IIITG-ADBU@HASOC-Dravidian-CodeMix-FIRE2020: Offensive Content Detection in Code-Mixed Dravidian Text

Arup Baruah[a], Kaushik Amar Das[a], Ferdous Ahmed Barbhuiya[a] and Kuntal Dey[b]

[a] *Indian Institute of Information Technology, Guwahati, India*
[b] *Accenture Technology Labs, Bangalore, India*

Abstract

This paper presents the results obtained by our SVM and XLM-RoBERTa based classifiers in the shared task "Dravidian-CodeMix-HASOC 2020". The SVM classifier trained using TF-IDF features of character and word n-grams performed the best on the code-mixed Malayalam text. It obtained a weighted F1 score of 0.95 (1st Rank) and 0.76 (3rd Rank) on the YouTube and Twitter dataset respectively. The XLM-RoBERTa based classifier performed the best on the code-mixed Tamil text. It obtained a weighted F1 score of 0.87 (3rd Rank) on the code-mixed Tamil Twitter dataset.

Keywords

SVM, XLM-RoBERTa, Offensive Language, Code-Mixed, Dravidian Language

1. Introduction

The use of offensive language in social media text has become a new social problem. Such language can have a negative psychological impact on the readers. It can have adverse effect on the emotion and behavior of people. Hate speech has fueled riots in many places around the world. As such, it is important to keep social media free from offensive language. Considerable research has been performed on automated techniques for detecting offensive language. Among the many challenges that such systems have to tackle, the use of code-mixed text is another. Code-mixing is the phenomena of mixing words from more than one language in the same sentence or between sentences.

The shared task "Dravidian-CodeMix-HASOC 2020" [1, 2] is an attempt to promote research on offensive language detection in code-mixed text. This shared task is held as a sub-track of "Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC)" at FIRE-2020. The shared task had two tasks. Task 1 required detection of offensive language in code-mixed Malayalam-English text from YouTube. Task 2 required detection of offensive language in code-mixed Tamil-English and Malayalam-English tweets. Both the tasks were binary classification problem where it was required to determine if the given text is offensive or not.

FIRE 2020: Forum for Information Retrieval Evaluation, December 16–20, 2020, Hyderabad, India

arup.baruah@gmail.com (A. Baruah); kaushikamardas@gmail.com (K. A. Das); ferdous@iiitg.ac.in (F. A. Barbhuiya); kuntal.dey@accenture.com (K. Dey)

© 2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

Table 1
Data set statistics

Label	Task 1 - Malayalam			Task 2 - Tamil		Task 2 - Malayalam	
	Train	Dev	Test	Train	Test	Train	Test
NOT	2633 (82.3%)	328 (82%)	334 (83.5%)	2020 (50.5%)	465 (49.5%)	2047 (51.2%)	488 (48.8%)
OFF	567 (17.7%)	72 (18%)	66 (16.5%)	1980 (49.5%)	475 (50.5%)	1953 (48.4%)	512 (51.2%)
Total	3200	400	400	4000	940	4000	1000

We participated in both the tasks. We used SVM and XLM-RoBERTa classifiers in our study. The SVM classifier was trained using TF-IDF features of character n-grams, word n-grams, and character and word n-grams combined.

2. Related Work

Offensive language detection in English has witnessed the use of SVM [3, 4, 5, 6, 7], Logistic Regression [8, 9, 10, 6, 11], and deep learning techniques [12, 13, 14, 15, 16, 17]. The main focus of [5] was to tackle the use of code words for obfuscating the hate words. Traditional machine learning and deep learning techniques have also been used in the detection of offensive language in code-mixed Hindi-English text [18, 19, 20, 21, 22, 23, 24]. Work performed on code-mixed Tamil-English and Malayalam-English text includes corpus created for sentiment analysis for these two languages [25, 26]. [27] focused on machine translation of code-mixed text in Dravidian languages. It was found that removal of code-mixing improves the quality of machine translation.

3. Dataset

Table 1 shows the statistics of the dataset provided as part of this shared task. The instances in the dataset were labeled as "not offensive" (NOT) or "offensive" (OFF). Task 1 was conducted for Malayalam language only. The source of the dataset for this task was YouTube. As can be seen from the table, this dataset is imbalanced with about 83% labeled as NOT. Task 2 was conducted for both Tamil and Malayalam languages. The source of the datasets for this task was Twitter. As can be seen from the tables, the dataset for this task was balanced. Train, development, and test set was provided for Task 1. For task 2, only train and test set was provided. We created the development set for Task 2, by doing a stratified split and retaining 85% of the dataset for training and 15% as development dataset.

