিন', 'আহমেদ', '*', '*', '২৪', 'জুন', 'দায়িত্ব', 'নেবেন', 'এস', 'এম', 'শফি', '##উদ্দিন', 'আহমেদ']

Encoded Text: [2, 14, 14, 1299, 21384, 1128, 18, 1683, 18, 1880, 1611, 28485, 4286, 3232, 14, 14, 3083, 3702, 2140, 7453, 1880, 1611, 28485, 4286, 3232, 3]

Figure 1: Representation of tokenized training text of id: 30960

tify the sentiment contained within a text. The dataset is consolidated from two distinct sources, i) MUBASE (Hasan et al., 2023b) and ii) Sent-NoB (Islam et al., 2021) consisting of social media tweets, posts, and comments. In this dataset, there are three columns, ID refers to sentence id, text refers to input text, and label containing Positive, Neutral, and Negative tags. In table 1, we present the class-wise official data distributions that are provided in the shared task.

3.2 Preprocessing

The dataset which is given for the Sentiment Analysis shared task at BLP-2023 was generated via social media, where it contains noise like emoticons, usernames, hashtags, URLs, invisible letters, and symbols. We went through numerous preprocessing stages to clear up these noisy data. We first removed unnecessary characters and URLs and then we removed the stopwords hashtags, and usernames from the data. We also used normalizer (Hasan et al., 2020b) before feeding into the pretrained language model.

3.3 Model

We run some traditional models and BERT-based models on the dataset. Several factors have been considered during the selection of these algorithms. The superior performance for the Bangla language is one of the main reasons for choosing BanglaBERT (Bhattacharjee et al., 2022) and BERT multilingual(Devlin et al., 2018) provides comparable results. We used two variants (base and large) of BanglaBERT. For the traditional models, we choose two popular algorithms such as Random Forest (RF) (Liaw et al., 2002) and SVM (Platt, 1998).

3.4 Experiments

BERT-based Models: Transformer toolkit (Wolf et al., 2020) is used in our study to fine-tune transformer-based models. We used a learning rate of $2e-5$ for optimizer Adam, batch size of 16, gradient accumulation of 1, and maximum sequence length of 256. BanglaBERT base version is trained on the BERT model, as a result, both BanglaBERT-base and BERT multilingual have 110M trainable parameters whereas BanglaBERT large is trained on the Electra model containing 335M parameters. For the transformer-based models, we run 3 epochs for all the models for better understanding. All models are trained on both tokenized and non-tokenized data and the change in performances is little to no on tokenized and non-tokenized data. To feed the non-tokenized data into the model, we added all the vocab of the dataset set to the pre-trained tokenizer which uses a Byte-Pair Encoding (BPE) tokenizer. As a result, we managed to ignore the default behavior of the BPE tokenizer, and the words were not tokenized by the BPE tokenizer. The representation of our tokenized and non-tokenized data is shown in Figure 1 and Figure 2 respectively.

Traditional Models: In order to train the traditional models, we first create tf-idf vectors with weighted n-gram from the preprocessed data. To use the contextual information, we utilized unigram, bigram, and trigram as part of weighted n-gram. We extract a fixed number of features (1,500) from the data and feed it to the models. Both models are trained on both tokenized and non-tokenized data and the performances remain the same on tokenized and non-tokenized data.

4 Results and Discussion

The official overall ranking and results determined by the lab organizers are presented in Table 2. The official evaluation metric for task 2 is F1-micro. We also presented the best system and baseline results (majority, random) including our system in Table 2. The last submission is considered for the leaderboard and our last submission is the BanglaBERT base model. In the competition, we officially ranked 5th position with an F1-micro score of 71.64 where the best system provides an

Input text: **নতুন সেনাপ্রধান লে. জে. এস এম শফিউদ্দিন আহমেদ **২৪ জুন দায়িত্ব নেবেন এস এম শফিউদ্দিন আহমেদ

Tokenized Text: ['**নতুন', 'সেনাপ্রধান', 'লে.', 'জে.', 'এস', 'এম', 'শফিউদ্দিন', 'আহমেদ', '**২৪', 'জুন', 'দায়িত্ব', 'নেবেন', 'এস', 'এম', 'শফিউদ্দিন', 'আহমেদ']

Encoded Text: [2, 32736, 21384, 32737, 32738, 1880, 1611, 32739, 3232, 32740, 3702, 2140, 7453, 1880, 1611, 32739, 3232, 3]

Figure 2: Representation of non-tokenized training text of id: 30960

Text: Pranoy Sen তখন পাকিস্তান ও আফগানিস্তান ভারতের হয়ে যাবে ।
Gold Label: Neutral
Predicted Label: Negative

Text: বিশ্বে উৎপাদিত করোনা টিকার বেশিরভাগই ব্যবহার করেছে... বিস্তারিত নিউজে
Gold Label: Neutral
Predicted Label: Positive

Figure 3: Example of sentences with wrong predictions for neutral class by BanglaBERT model.

