VEL@LT-EDI-2023: Detecting Homophobia and Transphobia in Code-Mixed Spanish Social Media Comments

Prasanna Kumar Kumaresan[1], **Kishore Kumar Ponnusamy**[2],
Kogilavani Shanmugavadivel[3], **Subalalitha Chinnaudayar Navaneethakrishnan**[4]
Ruba Priyadharshini[5], **Bharathi Raja Chakravarthi**[6]
[1]Insight SFI Research Centre for Data Analytics, University of Galway, Ireland
[2]Guru Nanak College, Chennai, Tamil Nadu, India. [3]Kongu Engineering College, Tamil Nadu, India
[4]Department Of Computer Science & Engineering, SRM Institute Of Science And Technology, Tamil Nadu, India
[5] Gandhigram Rural Institute-Deemed to be University, Tamil Nadu, India
[6] Insight SFI Research Centre for Data Analytics, School of Computer Science, University of Galway, Ireland.
prasanna.kumaresan@insight-centre.org
{kishorep16002, kogilavani.sv}@gmail.com
{subalalitha, rubapriyadharshini.a}@gmail.com
bharathi.raja@universityofgalway.ie

Abstract

Our research aims to address the task of detecting homophobia and transphobia in social media code-mixed comments written in Spanish. Code-mixed text in social media often violates strict grammar rules and incorporates non-native scripts, posing challenges for identification. To tackle this problem, we perform pre-processing by removing unnecessary content and establishing a baseline for detecting homophobia and transphobia. Furthermore, we explore the effectiveness of various traditional machine-learning models with feature extraction and pre-trained transformer model techniques. Our best configurations achieve weighted F1 scores of 0.86 on the test set and 0.86 on the development set for Spanish, demonstrating promising results in detecting instances of homophobia and transphobia in code-mixed comments.

1 Introduction

Hate speech is speech that is directly or indirectly against a person or group and contains animosity because of something inherent to that person or group (Schmidt and Wiegand, 2017; Chetty and Alathur, 2018; Fortuna and Nunes, 2018). Locating hate speech on the Internet is a difficult task that even the most advanced models struggle to complete (Govers et al., 2023; Chakravarthi et al., 2023a). As a result of the rapid development in user-generated online content, which has not only resulted in a vast increase in the accessibility of information but has also delivered a vast increase in the accessibility of information (Subramanian et al., 2022a), individuals have been provided with a simple platform on which to express their opinions and communicate with others in a public forum (Jahan and Oussalah, 2023; Chakravarthi, 2022a). This has resulted in some undesirable uses of online spaces, such as the dissemination of hate speech, which is regrettable. The use of abusive language frequently accompanies the dissemination of hate speech in everyday life, especially on social media (Chakravarthi et al., 2023c; Subramanian et al., 2022b; Chakravarthi, 2022b).

In studies conducted under the headings of hate speech, offensive language, and aggressive language, the examination of homophobic language is typically grouped with analyses of other forms of hostility (Waseem and Hovy, 2016; Espinosa Anke et al., 2019; Priyadharshini et al., 2022). ”Emotional disgust towards individuals who do not conform to society’s gender expectations” is one definition of transphobia (Nagoshi et al., 2008). Homophobia is the unreasonable fear, loathing, and intolerance of homosexual men and women in close proximity (Chakravarthi, 2023). Typically, hate speech detection models are evaluated by measuring how well they perform on data set aside for testing. Most of the evaluation, accuracy, and F1 score are used as metrics (Chakravarthi et al., 2021; Santhiya et al., 2022; Priyadharshini et al., 2022). The overview paper (Chakravarthi et al., 2023b), described the overall descriptions of the participants participated and the dataset of the shared task on Homophobia and Transphobia Detection in so-

LT-EDI 2023 – Third Workshop on Language Technology for Equality, Diversity and Inclusion, pages 233–238, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-084-7_035

cial media comments.

