UMUTeam at SemEval-2023 Task 12: Ensemble Learning of LLMs applied to Sentiment Analysis for Low-resource African Languages

José Antonio García-Díaz¹, Camilo Caparros-Laiz¹, Ángela Almela², Gema Alcaráz-Mármol³, María José Marín-Pérez², Rafael Valencia-García¹

¹ Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain

{joseantonio.garcia8, camilo.caparros1, valencia}@um.es

² Facultad de Letras, Universidad de Murcia, Campus de la Merced, 30001 Murcia, Spain {angelalm, mariajose.marin1}@um.es

³ Facultad de Educación, Universidad de Castilla La Mancha, 45071 Toledo, Spain Gema.Alcaraz@uclm.es

Abstract

These working notes summarize the participation of the UMUTeam in the SemEval 2023 shared task: AfriSenti, focused on Sentiment Analysis in several African languages. Two subtasks are proposed, one in which each language is considered separately and another one in which all languages are merged. Our proposal to solve both subtasks is grounded on the combination of features extracted from several multilingual Large Language Models and a subset of language-independent linguistic features. Our best results are achieved with the African languages less represented in the training set: Xitsonga, a Mozambique dialect, with a weighted f1-score of 54.89%; Algerian Arabic, with a weighted f1-score of 68.52%; Swahili, with a weighted f1-score of 60.52%; and Twi, with a weighted f1-score of 71.14%.

1 Introduction

The underlying objective of AfriSenti shared task (Muhammad et al., 2023b) is to promote methods, tools, and curated datasets for African languages (Muhammad et al., 2023a). Specifically, the organizers of this task propose a Sentiment Analysis (SA) task focused on 16 African languages, including several languages from Nigeria, Ethiopia, Kenya, Tanzania, Algeria, Rwanda, Ghana, Mozambique or South Africa. This task is divided into two subtasks. The first one is a monolingual sentiment classification, in which the polarity of a sentence should be determined. In this task, the documents from different languages are not mixed. The second subtask is a multilingual sentiment classification. In both cases, polarity is a multi-classification problem of three labels, namely, positive, neutral and negative.

To solve these subtasks, our research group proposes a deep-learning classifier based on ensemble

learning. This ensemble combines the predictions of a subset of language-independent linguistic features with features from several Large Language Models (LLMs). The evaluated LLMs include multilingual general purpose models such as BERT, XLM, and DeBERTA; DziriBERT, focused on the Algerian dialect, and AfriBERTa, a multi-lingual LLM trained on 11 languages from Africa. We also included the original BERT as some of the documents in the dataset are written in English.

For subtask 1, all languages are evaluated separately. However, we decided to train all languages together in order to add some instances of underrepresented languages in the provided corpora. Moreover, the majority of the evaluated LLMs are multilingual so they can handle documents from different tracks in the same batch. This idea has worked well, as our best results are achieved in languages with not so many instances. We consider that training all the documents for all the tasks involved in the first subtask has been beneficial for those languages. However, we obtained better results, but worse classification, in the best represented languages. In this sense, our general participation is far from the top-ten best systems.

Additional resources concerning our participation in this shared task can be found at https://github.com/NLP-UMUTeam/ semeval-2023-afrisenti.

2 Background

The dataset of the AfriSenti 2023 shared task consists in 107,549 documents annotated as positive, negative, or neutral. The dataset contains documents written in 14 African languages. According to the description of the task, each document in the dataset was annotated by three people, following the annotation guidelines described at (Mohammad, 2016), and the final sentiment was determined by a majority vote. More details of the dataset can be found at (Muhammad et al., 2022).

From this dataset, we extracted a custom validation split from the training instances in a proportion of 80-20. Table 1 depicts the dataset statistics including all the splits. We can observe that the dataset is almost balanced for all labels. However, the number of instances for each language is not balanced, as it can be observed from Figure 1. This strong imbalance in the data makes some tracks a real challenge.

