Hate Speech Detection in Code-Mixed Indonesian Social Media: Exploiting Multilingual Languages Resources

Endang Wahyu Pamungkas
Informatics Department
Universitas Muhammadiyah Surakarta
Surakarta, Indonesia
ewp123@ums.ac.id

Azizah Fatmawati
Informatics Department
Universitas Muhammadiyah Surakarta
Surakarta, Indonesia
af157@ums.ac.id

Yusuf Sulistyo Nugroho
Informatics Department
Universitas Muhammadiyah Surakarta
Surakarta, Indonesia
ysn222@ums.ac.id

Dedi Gunawan
Informatics Department
Universitas Muhammadiyah Surakarta
Surakarta, Indonesia
dg163@ums.ac.id

Endah Sudarmilah
Informatics Department
Universitas Muhammadiyah Surakarta
Surakarta, Indonesia
es132@ums.ac.id

2022 Seventh International Conference on Informatics and Computing (ICIC) | 979-8-3503-4571-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIC56845.2022.10006940

***Abstract*—Hate speech in social media is becoming a relevant issue recently. Several studies have been proposed to deal with the hate speech phenomena in online communication. However, detecting hate speech messages from social media data is not a trivial task. Previous works have mentioned the problem of code-mixed languages in hate speech detection. As a matter of fact, Indonesia consists of several regions, each with its own local languages. Naturally, Indonesians tend to mix their own local language with Bahasa Indonesia when communicating in everyday conversation, including in social media communication, which contributes to the difficulty of processing Indonesian social media data. In this study, we plan to investigate hate speech detection in code-mixed Indonesian social media by exploiting several available multilingual language resources. Our experiment shows that the current available multilingual language model could not improve the model performance compared to the models which utilized the monolingual Indonesian language model. We also found that the most recent neural-based models are able to obtain better performance than the traditional model. For future work, we plan to implement a transfer learning approach to detect hate speech in Indonesian social media, specifically to deal with the code-mixed issue.**

***Index Terms*—abusive language detection, code-mixed, hate speech detection, social media**

I. Introduction

Hate speech in social media is becoming a relevant issue recently, along with the use of social media by people as the primary medium for communicating, gathering information, and even entertainment. Due to the anonymity has given to their users and also the lack of effective regulation, social media platforms have become a convenient medium for an abusive person or even hate groups to promote and propagate their abusive view in online communication, even with higher reachability when their posts become viral [1]. Therefore, combating hate speech online is an essential issue since it was already proven to be detrimental, not only for the victims of the abuse but also for society [2]. Several studies have been proposed to deal with the hate speech phenomena in online communication. Some studies focused on detecting hate speech online by building machine learning models to predict whether a message is hate speech or not [3], [4]. Some other studies propose not only detecting but also countering hate speech [5], [6].

However, detecting hate speech messages from social media data is not a trivial task. Various factors contribute to the difficulties of this task, including the nature of social media data which is full of noise [7], the use of informal language [8], and also multilinguality issue [9]. As a matter of fact, Indonesia consists of several regions, each with its own local languages. A recent report shows 718 local languages used by different regions and tribes in Indonesia [1]. Naturally, Indonesians tend to mix their own local language with Bahasa Indonesia when communicating in everyday conversation, including in social media communication. These phenomena contribute to the difficulty of processing Indonesian social media data, which contain code-mixing use of languages. Previous works have mentioned the problem of code-mixed languages in hate speech detection in several languages, including Hindi-English [10], Tamil-Engish [11], and Swahili-English [12]. However, to the best of my knowledge, there still needs to be a study exploring the code-mixed issue in Indonesian languages. We only found some hate speech studies in several local languages, which will become the basis of this study [13], [14]. In Example 1, we can see an example of a Javanese-Indonesian code-mixed tweet. The Indonesian words are written in black, while the Javanese words are in blue. The

[1]https://petabahasa.kemdikbud.go.id/databahasa.php

use of the word 'picek' is the reason why this tweet is abusive. However, the translation to English, ' small', totally changes the meaning of the tweet, including the abusive context of the overall text. This example shows how code-mixed language use could have an impactful issue, specifically for detecting hate speech from text.