We participated in Task A for Spanish, which focused on the detection of Homophobia and Transphobia in social media comments. The task was organized by LT-EDI@RANLP-2023. With the provided dataset, we developed machine learning models using feature extraction as baselines, as well as the MuRIL transformer model. Among our models, the MuRIL model yielded the best results, achieving a weighted F1 score of 0.86. These scores indicate the effectiveness of our approach in accurately identifying instances of Homophobia and Transphobia in Spanish social media comments. Our participation in this shared task has provided valuable insights into the detection and understanding of discriminatory behavior in online platforms.

2 Related Work

Researchers examined the linguistic behaviors of homosexual individuals in China by compiling a corpus of their texts (Espinosa Anke et al., 2019). (Chakravarthi et al., 2022a) created fine-grained taxonomy for homophobia and transphobia for English and Tamil languages. (Chakravarthi et al., 2022b) conducted a shared task to the identification of homophobia, transphobia, and non-anti-LGBT+ content from the given corpus. This task was centered on three subtasks for the Tamil, English, and Tamil-English (code-mixed) languages. It received 10 Tamil systems, 13 English systems, and 11 Tamil-English systems. The average macro F1-score for the top systems for Tamil, English, and Tamil-English was 0.570, 0.877, and 0.610, respectively.

(Chinnaudayar Navaneethakrishnan et al., 2022) conducted sentiment analysis and homophobia detection shared task in code-mixed Dravidian language YouTube comments for Tamil, Malayalam and English. At FIRE 2022 the DravidianCodeMix organized task A for detecting sentiment analysis and task B for detecting homophobia. 95 individuals signed up for the shared task, 13 teams submitted their results for task-A a, and 10 teams submitted their results for task B. Traditional machine learning and deep learning models were used to investigate tasks A and B.

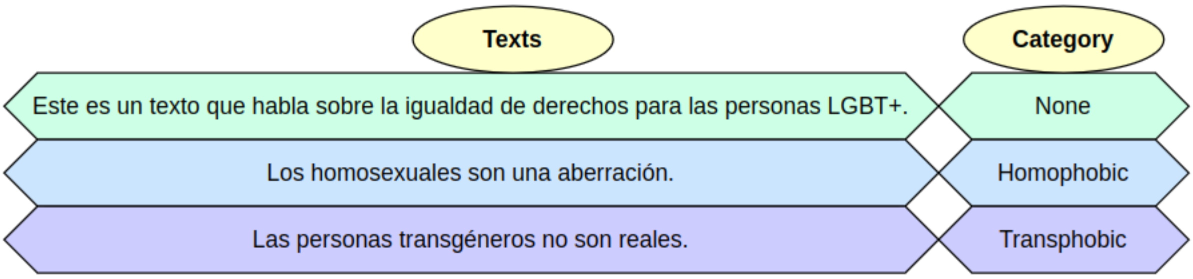
Transphobic and homophobic insults directed at LGBTQI+ persons for the shared task have been identified using transformer-based model methodologies such as BERT and XLMROBERTa models by (Manikandan et al., 2022). BERT offers 91%, while XLM-RoBERTa offers 93%. The content was predicted using the IndicBERT and LaBSE machine learning models. The following were the results: IndicBERT was utilized to train Tamil, Malayalam, and Tamil-English, whereas LaBSE was utilized to predict the English content. The weighted average F1 scores for English, Malayalam, Tamil-English, and Tamil were 0.46, 0.54, 0.39, and 0.28, respectively by (Pranith et al., 2022). (Varsha et al., 2022) participated in both sentiment analysis and homophobia detection tasks. Under the feature extraction techniques of Count Vectorizer and TF-IDF, pre-trained models such as BERT, XLM, and MPNet were used alongside classifiers such as SVM, MLP, and Random Forest. The rankings for sentiment analysis assignment are rank 1 in the Tamil dataset, rank 6 in the Malayalam dataset, and rank 7 in the Kannada dataset. The sentiment analysis task in the Malayalam dataset yielded the highest F1 score of 0.63, while the homophobia detection task yielded 0.95. Various machine learning algorithms are contrasted with the proposed system's performance (Shanmugavadivel et al., 2022; Kumaresan et al., 2022).