Table 1: Dataset statistics of the AfriSenti shared task

label	train	val	test	total
negative neutral positive	16,085 18,232 16,631	4,023 4,562 4,152	4,341 4,899 4,413	24,449 27,693 25,196
Total	50,948	12,737	13,653	77,338



Figure 1: Percentage of instances by track, including: Hausa (ha), Yoruba (yo), Igbo (ig), Nigerian Pidgin (pcm), Amharic (am), Algerian Arabic (dz), Moroccan Arabic/Darija (ma), Swahili (sw), Kinyarwanda (kr), Twi (twi), Mozambican Portuguese (pt), Xitsonga (ts), Setswana, isiZulu, and Xitsonga (ts).

Our first analysis of the dataset revealed that some documents are written in languages other than the track indicates. Accordingly, we decided to use the language identification module of fast-Text¹ (Joulin et al., 2016) to confirm the language of each document individually. FastText's identification module outputs a list of languages and its probabilities. As the documents are short texts from Twitter, we consider that the main language is the one with higher probability; however, we set a threshold of .75. That is, we consider that the fastText output is not reliable for those documents in which the maximum probability does not reach this threshold. We observed that 6,900 documents were annotated as English documents. There was also a small portion of the documents which were identified as Spanish (es), French (fr), Italian (it). Out of the documents identified as English, 5,882 were from the Nigerian Pidgin (pcm) track, 417 from the Igbo (ig) track, 281 from Hausa (ha) track and 156 from Yoruba (yo).

3 System Overview

Our participation can be summarized as the training and submission of three ensemble learning models based on language-independent Linguistic Features (LF) and several LLMs.

The linguistic features are obtained from UMU-TextStats (García-Díaz et al., 2022c). This tool is designed for the Spanish language; however, a subset of the features that incorporate stylometric and morphosyntactic features is language independent. It is worth mentioning that these LFs have been evaluated in NLP tasks, such as Author Profiling (García-Díaz et al., 2022a), satire identification (García-Díaz and Valencia-García, 2022), or hate-speech identification (García-Díaz et al., 2022b), among others. To extract the LF concerning stylometric and morphosyntactic features UMUTextStats relies on the Named Entity Recognition (NER) and Part-of-Speech (PoS) models from Stanza (Qi et al., 2020). However, not all the languages involved in the AfriSenti shared task have models in Stanza, and we could extract only the entities for the models based on the language identified by fastText.

The second type of features are obtained from several LLMs. Our method concerning the usage of the LLM applied to a SA task can be described as follows. First, we conduct a hyperparameter optimization stage by using RayTune (Liaw et al., 2018). For each LLM, we evaluate 10 models and the parameters are selected using Tree of Parzen Estimators (TPE) (Bergstra et al., 2013). The evaluated hyperparameters are weight decay, batch size, warm-up, the number of epochs and the learning rate. Once we obtained the best combination of parameters for each LLM, we extract the sentence embeddings from the documents in the corpus. To that end, we extract the weights of the [CLS] token of each document (Reimers and Gurevych, 2019). We follow this approach as it simplifies the combi-

¹https://fasttext.cc/docs/en/ language-identification.html

nation of the LFs with the LLMs. After that, the evaluated LLMs were described.

- AfriBERTa-large (Ogueji et al., 2021). This LLM is trained with 11 languages, namely Oromo, Amharic, Gahuza, Hausa, Igbo, Nigerian Pidgin, Somali, Swahili, Tigrinya and Yorùbá. This model is trained using Masked-Language Modeling but without Next Sentence Prediction. We use the large version of this model, with 10 layers, 6 attention heads, 768 hidden units and 3072 feed forward size.
- AfroXLMR-base (Alabi et al., 2022). This LLM is based on XLM and trained using 17 african languages such as Amharic, Hausa, Oromo, Naija, Somali, Swahili and isiXhosa among others. We use the base version of this LLM due to hardware limitations.

It is worth noting that This LLM is not considered in our official submission as it was suggested during the revision process. Accordingly, we include the results of this model and as part of the ensemble learning models only with the custom validation split.

