Deep Learning Miniproject on Hate Speech Detection in Translated Text: Realted Work

Romino Steiner, Marc Schenk, Moritz Widmer

April 9, 2024

Abstract

This document contains a short summary of the contents and approaches used in various research papers that tackle hate speech detection and classification in translated texts.

1 Cohen et al. - 2023 - Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time [1]

1.1 Summary

This study presents an ensemble approach that leverages DeBERTa models, integrating back-translation (BT) and GPT-3 augmentation techniques during both training and test times. The proposed method significantly enhances hate speech detection performance across various metrics and datasets. For reproducibility and further research, the code is publicly available.

1.2 Explored Problem

The research tackles the challenge of accurately detecting hate speech on social media platforms. This is difficult due to the use of slang, implicit hate speech, and the ever-evolving nature of language online. The study aims to improve hate speech detection models' performance, making them more robust against these challenges.

1.3 Approach

The approach combines DeBERTa models with BT and GPT-3 augmentations. During training and inference, BT is used to generate diverse training examples by translating sentences back and forth between languages, thus neutralizing slang and reducing bias. GPT-3 is employed to rephrase sentences, further increasing the diversity and quality of training data. Test-Time Augmentation (TTA) is also utilized, where predictions for original and augmented sentences are combined to enhance detection accuracy. The effectiveness of this methodology was demonstrated through extensive empirical experiments on the Parler and GAB datasets, showing significant improvements in hate speech detection performance.

2 Kar and Debbarma - 2023 - Sentimental analysis & Hate speech detection on English and German text collected from social media platforms using optimal feature extraction and hybrid diagonal gated recurrent neural network [2]

2.1 Summary

The study introduces a novel approach for hate speech detection and sentiment analysis in code-mixed texts across English and German languages, employing an optimal feature extraction method and a hybrid diagonal gated recurrent neural network (FE-DGRNN). This method addresses the

challenges posed by the rapid spread and harmful impact of hate speech in multilingual online environments, demonstrating high accuracy in distinguishing hate speech from non-offensive content.

2.2 Explored Problem

The paper addresses the challenge of detecting and classifying hate speech and harassment on social media platforms, particularly in non-English and code-mixed language environments. The complexity of hate speech, including its thematic focus and target orientation, complicates binary categorization methods and presents a significant challenge in multilingual contexts.

2.3 Approach

The approach outlined in the study combines several innovative techniques: preprocessing, improved seagull optimization (ISO) for optimal feature extraction, and a hybrid diagonal gated recurrent neural network (Hyb-DGRNN) for the classification and analysis of hate speech and sentiment. The method is tested using the HASOC 2019 dataset, focusing on English and German languages, demonstrating its effectiveness with high accuracy, precision, and F-measure scores. The research methodology showcases a three-fold process that significantly enhances hate speech detection and sentiment analysis capabilities in social media texts.

By integrating advanced optimization algorithms for feature extraction and employing a hybrid neural network architecture, this study contributes to the ongoing efforts to combat hate speech online, providing a robust tool for social media platforms and researchers alike.

3 Lee et al. - 2023 - Hate Speech Classifiers are Culturally Insensitive [3]

3.1 Summary

This study critically examines the cultural insensitivity of hate speech classifiers trained on monolingual datasets (Korean, English, Arabic) when applied to translated data from other languages. Through quantitative and qualitative analyses, it finds significant performance drops—up to nearly 50% in F1 scores and up to five-fold increases in false negative rates—demonstrating a substantial cultural gap. The paper underscores the importance of incorporating cultural nuances in hate speech detection models to improve their efficacy across diverse linguistic and cultural landscapes.

3.2 Explored Problem

The primary issue addressed is the cultural insensitivity of hate speech classifiers, which significantly hinders their effectiveness when detecting hate speech across different languages and cultural contexts. The research reveals that classifiers perform poorly on datasets translated from other cultures, highlighting a crucial gap in current hate speech detection methodologies that fail to account for cultural nuances and linguistic variations.