Example 1

[Original tweet] : `kok lu gini sih, kok lu ga gini aja, gina gini matamu picek!`

[English translation] : `why are you like this, why are you not like this, how are your eyes so small!`

In this study, we plan to investigate hate speech detection in code-mixed Indonesian social media by exploiting several available multilingual language resources. Specifically, we will experiment with three variants of machine learning classifiers: the traditional machine learning model, the neural-based model, and the recent transformer-based model. These models will be coupled with the available multilingual language models. We focus on two code-mixed languages, Javanese-Indonesian and Sundanese-Indonesian. Our main objective in this research is to investigate the ability of current available multilingual language representation ability to process code-mixed data, especially in low resources language settings such as Bahasa Indonesian. This paper consists of several sections. Section 2 presents several studies which focus on hate speech detection in Indonesian. Section 3 contains the dataset description and statistics. Section 4 provides the experimental settings, results, and analysis. Finally, Section 5 elaborates on the conclusion and future work of the present study.

II. Related Works

The study of hate speech detection from social media data in Indonesia is still new, and the available resources are still limited. In the early stage of this research area, [15] proposed a new hate speech dataset gathered from Twitter using several keywords related to the political election. Then, the study by [16] proposed another hate speech corpus sourced from Twitter and annotated by finer-grained annotation, including hatefulness, target, and aggressiveness levels. Various approaches have been exploited to deal with the hate speech classification task. The earlier studies used a traditional machine learning model to detect hate speech messages [15], [17]. The more recent works utilized deep learning technology [18], [19], which turned out to obtain better performance compared to the traditional models. The study by [20] also tried to implement an ensemble method to optimize the model performance to obtain the best results.

The code-mixed issue in the hate speech detection task, specifically in the Indonesian language, is still not yet explored. We only found some research focusing on detecting hate speech in Indonesian local languages. The study of [14] proposed the first hate speech dataset, which focused on two Indonesian local languages, including Javanese and Sundanese. Furthermore, they experimented with several machine learning models such as support vector machine, naive Bayes, and random forest classifier, coupled with some lexical features, including character n-grams and word n-grams. Then, further studies by [13], [21] proposed other datasets in Indonesian local languages by providing three extra languages other than Javanese and Sundanese, including Madurese, Minangkabau, and Musi. The dataset from [13] also provides a multi-label annotation, as proposed by [16]. Both studies also experimented with a wide range of traditional machine-learning models. In this study, we use all mentioned hate speech datasets in Indonesian local language as the basis to build our code-mixed dataset.

III. Dataset Description

As mentioned in the previous section, we intend to use datasets from earlier studies [13], [14] as the basis for building our Indonesian code-mixed hate speech datasets . We expect to find code-mixed instances from these available datasets. In this study, we plan to focus on two widely used local languages in Indonesian social media: Javanese and Sundanese. For this purpose, we initiate to conduct a manual check on the available datasets and filter out the code-mixed instances for our experiment. To extend the number of data, we combine the dataset instances from [14] and [13] for both Javanese and Sundanese languages. Then we start the manual check by employing three bachelor students (A, B, and C), two students (A and B) to check the Javanese data and one student (C) to check the Sundanese data. The two annotators of Javanese data (A and B) are native speakers of Javanese and Indonesian, while annotator C is a native speaker of Sundanese and Indonesian. We classify a tweet as code-mixed when a tweet contains words in more than one language, even though only one word is different. For example, a tweet which is dominantly written in Indonesian but also contains one word in Javanese, we classify this tweet as a code-mixed tweet.

Table I present the dataset statistics, which describes the number of instances before and after the manual check process. We can observe that the number of code-mixed data contained in the original data from both datasets ([14] and [13]) are quite low. Based on our manual investigation, most of the non-code-mixed tweets are very short tweets or tweets written only in Indonesian. Tweets written only in local languages, either Javanese or Sundanese, are rare. Furthermore, we also found that the use of the Indonesian word is more dominant in the code-mixed tweet in both Javanese and Sundanese languages. Finally, after the manual check process, we obtained 6,115 tweets for Indonesian-Javanese and 2,945 tweets for Indonesian-Sundanese, which will be used for our experiment.

TABLE I
COMPARISON BETWEEN ORIGINAL AND MANUALLY CHECKED DATASET.

Javanese Data from [13]		
	Original	**Code-mixed**
Hate Speech	2,435	1,234
Not Hate Speech	7,017	2,804
Total	9,452	4,038
Javanese Data from [14]		
	Original	**Code-mixed**
Hate Speech	173	93
Not Hate Speech	3,304	1,984
Total	3,477	2,077
Sundanese Data from [13]		
	Original	**Code-mixed**
Hate Speech	3,168	589
Not Hate Speech	4,960	1,555
Total	8,128	2,144
Sundanese Data from [14]		
	Original	**Code-mixed**
Hate Speech	537	93
Not Hate Speech	1,671	660
Total	2,208	753

IV. EXPERIMENT AND RESULTS

In this section, we present the experiment part of this study, which includes the experimental settings and the analysis of the results.