- BERT (Devlin et al., 2018). BERT is the first architecture based on Transformers. This model was developed by Google and it is based on bidirectional representations of natural language. The bidirectional strategy allowed the user to analyze text from left to right and vice versa, allowing to better understand the context of a sentence. BERT has a multilingual version called multilingual BERT (Devlin et al., 2018). We used both models in this research. We decided to incorporate the English version of BERT due to the large number of documents identified as English by fastText's language detection model. Both for English and multilingual BERT we use the base cased version.
- DziriBERT (Abdaoui et al., 2021) is a pretrained language model for the Algerian language, containing both Arabic and Latin symbols. The dataset used for training this model, however, is not very large, as is less than 1 million tweets.
- multilingual DeBERTa (He et al., 2021). This multilingual LLM is based on DeBERTa architecture although it incorporates a disentangled attention mechanism, whereby each

token is represented with its content and its position and the attention weights among words are computed with disentangled matrices. DeBERTa significantly improved the performance of both Natural Language Understanding (NLU) and Generation downstream tasks.

- XLM (Conneau et al., 2019). This multilingual LLM is based on the RoBERTa architecture and it was trained on data from 100 languages extracted from the CommonCrawl dataset.
- XLM-Twitter (Barbieri et al., 2021). This LLM is based on XLM but was trained on almost 200 millions of tweets from almost 30 languages. We use the base model.

The results of the hyper optimization stage for subtasks 1 and 2 are depicted in Table 2. It can be observed that the best results vary from LLM and task; however, some LLMs such as BERT, mBERT, XLM and XLM-T benefit from larger batch sizes.

Once we obtained the LFs and fine-tuned the LLMs, we created three ensemble learning models that combine all the predictions. Each ensemble is based on a different strategy to combine the features: (1) highest probability, where the final prediction is based on the model which outputs the larger probability from certain sentiment; (2) average probabilities, where the probabilities for each LLM and sentiment are averaged and the final output is based on the highest value; and (3) hard-voting, which is the mode of the predictions.

4 Results

4.1 Results with the custom validation split

Firstly, we have reported the results obtained with the custom validation split. These results have been reported with the weighted F1-score for both subtasks (see Table 3). The results in these tables are grouped for each model architecture and feature set. The first row concerns the LFs. The next set of rows are the LLMs: (1) AfriBERTa-large, (2) AfroXLMR-base, (3) BERT-base, (4) DZIRIBERT, (5) mBERT base (multilingual BERT), (6) multilingual DEBERTA-base, (7) XLM-base, and (8) XMLTwitter-base. The two bottom rows show the results of the ensemble learning strategies (EL) that combine the predictions of the LLMs and the LFs. The first set of ensembles are only based on the

model	learning_rate	epochs	batch_size	warmup_steps	weight_decay	
	Subtask 1					
afriberta-large	1.3e-05	2	8	500	0.12	
afroxlmr-base	2.1e-05	5	8	0	0.15	
bert-base	2.3e-05	3	16	0	0.17	
dziribert	2.4e-05	3	8	1000	0.061	
mbert-base	4.5e-05	3	16	1000	0.12	
mdeberta-base	2.4e-05	5	8	0	0.079	
xlm-base	1.2e-05	3	16	500	0.045	
xlm-twitter-base	2.9e-05	3	16	500	0.25	
Subtask 2						
afriberta-large	4.3e-05	2	16	0	0.068	
afroxlmr-base	2.4e-05	5	16	0	0.21	
bert-base	2.9e-05	2	16	250	0.021	
dziribert	3.5e-05	4	8	500	0.14	
mbert-base	2.6e-05	4	16	250	0.29	
mdeberta-base	1.2e-05	5	16	500	0.15	
xlm-base	1.6e-05	4	16	500	0.1	
xlm-twitter-base	3.2e-05	4	16	0	0.25	

Table 2: Results of the hyperparameter optimization stage

Table 3: Classification report with the custom validation split for subtask 1 (left) and subtask 2 (right). The models are stacked by groups. The first group are the LFs, the second group is constituted by the LLMs, the third group corresponds to the ensemble learning models of the LLMs and the fourth group are the ensembles of the LLMs and the LFs. All the metrics are weighted