3.3 Approach

The methodology involved translating three monolingual hate speech datasets (Korean, English, Arabic) into each other's languages and then evaluating the performance of hate speech classifiers on these translated datasets. The study utilized both quantitative measures, such as F1 scores and false negative rates, and qualitative sample analyses to understand the reasons behind the classifiers' performance degradation. The findings point to a stark cultural divide, with classifiers unable to accurately interpret and classify hate speech that originates from cultural contexts different from the one they were trained on.

This work not only highlights the challenges in cross-cultural hate speech detection but also calls for the development of more nuanced and culturally aware models. By demonstrating the extent of cultural insensitivity in current classifiers, Lee, Jung, and Oh push for a paradigm shift towards incorporating cultural understanding in the design and training of hate speech detection systems.

4 Biradar et al. - 2022 - Fighting hate speech from bilingual hinglish speaker's perspective, a transformer- and translation-based approach. [4]

4.1 Summary

The study introduces a novel method for detecting hate speech in Hinglish (a mix of Hindi and English), leveraging transformer models like IndicBERT and mBERT, along with transfer learning techniques from ULMFiT and BERT. It proposes a Transformer-based Interpreter and Feature extraction model on Deep Neural Network (TIF-DNN), demonstrating significant improvements in identifying hate speech within Hinglish content with an accuracy of 73%.

4.2 Explored Problem

The research tackles the detection of hate speech in bilingual code-mixed Hinglish data. This is particularly challenging due to the lack of large datasets for such language combinations and the inherent complexity of accurately identifying hate speech across mixed linguistic and cultural contexts.

4.3 Approach

The approach involved experimenting with various transformer models and introducing the TIF-DNN model, which integrates translation and feature extraction techniques to better process and analyze Hinglish text for hate speech detection. The paper details the methodology for data preprocessing, model architecture, and the ensemble strategy for improving hate speech detection accuracy.

5 Ali et al. - 2022 - Hate speech detection on Twitter using transfer learning [5]

5.1 Summary

This paper outlines the development of an Urdu language hate lexicon and the creation of a dataset containing 10,526 Urdu tweets, annotated for hate speech detection. The study employs transfer learning techniques leveraging pre-trained FastText Urdu word embeddings and multilingual BERT embeddings, alongside experimenting with various machine learning models to detect hate speech effectively. Notably, it achieves encouraging F1-scores using BERT, xlm-roberta, and distil-BERT, outperforming several baseline models.

5.2 Explored Problem

The primary challenge addressed is the detection of hate speech in Urdu tweets, a low-resource language in the domain of natural language processing (NLP) for social media content. Given the growth of social media use and the proliferation of hate speech, developing automated tools for its detection, particularly in less commonly studied languages like Urdu, is crucial.

5.3 Approach

The approach includes creating an Urdu hate speech lexicon, compiling a large dataset of Urdu tweets, and applying transfer learning with pre-trained word embeddings (FastText and BERT) to train models for hate speech detection. The paper explores various machine learning models, including BERT variants, to evaluate their performance on the task. The use of transfer learning and multilingual models demonstrates the effectiveness of these methods in dealing with the challenges presented by low-resource languages and the nuanced nature of hate speech.

6 Bigoulaeva et al. - 2022 - Addressing the Challenges of Cross-Lingual Hate Speech Detection [6]

6.1 Summary

The paper explores the use of cross-lingual transfer learning to mitigate the scarcity of annotated hate speech data in low-resource languages. By leveraging cross-lingual word embeddings and neural network systems trained on source language data (English), the study applies these models to target languages (German) that lack labeled examples. The research demonstrates the feasibility and effectiveness of this approach, further enhanced by incorporating unlabeled target language data through bootstrapping techniques. It also examines the impact of class imbalance within hate speech datasets and proposes strategies to address this issue.

6.2 Explored Problem

The main challenge addressed is the limited availability of labeled hate speech datasets in languages other than English, making it difficult to develop hate speech detection systems for a wide range of languages. The study aims to overcome this challenge by employing cross-lingual transfer learning techniques, enabling the adaptation of models trained on well-resourced languages to those with fewer resources.

