zhangpeng at SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization

Peng Zhang[1,2], **Gehao Lu**[1,2]
[1]School of Information Science and Engineering, Yunnan University
[2]Yunnan Province Smart Tourism Engineering Research Center, Yunnan University
Kunming 650500, Yunnan, China
[1]zpp1219@gmail.com, [2]glu@ynu.edu.cn

Abstract

This paper presents our system developed for the SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization (Ghosh et al., 2026). On Subtask 1: Multilingual Text Classification Challenge - Polarization Detection. On Subtask 2: Multilingual Text Classification Challenge - Polarization Type Classification. On Subtask 3: Multilingual Text Classification Challenge - Manifestation Identification. For Subtask 1, we explored classical text representation approaches including Bag-of-Words, Word2Vec Average Vectors, and Bag-of-Centroids. Among these methods, the Bag-of-Centroids model achieved the best performance on both development and test datasets. For Subtask 2 and Subtask 3, we fine-tuned four different pre-trained language models: google-bert, FacebookAI-roberta, dccuchile-bert, and distilbert-multi. We experiment with 1) the training set data is analyzed visually, 2) multiple numbers of single models are trained on the training set data, and 3) multiple number of single models for voting weight ensemble learning. We further study the influence of different hyperparameters on the integrated model and select the best integration model for the prediction of the test set. On the official test set, our system achieved Macro-F1 scores of 0.6882 (EN) and 0.6711 (SP) for Subtask 1, 0.3752 (EN) and 0.6386 (SP) for Subtask 2, and 0.3561 (EN) and 0.4366 (SP) for Subtask 3. For the final ranking, organizers will use the Macro F1 score. These approaches have yielded good results.

1 Introduction

Online polarization has become a defining characteristic of contemporary digital discourse, influencing political debates, social movements, and intergroup relations. SemEval-2026 Task 9 addresses this challenge by proposing multilingual polarization detection (Naseem et al., 2026a) and analysis across 22 languages. The task consists of three subtasks: binary polarization detection, polarization target classification, and manifestation identification. The dataset covers diverse platforms and sensitive topics such as elections, conflicts, migration, and gender rights. Automatic identification of polarized content is essential for understanding online discourse dynamics and supporting moderation systems.

In this work, we focus on English and Spanish, representing both a high-resource global language and a widely spoken language with distinct sociopolitical contexts. Our system adopts different modeling strategies across subtasks. For Subtask 1, we employ a traditional Word2Vec-based approach inspired by the Kaggle. We train distributed word embeddings and construct document representations by aggregating word vectors, followed by a supervised classifier for binary prediction. This approach allows us to evaluate how classical distributional representations perform in polarization detection.

For Subtask 2 and Subtask 3, which involve multi-label classification and finer-grained semantic distinctions, we adopt transformer-based pre-trained language models. We fine-tune four different models: google-bert, FacebookAI-roberta, dccuchile-bert (Spanish-specific), and distilbert-multi. All models are adapted using a sigmoid output layer for multi-label prediction. This design enables us to compare general multilingual models with language-specific pre-trained representations, particularly for Spanish.The code of this experiment method is available on my GitHub website. [1]

2 Related Work

SemEval competition in previous years has introduced tasks focusing on multi-label text classifi-

[1]https://github.com/zpp1219/SemEval-2026-Task-9

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 95–100

cation and text binary classification (Wang et al., 2024) to evaluate internal potential elements and potential content of the text. These tasks provided datasets with human labeled similarity scores, which have been extensively utilized for training sentence embedding models and conducting semantic evaluations.

2.1 Sentence Embeddings

Word embedding models such as BERT (Devlin et al., 2019), GloVe (Pennington et al., 2014), RoBeRTa (Liu et al., 2019) and Word2Vec (Mikolov et al., 2013) are frequently employed to assess the semantic distance between words. They are also some of the more commonly used methods in text classification tasks. Sentence embeddings with a fixed length are often generated via mean/max pooling of word embeddings or employing CLS embedding in BERT. The semantic distances are commonly measured using the cosine similarity of embeddings of two expressions. Siamese or triplet network architectures are frequently employed in sentence embedding training (Reimers and Gurevych, 2019). For example, models such as Sentence-BERT utilize a dual-encoder architecture with shared weights for predicting sentence relationships (e.g., semantic contradiction, entailment, or neutral labeling) or for similarity score prediction using regression objectives, e.g., the difference between human annotated similarity score (sim) of two sentences and the cosine of two sentence embeddings.