Table 2
Dev Set Results

Task	System	Precision (Weighted)	Recall (Weighted)	F1 (Weighted)
Task 1 Malayalam	SVM (char)	0.9187	0.9175	0.9096
Task 1 Malayalam	SVM (word)	0.9138	0.9075	0.8950
Task 1 Malayalam	SVM (char + word)	**0.9330**	**0.9325**	0.9278
Task 1 Malayalam	XLM-RoBERTa	0.9305	**0.9325**	**0.9307**
Task 2 Tamil	SVM (char)	0.8650	0.8633	0.8630
Task 2 Tamil	SVM (word)	**0.8733**	**0.8717**	**0.8714**
Task 2 Tamil	SVM (char + word)	0.8617	0.8600	0.8597
Task 2 Tamil	XLM-RoBERTa	0.8651	0.8650	0.8650
Task 2 Malayalam	SVM (char)	0.7519	0.7500	0.7490
Task 2 Malayalam	SVM (word)	0.7190	0.7100	0.7056
Task 2 Malayalam	SVM (char + word)	**0.7630**	**0.7617**	**0.7610**
Task 2 Malayalam	XLM-RoBERTa	0.5732	0.5483	0.5171

4. Methodology

In this study we used SVM and XLM-RoBERTa based classifiers. The SVM classifier was trained using TF-IDF features of character n-grams, word n-grams, and combination of character and word n-grams. In our study, we used character n-grams of size 1 to 6, and word n-grams of size 1 to 3.

XLM-RoBERTa model [28] is based on the RoBERTa model [29]. RoBERTa model is based on the transformer architecture. XLM-RoBERTa is a multi-lingual model trained on 100 different languages including Tamil and Malayalam. In our study, we used the pre-trained base model. The Adam optimizer with weight decay was used during training. The learning rate and epsilon parameter for the optimizer were set to 2e-5 and 1e-8 respectively. We used the class provided by HuggingFace Transformers library [1] for sequence classification in our study. This class provides a linear layer on top of the pooled output to perform the binary classification.

5. Results

Table 2 shows the results obtained by our SVM and XLM-RoBERTa classifiers on the development set. For task 1, the development set was provided as part of the dataset. For task 2, the development set was created by performing a stratified split on the train set. 15% of the train set was set aside as the development set. The XLM-RoBERTa classifier performed the best with a weighted F1 score of 0.9307 in the development set for task 1. Among the SVM classifiers, the one trained using the combination of TF-IDF features of character and word n-grams performed the best with a weighted F1 score of 0.9278.

In task 2 dev set, the SVM classifier trained using the TF-IDF features of word n-grams performed the best for code-mixed Tamil-English text. It obtained a weighted F1 score of 0.8714.

[1] https://huggingface.co/transformers/

Table 3
Test Set Results

Task	System	Precision (Weighted)	Recall (Weighted)	F1 (Weighted)	Rank
Task 1 Malayalam	SVM (char + word)	**0.9505**	**0.9500**	**0.9471**	1st
Task 1 Malayalam	XLM-RoBERTa	0.9241	0.9250	0.9245	-
Task 2 Tamil	SVM (word)	0.8524	0.8521	0.8520	-
Task 2 Tamil	XLM-RoBERTa	**0.8680**	**0.8670**	**0.8669**	3rd
Task 2 Malayalam	SVM (char + word)	**0.7686**	**0.7630**	**0.7623**	3rd
Task 2 Malayalam	XLM-RoBERTa	0.6181	0.5800	0.5337	-

Table 4
Confusion Matrices of the submitted classifiers on the Test Set

	Task 1 (Malayalam)				**Task 2 (Tamil)**				**Task 2 (Malayalam)**			
	SVM char+word pred		**XLM-RoBERTa pred**		**SVM word pred**		**XLM-RoBERTa pred**		**SVM char+word pred**		**XLM-RoBERTa pred**	
	NOT	**OFF**	**NOT**	**OFF**	**NOT**	**OFF**	**NOT**	**OFF**	**NOT**	**OFF**	**NOT**	**OFF**
NOT	332	2	320	14	389	76	390	75	403	85	127	361
OFF	18	48	16	50	63	412	50	425	152	360	59	453