We discuss the existing research in the field of text classification, particularly focusing on the specific context of Indian languages. We implemented the MuRIL (Multilingual Representations for Indian Languages)(Khanuja et al., 2021) pre-trained transformer model, which we utilize in our study. MuRIL, available through the Hugging Face model repository, is specifically designed to handle the linguistic complexities and nuances of Indian languages, enabling effective text classification tasks. While previous works have explored various approaches for text classification in Indian languages, our paper distinguishes itself by leveraging the MuRIL model and fine-tuning it on a diverse range of downstream tasks. By highlighting the advantages of MuRIL and showcasing the results of our fine-tuning experiments, we contribute to the growing body of research focused on enhancing the performance of text classification in Indian languages.

3 Task and Dataset Description

This research paper discusses our participation in the shared task Homophobia/Transphobia Detection in social media comments[1], which was or-

[1] https://codalab.lisn.upsaclay.fr/competitions/11077



Texts	Category
Este es un texto que habla sobre la igualdad de derechos para las personas LGBT+.	None
Los homosexuales son una aberración.	Homophobic
Las personas transgéneros no son reales.	Transphobic

Figure 1: Examples comments from the datasets in Spanish

ganized by LT-EDI@RANLP-2023. The objective of this task was to identify instances of homophobia and transphobia in social media comments across multiple languages, including English, Hindi, Tamil, Spanish, and Malayalam. Our focus was specifically on Task A, which involved detecting these forms of discrimination in Spanish comments using a 3-class classification approach. The dataset provided for this task consisted of 1586 social media comments, each annotated as Homophobic, Transphobic, or None. We divided this dataset into a training set of 850 comments, a development set of 236 comments, and a test set of 500 comments. The class distribution for each set is presented in Table 1. Further details about the dataset can be found in the study by (Chakravarthi et al., 2022a). The shared task consisted of three phases: in the first phase, a training and development set was provided to train our model; in the second phase, only the test set comments were released, and we were required to make predictions using the model trained in the first phase; finally, in the last phase, the test set with labels was released to assess the performance of our model. We will submit our predictions based on the test set comments to the organizers for evaluation.

Table 1: Data statistics for Spanish in Task A

Category	Train	Test	Dev
None	450	300	150
Homophobic	200	100	43
Transphobic	200	100	43
Total	**850**	**500**	**236**

4 Methodology

The methodology section outlines the step-by-step process we employed to identify instances of homophobia and transphobia in code-mixed text in the Spanish language. This involved utilizing feature extraction with machine learning and transformer-based approaches for text classification.

4.1 Machine learning

In this task, we employed traditional machine learning models as our baseline, along with CountVectorizer[2] feature extraction. Before proceeding with the models, we executed essential preprocessing steps, which involved removing tags, punctuation, URLs, and other unwanted elements. Additionally, we converted the labels into numerical values using a LabelEncoder[3]. To facilitate effective machine learning, we utilized the CountVectorizer technique to transform the text data into vectorized representations, which would be conducive to the performance of our models. Consequently, we implemented several popular algorithms including Naive Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR), Decision Trees (DT), and Random Forests (RF), by leveraging the sci-kit-learn library[4]. By combining these steps, we were able to establish a strong foundation for our machine-learning approach using traditional models and CountVectorizer extraction.

4.2 Transformers

We utilized the MuRIL (Multilingual Representations for Indian Languages) pre-trained transformer model[5], trained on BERT Large (24L) with 17 Indian languages. Categorical labels were encoded using LabelEncoder from sci-kit-learn. The Hugging Face library trained a transformer model,

[2] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

[3] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

[4] https://scikit-learn.org/stable/index.html

[5] https://huggingface.co/google/muril-large-cased

”google/muril-large-cased”[6]. Tokenizer encoded train and test texts into tensors. A TensorDataset organized encoded data, and a data loader batched the data. The model was trained on a preferred device, optimized using AdamW, and adjusted learning rate with a scheduler. After training, the model was evaluated in the test mode. Predictions were generated using test data, and the highest probability class was chosen. Classification report printed with precision, recall, F1-score, and support. Confusion matrix computed using sci-kit-learn’s confusion_matrix function. Visualization was created using Matplotlib and seaborn, representing correctly and incorrectly classified samples. This approach provided insights into the performance of the MuRIL transformer model for the task.