A. Experimental Settings

In this part, we will explain the setting of our experiment. As mentioned in Section I, our study will focus on investigating the ability of current available multilingual language resources to deal with Indonesian code-mixed data. Therefore, we propose implementing three different models using three available multilingual language models, including LASER, Multilingual FastText (M-FastText), and Multilingual BERT (M-BERT). Specifically, we implement Linear Regression with LASER, LSTM with M-FastText, and M-BERT. Additionally, we also propose three models, including Linear Regression with n-grams feature, LSTM with Indonesian FastText, and Indonesian BERT models. Therefore, we will be able to analyze the ability of multilingual language models by comparing their performance with monolingual Indonesian language models. Below is the description of each model implemented for this experiment.

Linear Regression with LASER: We implement a basic Linear Regression model coupled with LASER Embedding as the feature representation. Language-Agnostic SEntence Representations (LASER) [22] is a multilingual language representation covering 93 languages, belonging to 30 different language families and written in 28 different scripts. We utilize Scikit-learn[2] to implement the models. We use the default parameter provided by Scikit-learn without performing any hyperparameter optimization.

Linear Regression with N-grams Feature: As a comparison, we also implement basic Linear Regression model with a simple n-grams feature. We also exploit Scikit-learn[3] to implement the models.

LSTM with M-FastText: We also adopt straightforward LSTM (Long Short-Term Memory) [23] networks for this experiment. Specifically, our architecture consists of several layers, starting with an embedding layer which consist of 300 dimensions, where we use the multilingual FastText embedding. Multilingual FastText[4] is a multilingual word embedding model obtained by aligning monolingual word embeddings in an unsupervised way. The embedding layer becomes the input to LSTM networks which consist of 128 units, then followed by 16 units of dense layer with ReLU activation function. The final layer is a prediction layer which consists of a dense layer with sigmoid activation function. We tune our architecture by varying the batch size (16, 32, 64) and the number of epochs (1-5) to obtain the best possible result.

LSTM with Indonesian FastText: Basically, we implement the same architecture as LASTM with M-FastText, but we only replace the use of M-FastText with Indonesian FastText model.

Neural Based with Multilingual BERT: Multilingual BERT is a multilingual version of original English BERT [24], which is trained on a Wikipedia dump (excluding user and talk pages) in 104 languages. This model also uses a pre-trained Multilingual BERT model available in TensorFlow-hub (`bert-multi-cased`). Our network starts with the BERT layer, which takes three inputs consisting of id, mask, and segment before passing into a dense layer with RELU activation (256 units) on top and an output layer with sigmoid activation. This model is also optimized with Adam optimizer with a learning rate of 2^{-5}. We vary the number of batch sizes (32, 64, 128) and epochs (1-5) to tune this model.

Neural Based with Indonesian BERT: Basically, this model is implemented similar to previously described model which uses M-BERT. However, we use available Indonesian BERT model available in HuggingFace. The rest of configuration is the same as neural based with multilingual BERT model.

For the experiment, we split both our Javanese-Indonesian and Sundanese-Indonesian into 70% for training and 30% for testing. Minimum preprocessing steps are performed, including changing text to lowercase and mentioning (@) normalization (change mention to "USER"). We evaluate our model performance by using several standard metrics, including macro-precision, macro-recall, macro-F1 and accuracy.

B. Results and Analysis

Table II presents the results of our experiment. In this first part, we will analyze the impact of using a multilingual

[2]https://scikit-learn.org/stable/

[3]https://scikit-learn.org/stable/

[4]https://github.com/facebookresearch/MUSE

TABLE II
RESULTS OF HATE SPEECH DETECTION EXPERIMENT IN CODE-MIXED DATA

Dataset	Javanese Dataset				Sundanese Dataset			
Model	P	R	F_1	Acc	P	R	F_1	Acc
Linear Regression with n-grams	0.652	0.650	0.651	0.736	0.657	0.681	0.665	0.724
Linear Regression with LASER Embeddings	0.698	0.655	0.670	0.797	0.652	0.650	0.651	0.736
LSTM with FastText (ID)	0.811	0.752	0.775	0.860	0.732	0.697	0.710	0.795
LSTM with FastText (Multi)	0.776	0.743	0.757	0.843	0.740	0.689	0.706	0.799
BERT with Indonesian Model	0.794	0.758	0.774	0.854	0.743	0.727	0.735	0.783
BERT with Multilingual Models	0.782	0.724	0.745	0.843	0.735	0.721	0.727	0.799