The XLM-RoBERTa classifier obtained a weighted F1 score of 0.8650 and was the second best performing classifier on the dev set for this task. For code-mixed Malayalam-English text of the task 2 dev set, the best performing classifier was the SVM classifier trained using the combination of TF-IDF features of character and word n-grams. It obtained a weighted F1 score of 0.7610. The XLM-RoBERTa classifier obtained a weighted F1 score of 0.5171 and was the worst performing classifier for this task.

Table 3 shows the results that our submitted classifiers obtained on the test set. The SVM classifiers mentioned in this table are the only one submitted for the tasks. These classifiers were selected based on their performance on the development set. As can be seen from the table, the SVM classifier trained on the combination of TF-IDF features of character and word n-grams performed the best in task 1 with as weighted F1 score of 0.9471. It obtained the 1st rank for the task. XLM-RoBERTa was the best performing classifier for the Tamil-English dataset of task 2. It was a weighted F1 score of 0.8669 and obtained the 3rd rank for the task. The SVM classifier trained on the combination of TF-IDF features of character and word n-grams again performed the best for the Malayalam-English dataset of task 2 with a weighted F1 score of 0.7623. It obtained the 3rd rank for the task. Table 4 shows the confusion matrices obtained on the test set by classifiers submitted for the shared task.

6. Conclusion

We used the SVM and XLM-RoBERTa based classifiers to detect offensive language in code-mixed Tamil-English and Malayalam-English text. In our study, the SVM classifier trained using combination of TF-IDF features of character and word n-grams performed the best for code-mixed Malayalam-English text (both YouTube and Twitter dataset). This classifier obtained the weighted F1 score of 0.95 (1st rank) and 0.76 (3rd rank) for Task 1 and Task 2 (Malayalam) respectively. The XLM-RoBERTa based classifier performed the best for the Tamil-English dataset of Task 2 and obtained an weighted F1 score of 0.87 (3rd rank) for the task. On comparing the performance of our SVM models on the YouTube and Twitter data for Malayalam language, we can observe that the performance of the classifier degraded considerably for the Twitter dataset. Whether this degradation is due to the type of language used in Twitter conversation, length of the text etc. can be performed as a future study.

Acknowledgments

Supported by Visvesvaraya PhD Scheme, MeitY, Govt. of India, MEITY-PHD-3050.

References

[1] B. R. Chakravarthi, M. A. Kumar, J. P. McCrae, P. B, S. KP, T. Mandl, Overview of the track on "hasoc-offensive language identification- dravidiancodemix", in: Proceedings of the 12th Forum for Information Retrieval Evaluation, FIRE '20, 2020.

[2] B. R. Chakravarthi, M. A. Kumar, J. P. McCrae, P. B, S. KP, T. Mandl, Overview of the track on "hasoc-offensive language identification- dravidiancodemix", in: Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2020). CEUR Workshop Proceedings. In: CEUR-WS. org, Hyderabad, India, 2020.

[3] S. Malmasi, M. Zampieri, Detecting Hate Speech in Social Media, in: RANLP 2017, Varna, Bulgaria, 2017, pp. 467–472.

[4] S. Malmasi, M. Zampieri, Challenges in discriminating profanity from hate speech, Journal of Experimental & Theoretical Artificial Intelligence 30 (2018) 187–202.

[5] R. Magu, K. Joshi, J. J.Luo, Detecting the Hate Code on Social Media, in: AAAI ICWSM 2017, Montreal, 2017, pp. 608–611.

[6] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated Hate Speech Detection and the Problem of Offensive Language, in: AAAI ICWSM 2017, Montreal, 2017, pp. 512–515.

[7] N. Samghabadi, S. Maharjan, A. Sprague, R. Diaz-Sprague, T. Solorio, Detecting Nastiness in Social Media, in: ALW1 at ACL 2017, Vancouver, 2017, pp. 63–72.

[8] E. Wulczyn, N. Thain, L. Dixon, Ex Machina: Personal Attacks Seen at Scale, in: WWW 2017, Perth, 2017, pp. 1391–1399.