6.3 Approach

To tackle the problem, the researchers adopted a cross-lingual transfer learning framework, utilizing cross-lingual word embeddings (CLWEs) to train neural network models on English hate speech data and then apply these models to German data without labeled examples. The study employs CNN, LSTM, and BERT-based architectures for the task. Additionally, it leverages unlabeled data in the target language (German) for model improvement through a bootstrapping process, involving ensemble predictions and majority voting to annotate the new data. The approach also tests data undersampling and oversampling techniques to address the issue of label imbalance prevalent in hate speech datasets.

7 Biradar et al. - 2021 - Hate or Non-hate: Translation based hate speech identification in Code-Mixed Hinglish data set [7]

7.1 Summary

This study introduces a Transformer-based Interpretation and Feature Extraction Model (TIF-DNN) designed to identify hate speech in code-mixed Hinglish social media content. The approach involves converting code-mixed data into monolingual data using transliteration and translation, followed by feature extraction with mBERT (multilingual BERT), and classification with deep neural networks (DNN) alongside traditional machine learning models. The model outperforms several baseline classifiers and demonstrates the effectiveness of translating multilingual data into monolingual data for hate speech detection in low-resource languages.

7.2 Explored Problem

The research tackles the problem of hate speech detection in Hinglish, a code-mixed language prevalent on Indian social media platforms. Code-mixing presents a unique challenge for computational linguistics due to its non-standard linguistic structure and the lack of labeled datasets for training models. Traditional deep learning techniques trained on monolingual data struggle with code-mixed content, underscoring the need for innovative approaches to handle multilingual data effectively.

7.3 Approach

The TIF-DNN model employs a three-layer architecture:

- **Interpretation Layer:** This initial layer uses the Microsoft LID-tool for language identification, followed by transliteration (for Hindi words) and translation (for English words) into Devanagari script, creating a monolingual representation of the original code-mixed tweet.
- **Feature Extraction Layer:** mBERT is utilized to extract features from the translated tweets. The model leverages the bidirectional capabilities of BERT within a multilingual context to generate embeddings that capture the semantic nuances of the input.
- **Classification Layer:** Both traditional machine learning models (e.g., Logistic Regression, SVM, Random Forest, etc.) and a custom DNN architecture are explored for classification. The DNN model includes multiple dense layers, dropout for regularization, and batch normalization, culminating in a sigmoid layer for binary classification (hate or non-hate).

This methodological framework demonstrates how converting code-mixed content into a monolingual format can facilitate more effective feature extraction and classification, leveraging the strengths of pre-trained multilingual models and deep learning architectures.

This study highlights the potential for transliteration and translation as preprocessing steps in enhancing hate speech detection in code-mixed languages, providing a foundation for future research in multilingual and low-resource language processing.

8 Wei et al. - 2021 - Offensive Language and Hate Speech Detection with Deep Learning and Transfer Learning [8]

8.1 Summary

This study introduces a methodology for classifying tweets using BI-LSTM models starting from empty embeddings and incorporating pre-trained GloVe embeddings. Further, it explores transfer learning with pre-trained language models like BERT, DistilBert, and GPT-2 to improve hate speech detection. The approach includes extensive data preprocessing, sentiment analysis, and data augmentation techniques to handle the challenges posed by the imbalanced dataset and the nuances of Twitter language. The study achieves over 92% accuracy on the test data, indicating the effectiveness of combining deep learning and transfer learning in detecting offensive content on social media.

8.2 Explored Problem

The research addresses the critical issue of identifying hate speech and offensive language in tweets. With the vast and growing use of social media, moderating content to maintain brand image and prevent the spread of harmful speech has become increasingly challenging. The study aims to automate the detection of such content, helping brands and community managers to efficiently monitor and control digital interactions related to them.