2.2 Ensemble Learning

In previous studies, ensemble learning has demonstrated several advantages (Sagi and Rokach, 2018). The ensemble approach can reduce errors from individual models by combining results from multiple sources, thereby improving robustness. In our study, using multiple pre-trained models can also leverage the rich information learned during large-scale pre-training while maintaining computational efficiency. Prior research has shown that ensemble methods can achieve strong performance across various tasks.

In this work, we aim to integrate multiple pre-trained models to assess semantic relatedness. When models are trained on diverse datasets with different architectures, they may produce varied predictions. Combining these predictions can potentially improve overall performance. We use sentence embeddings primarily from the following models: Multilingual BERT (cased and uncased), RoBERTa, BETO (cased, uncased) (Cañete et al., 2023), DistilBERT.

3 Methodology

3.1 Overall Architecture

The proposed approach adopts a weighted voting ensemble composed of different transformer-based models. Several state-of-the-art NLP models are trained on a large dataset of annotated tweets to construct ensembles with diverse architectures and configurations. The predictions generated by these models are then combined using a weighted voting strategy to produce the final outputs. The ensemble includes Multilingual BERT (cased and uncased), RoBERTa, BETO (cased and uncased), and DistilBERT. For each instance, the final classification decision is computed as a weighted sum of the outputs of the individual models. The weighted voting mechanism assigns importance to each transformer according to its normalized performance metric (F1-score or RMSE, depending on the task), rather than relying on the simple arithmetic mean commonly used in conventional voting systems.

3.2 Implementation Details

For Subtask 1 is formulated as a binary classification task to identify whether a post contains polarized content. The data are preprocessed and represented using three feature extraction methods: Bag-of-Words, Word2Vec Average Vectors, and Bag-of-Centroids. For each representation, a supervised classifier is trained on the training set. Specifically, we employ a Random Forest classifier as the supervised learning model for Subtask 1. The classifier is implemented using the scikit-learn library, with the number of trees set to 100 and default hyperparameters for other settings. Random Forest is chosen due to its robustness to overfitting and its effectiveness in handling sparse and high-dimensional text features. The Bag-of-Centroids representation clusters word embeddings into a fixed number of centroids using k-means, and represents each document as a histogram over these clusters. This approach captures semantic groupings of words and provides a more structured representation compared to simple averaging. Model performance is evaluated on the development set using accuracy, precision, recall, and Macro F1 score. The model achieving the highest Macro F1 score is selected. The selected model is then applied to the test set

for final evaluation.

For Subtask 2 and Subtask 3, we adopt an ensemble-based transformer framework implemented using the simpletransformers Python library. To meet the library requirements, the dataset is first reformatted so that each data split contains two columns: text and labels, where labels is represented as an array corresponding to the target categories of the multi-label classification task.

We evaluate multiple transformer-based pretrained language models within a unified training framework, including google-bert, FacebookAI RoBERTa, dccuchile-bert (Spanish-specific), and distilbert-multi. All models are initialized and fine-tuned using the simpletransformers library, which enables efficient model training and evaluation with minimal implementation overhead. Each model is trained independently on the entire training set, and the trained models are stored in a structured dictionary for convenient access during ensemble construction.

After training individual models, different ensemble configurations are constructed by combining the previously trained transformers. Each ensemble is uniquely identified and stored for systematic comparison. Model evaluation is first conducted at the individual level using the validation split, where standard evaluation metrics—including accuracy, precision, recall, and F1-score—are recorded. Subsequently, ensemble predictions are generated for the validation set.

Ensemble predictions are obtained through a hard voting strategy. For each instance, predictions produced by all models in an ensemble are collected, and the final label is determined by majority voting. The voting process can be either unweighted or weighted. In the weighted setting, each model's contribution is scaled according to its normalized F1-score on the validation set, allowing stronger models to have greater influence while still preserving information from all ensemble members.

Based on validation performance, the ensemble achieving the highest F1-score is selected as the final system. This selected ensemble is then applied to the test set to generate final predictions. The resulting outputs are further analyzed and visualized using evaluation tools such as the confusion matrix and the ROC curve.