F1 score of 73.10. Our system also performed better than both the majority and random baseline with a large margin of 21.87 and 38.08 respectively.

The detailed results of all the performed experiments are presented in Table 3. Once the submission period was over and the test set with labels became available, we conducted all the experiments again and reported the comprehensive findings. As shown in the reported results, we can state that the BanglaBERT approach with tokenized data outperforms other experiments by providing an accuracy of 71.64 with respect to the positive class F1 score of 75.59. With non-tokenized data, BanglaBERT gives an accuracy of 71.49 where the F1 score with respect to positive class is 75.18. Across the datasets, there is a definite tendency for the tokenized dataset to give better performances while evaluating than non-tokenized data. The performance between tokenized and non-tokenized data before feeding into networks for BERT-based models is little to none and for the traditional models, the performances remain the same. As a result of this, we can conclude that the importance of tokenization before feeding into the models is diminishing in the realm of the pretrained language models because each pretrained language model uses a model-specific tokenizer.

In table 3, all the models struggle to identify whether the data is in the neutral class because neutral class data are highly correlated with either positive class or negative class data, making it difficult for the models. Among all the models, both traditional models poorly perform to predict the neutral class. Although the BERT-based models perform well in comparison with traditional models on neutral class, the results are not comparable with the other two classes. We also explored the model performances for predicting neutral class and we came up with interesting findings which include if the text contains the words from frequently occurring words of the positive class, the text is classified as positive or negative if the frequently occurring words belong to negative class. We present two examples where our model couldn't predict neutral classes in Figure 3 for a better understanding of our findings.

Model	F1-micro	Rank
BanglaBERT	**71.64**	5^{th}
Best system	73.10	1^{st}
Baseline (Majority)	49.77	25^{th}
Baseline (Random)	33.56	29^{th}

Table 2: Official results on the test set and overall ranking of Task 2: Sentiment Analysis. **Bold** indicates our systems.

In our study, we found that monolingual pretrained language provides superior performance compared with the multilingual pretrained language model. We achieved an F1 score of 66.81 with respect to the positive class using BERT-multilingual while the BanglaBERT-base model has an F1 score of 75.59 with respect to the positive class which demonstrates the superior performance of the monolingual pretrained language model. We also observed that the base model outperforms the large model. The large model has more trainable parameters than the base model and the amount of

data is not sufficient to train and overfit the large model.

L	Model	Acc	P	R	F1
Neg	SVM*	54.76	58.72	73.46	65.26
Neu			37.08	09.55	15.19
Pos			49.91	52.53	51.19
Neg	SVM	54.76	58.72	73.46	65.26
Neu			37.08	09.55	15.19
Pos			49.91	52.53	51.19
Neg	RF*	55.42	58.65	76.18	66.28
Neu			41.33	12.69	19.41
Pos			51.14	48.37	49.72
Neg	RF	55.42	58.65	76.18	66.28
Neu			41.33	12.69	19.41
Pos			51.14	48.37	49.72
Neg	M1*	71.49	77.16	79.66	78.39
Neu			49.16	41.19	44.82
Pos			73.48	76.96	75.18
Neg	M1	71.64	78.77	80.77	78.39
Neu			48.88	37.59	42.50
Pos			73.44	77.87	**75.59**
Neg	M2*	70.61	77.30	78.16	77.73
Neu			48.53	38.84	43.15
Pos			70.61	77.96	74.10
Neg	M2	70.66	76.60	78.34	77.46
Neu			48.83	40.72	44.41
Pos			71.99	76.67	74.26
Neg	M3*	64.95	71.49	73.10	72.29
Neu			43.72	39.55	41.53
Pos			65.97	67.45	66.70
Neg	M3	65.01	71.50	73.07	72.28
Neu			44.24	38.76	41.32
Pos			65.50	68.16	66.81

Table 3: Detail results on the test set of **Task 2: Sentiment Analysis**. **Bold** indicates the best F1 score for positive class. * indicates the model trained and evaluated on non-tokenized data. L: Label, P: Precision, R: Recall, F1: F1-score, Neg: Negative, Neu: Neutral, Pos: Positive, M1: BanglaBERT, M2: BanglaBERT large, M3: BERT multilingual.

5 Conclusion

In this study, we run comparative experiments and analysis on the Bangla sentiment dataset provided by the task organizers of BLP-2023. We presented a detailed comparison of the fine-tuned models along with traditional models. Comparing the traditional model, we found that SVM outperforms RF with a margin of 1.47%. BanglaBERT outperforms all the models we used in our study. Our study also reveals that tokenization has little to no control over performance during the use of pretrained language models. In the submission of task 2 on the Sentiment Analysis dataset, we ranked 5^{th} position among all the participants. To extend this work, we will employ large language models (LLMs) and GPT-based models for comparative and in-depth sentiment analysis.