8.3 Approach

The methodology employed in this research involves several key steps:

- **Data Preprocessing and Augmentation:** The tweets dataset undergoes cleaning to remove noise and standardize text. Techniques like sentiment analysis are applied to understand the sentiment polarity of tweets. To address class imbalance, data augmentation methods are used to enrich the dataset with synthetic examples of underrepresented classes.
- **Modeling with BI-LSTM and Transfer Learning:** Initially, BI-LSTM models are developed with and without pre-trained GloVe embeddings to classify tweets. Subsequently, transfer learning is applied using pre-trained models (BERT, DistilBert, and GPT-2), with further fine-tuning to adapt to the hate speech detection task. The models' architectures are designed to capture the sequential nature of text data effectively.
- **Hyperparameter Tuning:** To optimize model performance, hyperparameters such as the number of memory cells in LSTM layers, dropout rates, and learning rates are tuned using Keras Tuner.

This comprehensive approach, combining deep learning models built from scratch with the leveraging of pre-trained models through transfer learning, showcases a robust mechanism for enhancing the automatic detection of hate speech and offensive language on Twitter.

9 Bigoulaeva et al. - 2021 - Cross-Lingual Transfer Learning for Hate Speech Detection [9]

9.1 Summary

This study introduces a novel approach to detecting hate speech in low-resource languages, specifically German, by using cross-lingual transfer learning from English, a high-resource language with abundant hate speech datasets. The method involves training classifiers with bilingual word embeddings on English data and then applying these trained models to German data. Additionally, the research explores bootstrapping on unlabeled German datasets to enhance performance, demonstrating the effectiveness of this method in improving hate speech detection without the need for annotated data in the target language.

9.2 Explored Problem

The main problem tackled by this research is the scarcity of annotated hate speech datasets in languages other than English, which poses a significant challenge in developing automatic hate speech detection systems for such languages. The study aims to mitigate this issue for the German language by employing cross-lingual transfer learning techniques that leverage existing English hate speech datasets, thus avoiding the costly and time-consuming process of collecting and annotating new hate speech data in German.

9.3 Approach

The approach used in this research involves several key steps:

- **Cross-lingual Transfer Learning:** Utilizing bilingual word embeddings and neural network classifiers trained on an English dataset to detect hate speech in German data, without requiring any German annotations.
- **Bootstrapping on Unlabeled Data:** Further enhancing model performance by predicting the labels of unlabeled German datasets using an ensemble of trained models and incorporating this newly labeled data into training. This step is designed to improve the classifiers' accuracy on the target language by exposing them to more German hate speech examples.
- **Model Architectures:** The study experiments with linear SVM, CNN, and CNN/BiLSTM models based on bilingual word embeddings to evaluate the effectiveness of different architectures in the cross-lingual hate speech detection task.

By combining cross-lingual transfer learning with a novel bootstrapping approach on unlabeled data, this research demonstrates a promising avenue for addressing the challenge of hate speech detection in low-resource languages.

10 Mishra et al. - 2021 - Exploring Multi-Task Multi-Lingual Learning of Transformer Models for Hate Speech and Offensive Speech Identification in Social Media [10]

10.1 Summary

The paper presents a multi-task and multi-lingual approach to detect hate speech and offensive content in social media posts across Indo-European languages. Utilizing transformer-based neural networks, specifically designed to handle multiple tasks (e.g., differentiating hate speech, offensive language, and neither) and languages simultaneously, the study expands on contributions to the 2019 shared task on hate speech (HASOC). The authors explored several models, including separate task heads, back-translation for data augmentation, and multi-lingual training, demonstrating that their approach can achieve competitive performance with reduced computational costs. The code and models have been made open source to support further research.

10.2 Explored Problem

The research addresses the problem of efficiently detecting hate speech and offensive content in multi-lingual social media texts. Given the volume of content produced online, manual moderation is impractical. The challenge is compounded when considering the variety of languages and the subtleties of language-specific expressions of hate and offense. The study aims to provide an efficient, scalable, and effective solution that leverages recent advancements in machine learning.

10.3 Approach

The approach involves the development of multi-task learning models capable of handling multiple classification tasks across different languages. The key strategies include:

- **Multi-task Learning with Separate Task Heads:** Each task (e.g., identifying hate speech vs. offensive language) has a dedicated output layer, allowing the model to specialize in each task while sharing lower-level representations.
- **Back-Translation for Data Augmentation:** To mitigate the issue of data scarcity, especially in less-resourced languages, back-translation is used. This involves translating text to a second language and back to the original language to create new training samples.
- **Multi-lingual Training:** The models are trained on datasets from multiple languages, enabling them to learn from a broader range of examples and apply this knowledge across language barriers.