[9] Z. Waseem, D. Hovy, Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter, in: NAACL-HLT 2016, California, 2016, pp. 88–93.

[10] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, N. Bhamidipati, Hate Speech Detection with Comment Embeddings, in: WWW 2015, Florence, Italy, 2015, pp. 29–30.

[11] J. Risch, R. Krestel, Delete or not Delete? Semi-Automatic Comment Moderation for the Newsroom, in: TRAC-1 at COLING 2018, Santa Fe, USA, 2018, pp. 166–176.

[12] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep Learning for Hate Speech Detection in Tweets, in: WWW 2017, Perth, 2017, pp. 759–760.

[13] B. Gamback, U. Sikdar, Using Convolutional Neural Networks to Classify Hate-Speech, in: ALW1 at ACL 2017, Vancouver, 2017, pp. 85–90.

[14] J. Park, P. Fung, One-step and Two-step Classification fro Abusive Language Detection on Twitter, in: ALW1 at ACL 2017, Vancouver, 2017, pp. 41–45.

[15] J. Pavlopoulos, P. Malakasiotis, I. Androutsopoulos, Deep Learning for User Comment Moderation, in: ALW1 at ACL 2017, Vancouver, 2017a, pp. 25–35.

[16] Y. Mehdad, J. Tetreault, Do Characters Abuse More Than Words?, in: SIGDIAL 2016, Los Angeles, 2016, pp. 299–303.

[17] A. Baruah, F. A. Barbhuiya, K. Dey, ABARUAH at semeval-2019 task 5 : Bi-directional LSTM for hate speech detection, in: Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019, 2019, pp. 371–376.

[18] T. Y. S. S. Santosh, K. V. S. Aravind, Hate speech detection in hindi-english code-mixed social media text, in: Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, COMAD/CODS 2019, Kolkata, India, January 3-5, 2019, ACM, 2019, pp. 310–313.

[19] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, M. Shrivastava, A dataset of hindi-english code-mixed social media text for hate speech detection, in: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, PEOPLES@NAACL-HTL 2018, New Orleans, Louisiana, USA, June 6, 2018, Association for Computational Linguistics, 2018, pp. 36–41.

[20] S. Kamble, A. Joshi, Hate speech detection from code-mixed hindi-english tweets using deep learning models, CoRR abs/1811.05145 (2018).

[21] K. Sreelakshmi, B. Premjith, K. Soman, Detection of hate speech text in hindi-english code-mixed data, in: Proceedings of the 3rd International Conference on Computing and Network Communications, 2019, India, Dec 18–21, 2019, Elsevier B.V., 2020, pp. 737–744.

[22] P. Mathur, R. R. Shah, R. Sawhney, D. Mahata, Detecting offensive tweets in hindi-english code-switched language, in: Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, SocialNLP@ACL 2018, Melbourne, Australia, July 20, 2018, Association for Computational Linguistics, 2018, pp. 18–26.

[23] A. Baruah, F. A. Barbhuiya, K. Dey, IIITG-ADBU at HASOC 2019: Automated hate speech and offensive content detection in english and code-mixed hindi text, in: Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019, volume 2517 of *CEUR Workshop Proceedings*, 2019, pp. 229–236.

[24] A. Baruah, K. A. Das, F. A. Barbhuiya, K. Dey, Aggression identification in english, hindi and bangla text using bert, roberta and SVM, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, TRAC@LREC 2020, Marseille, France, May 2020, European Language Resources Association (ELRA), 2020, pp. 76–82.

[25] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, J. P. McCrae, Corpus creation for sentiment analysis in code-mixed Tamil-English text, in: Proceedings of the 1st Joint

Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), European Language Resources association, Marseille, France, 2020, pp. 202–210.

[26] B. R. Chakravarthi, N. Jose, S. Suryawanshi, E. Sherly, J. P. McCrae, A sentiment analysis dataset for code-mixed Malayalam-English, in: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), European Language Resources association, Marseille, France, 2020, pp. 177–184.

[27] B. R. Chakravarthi, Leveraging orthographic information to improve machine translation of under-resourced languages, Ph.D. thesis, NUI Galway, 2020.

[28] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://www.aclweb.org/anthology/2020.acl-main.747/.

[29] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.