Table 1: The text experiment data situation and the number of Subtask 2: Multilingual Text Classification Challenge - Polarization Type Classification labels are described.

Training Set Text	Value
count	2676.000000
mean	13.667040
std	9.513951
min	5.000000
25%	8.000000
50%	10.000000
75%	16.000000
max	62.000000

Training Set Label	Value
political	996
racial/ethnic	264
religious	106
gender/sexual	67
other	121

4 Results and Analysis

4.1 Training Dataset Analysis

The training dataset consists of text and corresponding label annotations, as summarized in Table 1. In this analysis, we focus on Subtask 2, namely the Multilingual Text Classification Challenge for Polarization Type Classification.

Figure 1 illustrates the distribution of text lengths and the number of training instances, providing an overview of the overall length characteristics of the training data. Figure 2 presents the distribution of text lengths across different polarization type classes, showing the relative proportions of instances for each class. This class-wise analysis is conducted exclusively on Subtask 2 to better understand the variability of text length among different polarization categories.

4.2 Experimentation Configuration

For the sake of completeness and to improve the robustness of our results, each experiment was repeated six times using different combinations of hyperparameters. The hyperparameter configurations explored in our experiments are summarized in Table 2.

Specifically, we investigated two optimizers, namely AdamW and Adafactor, and three learning rates (2×10^{-5}, 4×10^{-5}, and 8×10^{-5}). All other training settings were kept identical across runs.

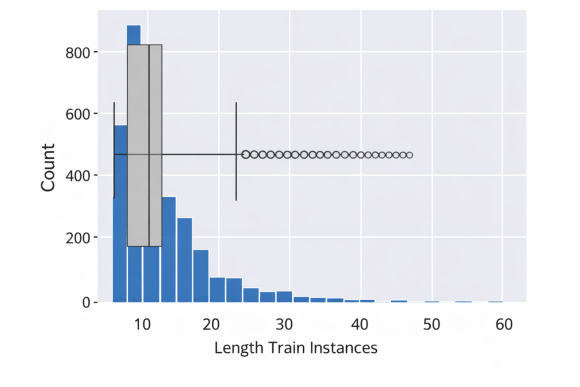

Figure 1: The length and quantity distribution of training text experiment data samples are analyzed.

Table 2: Hyperparameter configurations used in the experiments.

Hyperparameter	Values
Optimizer	AdamW, Adafactor
Learning Rate	2×10^{-5}, 4×10^{-5}, 8×10^{-5}

4.3 Development Dataset Result

Table 3 reports the official development set results of SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization, covering Subtask 1 (Polarization Detection), Subtask 2 (Polarization Type Classification), and Subtask 3 (Manifestation Identification).

For each subtask, evaluation results are provided for both English and Spanish.

4.4 Test Dataset Result

Table 4 presents the official test set results of SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization, covering Subtask 1 (Polarization Detection), Subtask 2 (Polarization Type Classification), and Subtask 3 (Manifestation Identification).

For each subtask, evaluation results are reported for both English and Spanish. The scores highlighted in bold correspond to the best-performing system on the test dataset for the respective language and subtask.

4.5 Text Length and Label Distribution Analysis

As shown in Figure 1, approximately 75% of the tweets in the training dataset contain no more than 20 words. This observation provides useful guidance for model design choices, such as selecting appropriate maximum sequence lengths and controlling model complexity.

Subtask 2 (Polarization Type Classification) is formulated as a multi-label classification problem, where each instance may be assigned zero to multiple categories (up to five). In the English and Spanish datasets considered in this study, the possible categories include political, racial/ethnic, religious, gender/sexual, and other. Table 1 reports the distribution of instances across these categories. Due to label overlap in the multi-label setting, independent per-class percentages cannot be directly interpreted, as a single instance may belong to multiple categories simultaneously.

5 Conclusion

Our system adopts an ensemble-based approach to estimate semantic relatedness, following the general motivation of combining multiple models to improve robustness and performance. For Subtask 1 (Polarization Detection), we employ traditional text representation methods to perform binary classification, demonstrating that classical approaches remain effective for identifying polarized content in short texts.