The study extensively tests these models on the HASOC 2019 dataset, which includes social media posts in English, Hindi, and German, assessing their performance and generalizability across languages and tasks. This paper contributes significantly to the field by demonstrating how multi-task and multi-lingual learning approaches can enhance hate speech detection in a diverse linguistic landscape, offering insights into the complexities of language and meaning in the context of social media.

11 Aluru et al. - 2021 - A Deep Dive into Multilingual Hate Speech Classification [11]

11.1 Summary

This study conducts a comprehensive analysis of hate speech detection across multiple languages using datasets in 9 languages from 16 different sources. It evaluates the performance of various deep learning models in monolingual and multilingual settings, observing that LASER embeddings with Logistic Regression excel in low-resource scenarios, while BERT-based models perform better with more extensive datasets. Translation to English and using BERT also achieves competitive results in several languages. The research proposes frameworks that could efficiently address hate speech detection in low-resource languages, serving as robust baselines for future multilingual hate speech detection tasks.

11.2 Explored Problem

The paper addresses the challenge of detecting hate speech across different languages, particularly focusing on the scarcity of annotated datasets for non-English languages. The study explores the effectiveness of deep learning models and embedding techniques in identifying hate speech, aiming to contribute to the development of tools that can be generalized for hate speech detection across various languages.

11.3 Approach

The methodology involves analyzing the performance of different models under monolingual and multilingual scenarios. Key approaches include:

- **Monolingual Setting:** Training and testing models using datasets from the same language, comparing LASER embeddings with Logistic Regression and BERT-based models in both low and high-resource settings.

- **Multilingual Setting:** Training models with data from multiple languages except one and testing on the remaining language, examining the effectiveness of including data from other languages in improving model performance, particularly in zero-shot or low-resource situations.

The study uses a variety of models, including LASER embeddings with Logistic Regression, MUSE embeddings with CNN-GRU, BERT, and mBERT, to explore their capabilities in hate speech detection. The research's findings highlight the potential of using cross-lingual transfer learning and the importance of choosing appropriate models based on the availability of training data across different languages.

12 Aluru et al. - 2020 - Deep Learning Models for Multilingual Hate Speech Detection [12]

12.1 Summary

This study is very similar to the previous one. It investigates the effectiveness of deep learning models in detecting hate speech across multiple languages. It emphasizes the challenges faced due to the scarcity of datasets in languages other than English and explores various models under different scenarios like low-resource settings and multilingual setups. The paper contributes a catalog indicating the most effective model based on the language and data availability, providing a valuable resource for future research in multilingual hate speech detection.

12.2 Explored Problem

The research tackles the significant problem of detecting hate speech in languages with limited datasets. Most available hate speech datasets are in English, leaving a gap in hate speech detection in other languages. This study aims to fill this gap by analyzing the performance of different deep learning models on multilingual hate speech detection, considering both low-resource and high-resource scenarios.

12.3 Approach

The approach taken involves extensive experimentation with different models under monolingual and multilingual settings across nine languages. The models explored include:

- **LASER embeddings with Logistic Regression** for low-resource settings.
- **BERT-based models** (including Translation + BERT and mBERT) for high-resource settings.
- The use of **CNN-GRU** and translation techniques to adapt models for multilingual detection.
- Performance evaluation in scenarios with varying amounts of training data, from very limited (low resource) to full datasets (high resource).

The study's findings suggest that LASER embeddings combined with Logistic Regression work well in low-resource situations, while BERT-based models excel in high-resource scenarios. Additionally, the paper discusses the potential benefits of translating datasets to English to leverage high-quality models developed for the English language.