		Ś	Subtask-1		Ś	Subtask-2	
Group	LLM	precision	recall	f1-score	precision	recall	f1-score
lf	lf	53.718	53.686	53.433	53.112	53.152	52.931
	afriberta-large	70.929	70.645	70.653	70.542	70.440	70.414
	afrox1mr-base	72.225	72.191	72.203	72.603	72.442	72.477
	bert-base	67.850	67.355	67.307	67.569	66.813	66.643
	dziribert	70.475	70.134	70.105	70.186	69.467	69.481
LLMs	mbert-base	68.680	68.014	67.947	68.347	68.030	67.969
	mdeberta-base	71.454	71.241	71.266	71.529	71.461	71.486
	xlm-base	69.110	68.729	68.770	70.240	70.024	70.010
	xlm-twitter-base	70.911	70.723	70.708	71.069	70.982	70.974
	highest probability	74.480	74.421	74.385	74.557	74.303	74.322
	mean	75.349	74.963	74.984	75.349	74.963	74.984
	mode	75.012	74.390	74.442	75.012	74.390	74.442
EL w/LF	highest probability	74.480	74.421	74.385	74.480	74.421	74.385
	mean	75.383	74.978	74.997	75.383	74.978	74.997
	mode	74.955	74.327	74.352	74.955	74.327	74.352

LLMs whereas the second group includes the LFs. These strategies are: (1) highest probability (HIGH-EST), (2) average of predictions (MEAN) and (3) hard voting (MODE).

Regarding the results for subtask 1 (see Table 3 -left-), it can be observed that all models display

similar performance concerning precision and recall. As expected, the language-independent LFs achieved limited results when compared with the LLMs. The reason for this is that the subset of linguistic features only captures a set of morphological and stylometric clues that are insufficient for an accurate sentiment analysis prediction. Concerning the performance of the LLMs, the results of the weighted F1-score is between 67.307 (BERT) and 72.203 (AFROXLMR), and the results achieved with the combination of the aforementioned features using ensemble learning strategies are superior, obtaining the best result averaging the probabilities of the models for the final classification, with a weighted f1-score of 74.994 considering both the LLMs and the LFs. It is worth noting that we have evaluated all tracks at once, rather than perform an evaluation for each track, as it is the official ranking of the AfriSenti shared-task. We adopted this scheme because it was faster to implement, it has a lower carbon footprint and and because it may favor languages with fewer instances in the training split.

Concerning the results for subtask 2 (see Table 3 -right-) and similarly to subtask 1, the models have similar performance concerning precision and recall. In addition, the performance of the LFs and the LLMs is also similar to the results achieved during task 1. Similarly, the best results are obtained using the ensemble learning strategy based on averaging the probabilities of all LLMs and the LFs; but the results compared with the ensemble learning of all LLMs without the LFs are very similar.

After that, we evaluated the predictions of the best model (ensemble learning with the averaging the probabilities strategy). Figure 2 depicts the confusion matrices for subtask 1 (left) and subtask 2 (right). The purpose of a confusion matrix is to compare the real and predicted values. In a multiclassification problem, such as the one proposed in the AfriSenti shared task, the confusion matrix can help to spot the importance of the neutral class or to identify whether there are wrong classifications in which negative documents are incorrectly labeled as positive or vice versa. As it can be observed from subtask 1 (left), the majority of wrong classifications for the negative and positive documents have occurred with the neutral class. In the case of neutral documents, we observed a larger number of incorrect classifications with the negative label (516) than with the positive label (377). A similar

behavior is observed in subtask 2 (right), whereby the neutral label is the typical error which appears instead of the positive and negative labels and that the majority of wrong classifications of the neutral classes are negative.

4.2 Official leader board

We participated in the final contest by submitting three runs, all based in ensemble learning but with different strategies. Our first strategy is the average of the probabilities, the second strategy is the highest probability and our last strategy is a hard voting scheme. It is important to notice that this ensemble learning does not contain the AfroXLMR-base model as this model was suggested during the revision of this working notes.