For Subtask 2 and Subtask 3, the system integrates predictions from multiple transformer-based pre-trained language models, including Multilingual BERT (cased and uncased), RoBERTa, BETO (cased and uncased), and DistilBERT. The ensemble framework combines the outputs of individual models to leverage complementary semantic representations learned during large-scale pre-training. The dataset usage for each subtask is summarized in Table 5.

Experimental results indicate that semantic relatedness can be inferred from diverse sources of information. Although certain features (e.g., lexical overlap ratio) may not perform as strongly as models explicitly designed for sentence representation learning, our findings show that combining heterogeneous features and model outputs within an ensemble framework can outperform many individual systems and achieve better agreement with human judgments on semantic relatedness of text classification (Reimers and Gurevych, 2019).

6 Limitation and Future Work

Our experiments are based on English and Spanish language datasets only. Constrained by the size of the training dataset and the availability of pre-train language models, it is regrettable that we did not offer insights into other Asian and African languages

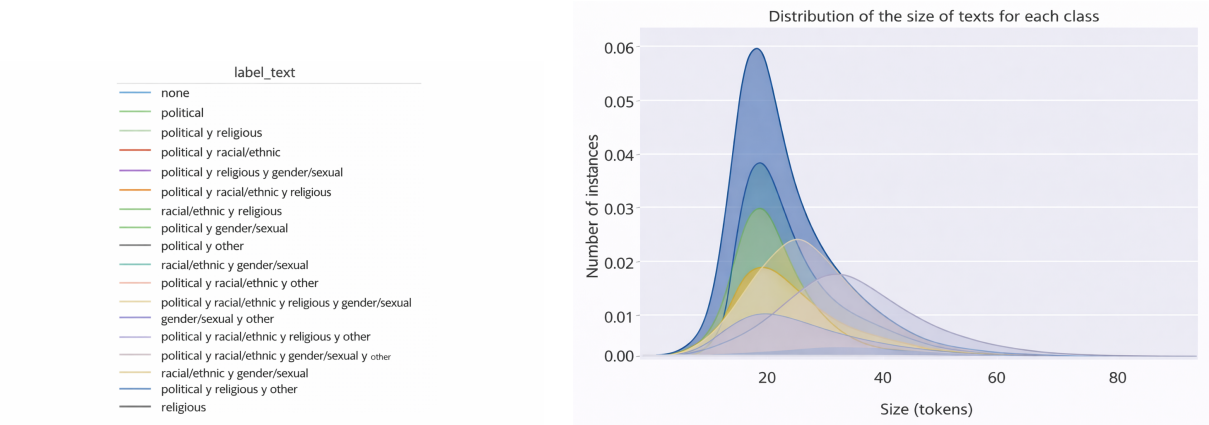

Figure 2: Distribution of the size of texts for each class.

Table 3: Development set results for all subtasks (For Subtask 2 and Subtask 3, transformer-based ensemble models are used, including multilingual BERT (cased and uncased), RoBERTa, BETO (Spanish cased and uncased), and DistilBERT-multilingual).

Subtask	Language	Method	Accuracy	Precision	Recall	F1 Binary	F1 Macro	F1 Micro	Exact Match
Subtask 1: Polarization Detection (Binary Classification)									
1	English	Bag-of-Words	0.7218	0.6750	0.5294	0.5934	0.6910	0.7218	–
1	English	Word2Vec Avg.	0.6015	0.4286	0.1176	0.1846	0.4605	0.6015	–
1	**English**	**Bag-of-Centroids**	**0.7293**	**0.6744**	**0.5686**	**0.6170**	**0.7039**	**0.7293**	**–**
1	Spanish	Bag-of-Words	0.6000	0.6286	0.5238	0.5714	0.5982	0.6000	–
1	Spanish	Word2Vec Avg.	0.4970	0.5059	0.5119	0.5089	0.4967	0.4970	–
1	**Spanish**	**Bag-of-Centroids**	**0.6667**	**0.6986**	**0.6071**	**0.6497**	**0.6659**	**0.6667**	**–**
Subtask 2: Polarization Type Classification (Multi-label)									
2	English	Transformer Ensemble	–	0.6786	0.4935	–	0.3089	0.5714	–
2	Spanish	Transformer Ensemble	–	0.6812	0.6351	–	0.6420	0.6573	–
Subtask 3: Manifestation Identification (Multi-label)									
3	English	Transformer Ensemble	–	0.5909	0.4088	–	0.4210	0.4833	0.5414
3	Spanish	Transformer Ensemble	–	0.5819	0.4769	–	0.4364	0.5242	0.4121

(Vaidya et al., 2024). In future research, studies on low-resource languages will be valuable. Future works including tasks such as data collection, annotation, and training models tailored to these languages.