Limitations

The pretrained language models show promising performances toward tackling the sentiment analysis problem presented for this shared task. However, our models keep failing to predict neutral class, and we overfit the larger models (i.e., BanglaBERT large). Although we perform different hyperparameter tuning and dropouts for all the models, we are not able to find the optimal hyperparameters for each model. As a result, we decided to use the constant hyperparameter for all the models which causes overfitting the large model.

References

Firoj Alam, Arid Hasan, Tanvirul Alam, Akib Khan, Janntatul Tajrin, Naira Khan, and Shammur Absar Chowdhury. 2021. A review of bangla natural language processing tasks and the utility of transformer models. *arXiv preprint arXiv:2107.03844*.

Vuk Batanović, Boško Nikolić, and Milan Milosavljević. 2016. Reliable baselines for sentiment analysis in resource-limited languages: The serbian movie review dataset. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2688–2696.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. Senticnet 7: A commonsense-based neurosymbolic ai framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3829–3839.

Rumman Rashid Chowdhury, Mohammad Shahadat Hossain, Sazzad Hossain, and Karl Andersson. 2019. Analyzing sentiment of movie reviews in bangla by applying machine learning techniques. In *2019 international conference on bangla speech and language processing (ICBSLP)*, pages 1–6. IEEE.

Shaika Chowdhury and Wasifa Chowdhury. 2014. Performing sentiment analysis in Bangla microblog posts. In *2014 International Conference on Informatics, Electronics Vision (ICIEV)*, pages 1–6.

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*.

Md. Arid Hasan, Firoj Alam, Anika Anjum, Shudipta Das, and Afiyat Anjum. 2023a. Blp-2023 task 2: Sentiment analysis. In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*, Singapore. Association for Computational Linguistics.

Md Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2023b. Zero-and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis. *arXiv preprint arXiv:2308.10783*.

Md Arid Hasan, Jannatul Tajrin, Shammur Absar Chowdhury, and Firoj Alam. 2020a. Sentiment classification in bangla textual content: A comparative study. In *2020 23rd international conference on computer and information technology (ICCIT)*, pages 1–6. IEEE.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020b. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online. Association for Computational Linguistics.

Asif Hassan, Mohammad Rashedul Amin, N Mohammed, and AKA Azad. 2016. Sentiment analysis on bangla and romanized bangla text (brbt) using deep recurrent models. *arXiv preprint arXiv:1610.00369*.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. Sentnob: A dataset for analysing sentiment on noisy bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271.

Md Ekramul Islam, Labib Chowdhury, Faisal Ahamed Khan, Shazzad Hossain, Sourave Hossain, Mohammad Mamun Or Rashid, Nabeel Mohammed, and Mohammad Ruhul Amin. 2023. SentiGOLD: A large bangla gold standard multi-domain sentiment analysis dataset and its evaluation. *arXiv preprint arXiv:2306.06147*.

Md Saiful Islam, Md Ashiqul Islam, Md Afjal Hossain, and Jagoth Jyoti Dey. 2016. Supervised approach of sentimentality extraction from bengali facebook status. In *2016 19th international conference on computer and information technology (ICCIT)*, pages 383–387. IEEE.

Mohsinul Kabir, Obayed Bin Mahfuz, Syed Rifat Raiyan, Hasan Mahmud, and Md Kamrul Hasan. 2023. Banglabook: A large-scale bangla dataset for sentiment analysis from book reviews. *arXiv preprint arXiv:2305.06595*.

Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomforest. *R news*, 2(3):18–22.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, et al. 2023. Afrisenti: A twitter sentiment analysis benchmark for african languages. *arXiv preprint arXiv:2302.08956*.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2515–2519.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.

Braja Gopal Patra, Dipankar Das, Amitava Das, and Rajendra Prasath. 2015. Shared task on sentiment analysis in indian languages (sail) tweets-an overview. In *Mining Intelligence and Knowledge Exploration: Third International Conference, MIKE 2015, Hyderabad, India, December 9-11, 2015, Proceedings 3*, pages 650–655. Springer.

J. Platt. 1998. *Fast Training of Support Vector Machines using Sequential Minimal Optimization*. MIT Press.

Md Atikur Rahman and Emon Kumar Dey. 2018. Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation. *Data*, 3(2):15.

Nafis Irtiza Tripto and Mohammed Eunus Ali. 2018. Detecting multilabel sentiment and emotions from bangla youtube comments. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–6. IEEE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, EMNLP '20, pages 38–45, Online. Association for Computational Linguistics.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Ves Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2022. Complementary explanations for effective in-context learning. *arXiv preprint arXiv:2211.13892*.