5 Results and Discussion

In this section, we evaluated the results of various models used to detect homophobia and transphobia in the Spanish language. The performance of these models was assessed using metrics such as Accuracy (ACC), Macro Precision (MP), Macro Recall (MR), Macro F1 (MF1), Weighted Precision (WP), Weighted Recall (WR), and Weighted F1 (WF1) scores. We experimented with five machine learning models, including NB, SVM, LR, DT, and RF utilizing CountVectorizer feature extraction. The weighted F1 scores obtained for these models were 0.60, 0.78, 0.82, 0.79, and 0.77, respectively.

Next, we explored the performance of a large language model, specifically the MuRIL large cased model, which was originally pre-trained on Indian languages. However, we adapted it for detecting homophobia and transphobia in the Spanish language. Comparing the results in Table 3, it was evident that the pre-trained transformer model outperformed the other models, achieving a weighted F1 score of 0.84 on the test set and 0.82 on the development set shown int he Table 2. This higher score indicates its superior performance in classifying instances of homophobia and transphobia. To gain further insights, we visualized the model’s predictions using a confusion matrix, which is shown in Figure 2. This visualization provides a clear representation of how well the best model performed for each class, demonstrating its ability to correctly identify instances of homophobia and transphobia. Overall, based on the evaluation metrics and the clear performance superiority of the pre-trained transformer model, we selected it as the best model for detecting homophobia and transphobia in the Spanish language.

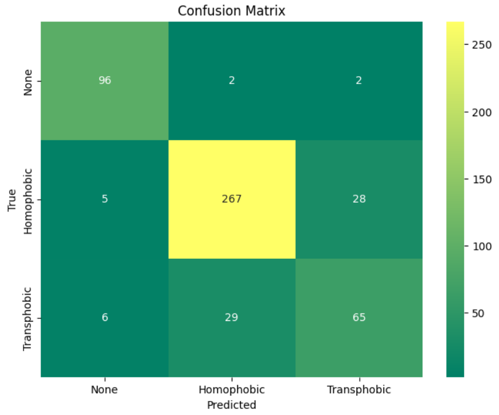


Figure 2: Confusion matrix for MuRIL transformer model

6 Conclusion

In this study, we examined the detection of homophobia and transphobia in the Spanish language using machine learning models and a pre-trained transformer model. Among the traditional models with CountVectorizer feature extraction, the MuRIL pre-trained transformer model outperformed them with a weighted F1 score of 0.84 on the test set. The transformer model’s superior performance demonstrates its effectiveness in classifying instances of homophobia and transphobia. The confusion matrix visualization further supported the model’s ability to correctly identify such instances. Consequently, we conclude that the pre-trained transformer model is a suitable choice for this task, offering the potential for addressing social issues and promoting inclusivity in online spaces. Future research can explore fine-tuning techniques and larger datasets to enhance the model’s performance.

References

Bharathi Raja Chakravarthi. 2022a. Hope speech detection in youtube comments. *Social Network Analysis and Mining*, 12(1):75.

Bharathi Raja Chakravarthi. 2022b. Multilingual hope speech detection in english and dravidian languages. *International Journal of Data Science and Analytics*, 14(4):389–406.

[6] https://huggingface.co/google/muril-large-cased

Table 2: Develpment set results for Homophobia and Transphobia in Spanish

Models	**ACC**	**MP**	**MR**	**MF1**	**WP**	**WR**	**WF1**
NB	0.78	0.79	0.65	0.69	0.78	0.78	0.76
SVM	0.82	0.80	0.77	0.78	0.82	0.82	0.82
LR	0.82	0.78	0.80	0.79	0.83	0.82	0.82
DT	0.81	0.76	0.79	0.78	0.83	0.81	0.82
RF	0.85	0.81	0.81	0.81	0.84	0.85	0.85
MuRIL	**0.86**	**0.81**	**0.84**	**0.82**	**0.86**	**0.86**	**0.86**

Table 3: Test set results for Homophobia and Transphobia in Spanish

Models	**ACC**	**MP**	**MR**	**MF1**	**WP**	**WR**	**WF1**
NB	0.74	0.77	0.60	0.62	0.75	0.74	0.70
SVM	0.82	0.80	0.77	0.78	0.82	0.82	0.81
LR	0.84	0.81	0.83	0.82	0.84	0.84	0.84
DT	0.82	0.79	0.80	0.79	0.82	0.82	0.82
RF	0.82	0.81	0.76	0.77	0.82	0.82	0.81
MuRIL	**0.86**	**0.83**	**0.85**	**0.84**	**0.86**	**0.86**	**0.86**

Bharathi Raja Chakravarthi. 2023. Detection of homophobia and transphobia in YouTube comments. *International Journal of Data Science and Analytics*.