13 Sohn and Lee - 2019 - MC-BERT4HATE: Hate Speech Detection using Multi-channel BERT for Different Languages and Translations [13]

13.1 Summary

This study introduces a multi-channel BERT (MC-BERT) model that leverages English, Chinese, and multilingual versions of BERT to improve hate speech detection in non-English languages. By incorporating translations of training and test sentences into corresponding languages

suitable for each BERT model, the approach addresses the challenge of detecting hate speech across diverse linguistic contexts. The model was evaluated using three non-English hate speech datasets (Spanish, German, and Italian) and achieved state-of-the-art or comparable performance.

13.2 Explored Problem

The paper addresses the increasing issue of online hate speech in various languages, exacerbated by the anonymity and mobility of social networking services (SNS). Given the costly and time-consuming nature of manual hate speech detection by human annotators, there's a pressing need for effective automatic detection algorithms. The challenge lies in the nuanced nature of hate speech, which may involve sarcasm, cultural references, and requires context understanding, making automatic detection difficult.

13.3 Approach

The MC-BERT model employs a novel multi-channel architecture that integrates three different BERT models (English BERT, Chinese BERT, and multilingual BERT) to capture a broader range of linguistic features for hate speech detection. Additionally, the study explores the effectiveness of using translated texts as additional input, hypothesizing that translations, despite their potential inaccuracies, can provide valuable context when processed through these sophisticated language models. This approach is tested against three datasets in Spanish, German, and Italian, showcasing its adaptability and effectiveness across languages. The methodology includes preprocessing steps tailored for social media texts, application of Google's Translation API for generating translated texts, and detailed model training procedures with specified hyperparameters.

By leveraging the strengths of multiple BERT models and incorporating translations as supplementary data, this research provides insights into improving automatic hate speech detection across different linguistic and cultural contexts, highlighting the potential of transfer learning and multi-channel architectures in addressing complex NLP challenges.

References

[1] S. Cohen, D. Presil, O. Katz, O. Arbili, S. Messica, and L. Rokach, "Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time," *INFORMATION FUSION*, vol. 99, p. 101887, Nov. 2023, num Pages: 9 Place: Amsterdam Publisher: Elsevier Web of Science ID: WOS:001028327900001. [Online]. Available: https://www.webofscience.com/api/gateway?GWVersion=2&SrcAuth=DynamicDOIArticle&SrcApp=WOS&KeyAID=10.1016%2Fj.inffus.2023.101887&DestApp=DOI&SrcAppSID=EUW1ED0E5ArMIeKKjXQ3S5SUq89WU&SrcJTitle=INFORMATION+FUSION&DestDOIRegistrantName=Elsevier

[2] P. Kar and S. Debbarma, "Sentimental analysis & Hate speech detection on English and German text collected from social media platforms using optimal feature extraction and hybrid diagonal gated recurrent neural network," *ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE*, vol. 126, p. 107143, Nov. 2023, num Pages: 12 Patent Number: D Place: Oxford Publisher: Pergamon-Elsevier Science Ltd Web of Science ID: WOS:001086669700001. [Online]. Available: https://www.webofscience.com/api/gateway?GWVersion=2&SrcAuth=DOISource&SrcApp=WOS&KeyAID=10.1016%2Fj.engappai.2023.107143&DestApp=DOI&SrcAppSID=EUW1ED0E5ArMIeKKjXQ3S5SUq89WU&SrcJTitle=ENGINEERING+APPLICATIONS+OF+ARTIFICIAL+INTELLIGENCE&DestDOIRegistrantName=Elsevier

[3] N. Lee, C. Jung, and A. Oh, "Hate Speech Classifiers are Culturally Insensitive," in *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, S. Dev, V. Prabhakaran, D. Adelani, D. Hovy, and L. Benotti, Eds. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 35–46. [Online]. Available: https://aclanthology.org/2023.c3nlp-1.5

[4] S. Biradar, S. Saumya, and A. chauhan, "Fighting hate speech from bilingual hinglish speaker's perspective, a transformer- and translation-based approach." *Social Network Analysis and Mining*, vol. 12, no. 1, p. 87, Jul. 2022. [Online]. Available: https://doi.org/10.1007/s13278-022-00920-w

[5] R. Ali, U. Farooq, U. Arshad, W. Shahzad, and M. O. Beg, "Hate speech detection on Twitter using transfer learning," *Computer Speech & Language*, vol. 74, p. 101365, Jul. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230822000110