Acknowledgments

We are very grateful for the assistance and discussions provided by Semeval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization leader and organizer.

References

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2023. Spanish pre-trained bert model and evaluation data. *arXiv preprint arXiv:2308.02976*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Debanjan Ghosh, Kai North, Ekaterina Kochmar, Mamoru Komachi, and Marcos Zampieri, editors. 2026. *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, San Diego, United States.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, "Ozge Alacam, Cengiz Acar"urk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, Dheeraj Kodati, Sahar Moradizeyveh, Firoj Alam, Ye Kyaw Thu, Shantipriya Parida, Ihsan Ayyub Qazi, Nelson Odhiambo Onyango, Clemencia Siro, Ibrahim Said Ahmad, Lilian Wanzare, Adem Chanie Ali, Martin Semmann,

Table 4: Test set results for all subtasks. The last column reports the Macro-F1 score of the official POLAR baseline for comparison. (For Subtask 2 and Subtask 3, transformer-based ensemble models are used, including multilingual BERT (cased and uncased), RoBERTa, BETO (Spanish cased and uncased), and DistilBERT-multilingual).

Subtask	Language	Method	Accuracy	Precision	Recall	F1 Binary	F1 Macro	F1 Micro	POLAR Baseline
Subtask 1: Polarization Detection									
1	English	Bag-of-Words	0.7149	0.6291	0.5441	0.5835	0.6834	0.7149	0.7802
1	English	Word2Vec Avg.	0.6398	0.5439	0.1163	0.1917	0.4800	0.6398	0.7802
1	**English**	**Bag-of-Centroids**	**0.7176**	**0.6300**	**0.5591**	**0.5924**	**0.6882**	**0.7176**	**0.7802**
1	Spanish	Bag-of-Words	0.6465	0.6799	0.5374	0.6003	0.6417	0.6465	0.7266
1	Spanish	Word2Vec Avg.	0.5370	0.5364	0.4612	0.4960	0.5339	0.5370	0.7266
1	**Spanish**	**Bag-of-Centroids**	**0.6815**	**0.7669**	**0.5102**	**0.6127**	**0.6711**	**0.6815**	**0.7266**
Subtask 2: Polarization Type Classification									
2	English	Transformer Ensemble	–	0.6396	0.5705	–	0.3752	0.6031	0.3333
2	Spanish	Transformer Ensemble	–	0.7090	0.5917	–	0.6386	0.6451	0.5935
Subtask 3: Manifestation Identification									
3	English	Transformer Ensemble	–	0.4772	0.3740	–	0.3561	0.4194	0.4100
3	Spanish	Transformer Ensemble	–	0.5700	0.4300	–	0.4366	0.4902	0.5088

Table 5: Use dataset supported by Semeval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization. The style is based on raw data without any manual alteration of the dataset.

Dataset Input	Description	Use or Not
(Naseem et al., 2026b)	`|POLAR, a multilingual, multicultural, and multi-event dataset with over 110K instances in 22 languages.|`	yes
other dataset	`|Use external or additional corpora.|`	no

Chris Biemann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2026a. SemEval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Kritesh Rauniyar, Tanmoy Chakraborty, Arfeen Zeeshan, Dheeraj Kodati, Satya Keerthi, Sahar Moradizeyveh, Firoj Alam, Arid Hasan, Syed Ishtiaque Ahmed, Ye Kyaw Thu, Shantipriya Parida, Ihsan Ayyub Qazi, Lilian Wanzare, Nelson Odhiambo Onyango, Clemencia Siro, Jane Wanjiru Kimani, Ibrahim Said Ahmad, Adem Chanie Ali, Martin Semmann, Chris Biemann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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 and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 8(4):e1249.

Ankit Vaidya, Aditya Gokhale, Arnav Desai, Ishaan Shukla, and Sheetal Sonawane. 2024. CLTeam1 at SemEval-2024 task 10: Large language model based ensemble for emotion detection in Hinglish. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 365–369, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.