Bharathi Raja Chakravarthi, Adeep Hande, Rahul Ponnusamy, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022a. How can we detect homophobia and transphobia? experiments in a multilingual code-mixed setting for social media governance. *International Journal of Information Management Data Insights*, 2(2):100119.

Bharathi Raja Chakravarthi, Manoj Balaji Jagadeeshan, Vasanth Palanikumar, and Ruba Priyadharshini. 2023a. Offensive language identification in dravidian languages using mpnet and cnn. *International Journal of Information Management Data Insights*, 3(1):100151.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Malliga Subramanian, Paul Buitelaar, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, Rafael Valencia-García, and Nitesh Jindal. 2023b. Overview of second shared task on homophobia and transphobia detection in english, spanish, hindi, tamil, and malayalam. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023c. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022b. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv. Association for Computational Linguistics.

Naganna Chetty and Sreejith Alathur. 2018. Hate speech review in the context of online social networks. *Aggression and violent behavior*, 40:108–118.

Subalalitha Chinnaudayar Navaneethakrishnan, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Malliga Subramanian, Prasanna Kumar Kumaresan, Bharathi, Lavanya Sambath Kumar, and Rahul Ponnusamy. 2022. Findings of shared task on sentiment analysis and homophobia detection of youtube comments in code-mixed dravidian languages. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 18–21.

Luis Espinosa Anke, Thierry Declerck, Dagmar Gromann, Ziqi Zhang, Lei Luo, Dagmar Gromann, Luis Espinosa Anke, and Thierry Declerck. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semant. Web*, 10(5):925–945.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Jarod Govers, Philip Feldman, Aaron Dant, and Panos Patros. 2023. Down the rabbit hole: Detecting online extremism, radicalisation, and politicised hate speech. *ACM Computing Surveys*.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, page 126232.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Elizabeth Sherly, Sangeetha Sivanesan, and Bharathi Raja Chakravarthi. 2022. Transformer based hope speech comment classification in code-mixed text. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 120–137. Springer.

Deepalakshmi Manikandan, Malliga Subramanian, and Kogilavani Shanmugavadivel. 2022. A system for detecting abusive contents against lgbt community using deep learning based transformer models. In *Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation (Hybrid). CEUR*.

Julie L Nagoshi, Katherine A Adams, Heather K Terrell, Eric D Hill, Stephanie Brzuzy, and Craig T Nagoshi. 2008. Gender differences in correlates of homophobia and transphobia. *Sex roles*, 59:521–531.

P Pranith, V Samhita, D Sarath, and Durairaj Thenmozhi. 2022. Homophobia and transphobia detection of youtube comments in code-mixed dravidian languages using deep learning.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

S Santhiya, P Jayadharshini, and SV Kogilavani. 2022. Transfer learning based youtube toxic comments identification. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 220–230. Springer.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Prasath Mahalingam, Malliga Subramanian, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. An analysis of machine learning models for sentiment analysis of tamil code-mixed data. *Computer Speech & Language*, 76:101407.

Malliga Subramanian, Ramya Chinnasamy, Prasanna Kumar Kumaresan, Vasanth Palanikumar, Madhoora Mohan, and Kogilavani Shanmugavadivel. 2022a. Development of multi-lingual models for detecting hope speech texts from social media comments. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 209–219. Springer.

Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022b. Offensive language detection in tamil youtube comments by adapters and cross-domain knowledge transfer. *Computer Speech & Language*, 76:101404.

Josephine Varsha, B Bharathi, and A Meenakshi. 2022. Sentiment analysis and homophobia detection of youtube comments in code-mixed dravidian languages using machine learning and transformer models. In *Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation (Hybrid). CEUR*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.