[6] I. Bigoulaeva, V. Hangya, I. Gurevych, and A. Fraser, "Addressing the Challenges of Cross-Lingual Hate Speech Detection," Jan. 2022, arXiv:2201.05922 [cs]. [Online]. Available: http://arxiv.org/abs/2201.05922

[7] S. Biradar, S. Saumya, and A. Chauhan, "Hate or Non-hate: Translation based hate speech identification in Code-Mixed Hinglish data set," in *2021 IEEE International Conference on Big Data (Big Data)*, Dec. 2021, pp. 2470–2475. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9671526?casa_token=sCS2MrE1DH8AAAAA:INYTF57vOdc-mTeQFlP2CVaRogxZiSPqzdXytBAtGS-Aj6k1iXEcB0nLjWCKDMzsBJTjXm85xA

[8] B. Wei, J. Li, A. Gupta, H. Umair, A. Vovor, and N. Durzynski, "Offensive Language and Hate Speech Detection with Deep Learning and Transfer Learning," Aug. 2021, arXiv:2108.03305 [cs]. [Online]. Available: http://arxiv.org/abs/2108.03305

[9] I. Bigoulaeva, V. Hangya, and A. Fraser, "Cross-Lingual Transfer Learning for Hate Speech Detection," in *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, B. R. Chakravarthi, J. P. McCrae, M. Zarrouk, K. Bali, and P. Buitelaar, Eds. Kyiv: Association for Computational Linguistics, Apr. 2021, pp. 15–25. [Online]. Available: https://aclanthology.org/2021.ltedi-1.3

[10] S. Mishra, S. Prasad, and S. Mishra, "Exploring Multi-Task Multi-Lingual Learning of Transformer Models for Hate Speech and Offensive Speech Identification in Social Media," *SN Computer Science*, vol. 2, no. 2, p. 72, Feb. 2021. [Online]. Available: https://doi.org/10.1007/s42979-021-00455-5

[11] S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee, "A Deep Dive into Multilingual Hate Speech Classification," in *MACHINE LEARNING AND KNOWLEDGE DISCOVERY IN DATABASES: APPLIED DATA SCIENCE AND DEMO TRACK,*

ECML PKDD 2020, PT V, Y. Dong, G. Ifrim, D. Mladenic, C. Saunders, and S. VanHoecke, Eds., vol. 12461. Cham: Springer International Publishing Ag, 2021, pp. 423–439, iSSN: 0302-9743, 1611-3349 Num Pages: 17 Series Title: Lecture Notes in Artificial Intelligence Web of Science ID: WOS:000716884800026. [Online]. Available: https://www.webofscience.com/api/gateway?GWVersion=2&SrcAuth=DynamicDOIConfProc&SrcApp=WOS&KeyAID=10.1007%2F978-3-030-67670-4_26&DestApp=DOI&SrcAppSID=EUW1ED0E5ArMIeKKjXQ3S5SUq89WU&SrcJTitle=MACHINE+LEARNING+AND+KNOWLEDGE+DISCOVERY+IN+DATABASES%3A+APPLIED+DATA+SCIENCE+AND+DEMO+TRACK%2C+ECML+PKDD+2020%2C+PT+V&DestDOIRegistrantName=Springer-Verlag

[12] ——, "Deep Learning Models for Multilingual Hate Speech Detection," Dec. 2020, arXiv:2004.06465 [cs]. [Online]. Available: http://arxiv.org/abs/2004.06465

[13] H. Sohn and H. Lee, "MC-BERT4HATE: Hate Speech Detection using Multi-channel BERT for Different Languages and Translations," in *2019 International Conference on Data Mining Workshops (ICDMW)*, Nov. 2019, pp. 551–559, iSSN: 2375-9259. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8955559?casa_token=DjXZzpkQB6cAAAAA:XmTLUvChJ3uBfyfCmAJTHgYeZojalS7mZvrveKFS4Uvy70vY7z_MWWVtywc3lEpria-R9clHWw