Table 4 depicts the results of the UMUTeam for each language for subtask 1 (tracks from 1 to 12) and the multilingual task (track 16). The ranking was established using the weighted F1 score. It is worth noting that we set the macro f1 score as the main metric during the hyperparameter evaluation stage, as we wanted to give the same importance to all labels. Our best result according to the ranking was obtained for track 12 (ts), with a weighted f1score of 54.89% and reaching position 7 out of 31 participants. We also achieved competitive results for track 6, with a weighted f1-score of 68.52% and position 13 out of 30 participants; track 8 (sw), with a weighted f1-score of 60.52% achieving position 14 out of of 30 participants; and track 11, with a weighted f1-score of 71.14% and reaching position 12. However, in track 11, we achieved a superior weighted f1-score, 73.92%, but in this case we are at the bottom of the ranking. As we used similar methods to participate in all the tasks, these results suggest that the degree of difficulty varies greatly across languages. Finally, in the multilingual track (16), we achieved a weighted f1-score of 65.47% with position 23 of a total of 33 participants.

We consider that our results are good in tracks with not so many documents, as we decided to face subtask 1 with all the tracks at once. However, our results are more limited as the number of documents increases since the other teams also achieve more competitive results.

5 Conclusion

Even though our results are not the best in the different tracks, we are very happy with our participation in this task. First of all, because we



Figure 2: Confusion matrices with the custom validation split for subtask 1 (left) and subtask 2 (right)

Track	Language (ISO)	Score	Ranking
01	HA	73.92	28/35
02	YO	66.10	26/33
03	IG	76.78	20/32
04	PCM	65.55	22/32
05	AM	55.39	21/29
06	DZ	68.52	13/30
07	MA	54.75	20/32
08	SW	60.52	14/30
09	KR	65.03	19/34
10	TWI	63.01	19/31
11	PT	71.14	12/30
12	TS	54.89	7/31
16	MUL	65.47	23/33

Table 4: Results for tracks from 1 to 12 and track 16, using the weighted f1-score

acknowledge the importance of the development of state-of-the-art linguistic resources for languages other than English. Second, because we could evaluate multilingual LLMs for a popular task such as Sentiment Analysis. However, as further work we will incorporate data augmentation techniques in order to give major support to underrepresented languages. Another further experiment is the training of models per individual tracks, instead to combine all languages together. This approach could reveal traits related to background and cultural differences for each dataset.

6 Acknowledgments

This work is part of the research projects (PDC2021-121112-I00) and LT-AIInFunds SWM (TED2021-131167B-I00) funded MCIN/AEI/10.13039/501100011033 by and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PSP (PID2019-107652RB-I00/AEI/ 10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033. In addition, José Antonio García-Díaz is supported by Banco Santander and the University of Murcia through the Doctorado Industrial programme.

References

- Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2021. Dziribert: a pre-trained language model for the algerian dialect. *arXiv preprint arXiv:2109.12346*.
- Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pretrained language models to African languages via

multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

- Francesco Barbieri, Luis Espinosa Anke, and José Camacho-Collados. 2021. XLM-T: A multilingual language model toolkit for twitter. *CoRR*, abs/2104.12250.
- James Bergstra, Daniel Yamins, and David Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
- José Antonio García-Díaz, Ricardo Colomo-Palacios, and Rafael Valencia-García. 2022a. Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020. *Future Generation Computer Systems*, 130:59–74.
- José Antonio García-Díaz, Salud María Jiménez-Zafra, Miguel Angel García-Cumbreras, and Rafael Valencia-García. 2022b. Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers. *Complex* & *Intelligent Systems*, pages 1–22.
- José Antonio García-Díaz and Rafael Valencia-García. 2022. Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers. *Complex & Intelligent Systems*, 8(2):1723–1736.
- José Antonio García-Díaz, Pedro José Vivancos-Vicente, Angela Almela, and Rafael Valencia-García. 2022c. Umutextstats: A linguistic feature extraction tool for spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6035– 6044.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651.

- Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.
- Saif Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 174–179.
- Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023a. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages.
- Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif M. Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023b. SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). In *Proceedings of the* 17th International Workshop on Semantic Evaluation (SemEval-2023). Association for Computational Linguistics.
- Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. 2022. NaijaSenti: A Nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.
- Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for lowresourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, Christopher D Manning, Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2020. Stanza: A {Python} natural language processing toolkit for many human languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.