language model based on our three variant machine learning model results. We observe that the use of LASER with a linear regression model could not obtain better performance than linear regression coupled with straightforward n-gram features in the Sundanese-Indonesian dataset. However, linear regression with LASER is able to achieve better performance than linear regression with n-grams in the Javanese-Indonesian dataset. Meanwhile, multilingual FastText could not perform better than the monolingual Indonesian FastText model. It can be observed from the results of LSTM with multilingual FastText and LSTM with Indonesian FastText, where LSTM with Indonesian FastText obtained better performance in both Javanese-Indonesian and Sundanese-Indonesian datasets. Similar results were also observed in BERT-based models. The Indonesian BERT model also outperformed the Multilingual BERT model on both Javanese-Indonesian.

For the machine learning model performance, we can observe that the BERT-based model achieve the best performance compared to the two other models in Sundanese-Indonesian datasets in term of macro-F-score. Meanwhile, LSTM-based models obtain the best results in Javanese-Indonesian datasets based on the macro-F-score. We also found that the traditional machine learning model obtains the lowest performance than other models, as seen from linear regression results. Therefore, the overall results show that the deep learning-based model is able to outperform the traditional one in all dataset and language settings.

Example 2

[Original tweet] : `USER USER USER congor loe sama pantat loe lebih pintar pantatnya loe ya.....`

[English translation] : `USER USER USER your mouth and your ass, your ass is smarter, huh.....`

Example 3

[Original tweet] : `USER USER USER apa sih gelo ga nyambung tolol'`

[English translation] : `USER USER USER what's wrong you crazy it does not make sense stupid!'`

Based on the experimental results, we can observe that the use of multilingual language models, ranging from the earlier LASER Embeddings to the recent pre-trained embeddings, including FastText and BERT models, were not effective for detecting hate speech in Indonesian code-mixed social media data, particularly on two languages variant namely Javanese-Indonesian and Sundanese-Indonesian languages. Our further analysis found that this result could relate to the nature of our code-mixed data collection. As we mentioned earlier in Section III, Indonesian words are more dominant than the local languages, either Javanese or Sundanese. These data characteristics could be why the use of the Indonesian language model, which contains a more complete vocabulary than the multilingual models, is more effective than the use of multilingual language models. As shown in example 1, example 2, and example 3, where the local languages used (in blue) are very limited compared to the Indonesian words (in black).

V. CONCLUSION AND FUTURE WORKS

In this study, we focus on investigating the ability of current available multilingual language models to deal with hate speech detection tasks in code-mixed Indonesian social media data, specifically on two language variants, namely Javanese-Indonesian and Sundanese-Indonesian languages. We employed various machine learning models coupled with several multilingual language models, including LASER Embeddings, multilingual FastText, and multilingual BERT model. Our experiment uncovers that the current available multilingual language model could not improve the model performance compared to the models which utilized the monolingual Indonesian language model. Based on our manual investigation of the dataset collection, these results could be related to the characteristics of the code-mixed tweet contained in the dataset, where the use of Indonesian words is more dominant than the local language words. From the experiment, we also found that the most recent neural-based models are able to obtain better performance than the traditional model. Based on our results, there is still a lot of room for improvement that could be further explored. Since we already found that using monolingual Indonesian achieve better performance, we can try to do a transfer knowledge approaches, including by using a simple use of multilingual hate lexicon [25] as proposed by [26] or by proposing a joint learning model as presented by [27].

REFERENCES

[1] B. Mathew, N. Kumar, Ravina, P. Goyal, and A. Mukherjee, "Analyzing the hate and counter speech accounts on twitter," *CoRR*, vol. abs/1812.02712, 2018. [Online]. Available: http://arxiv.org/abs/1812.02712

[2] J. Langham and K. Gosha, "The classification of aggressive dialogue in social media platforms," in *Proceedings of the 2018 ACM SIGMIS Conference on Computers and People Research, SIGMIS-CPR 2018, Buffalo-Niagara Falls, NY, USA, June 18-20, 2018*, R. Kishore, D. Beimborn, R. K. Bandi, B. Aubert, D. Compeau, and M. Tarafdar, Eds. ACM, 2018, pp. 60–63. [Online]. Available: https://doi.org/10.1145/3209626.3209720

[3] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, A. Gangemi, S. Leonardi, and A. Panconesi, Eds. ACM, 2015, pp. 29–30. [Online]. Available: https://doi.org/10.1145/2740908.2742760

[4] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on Twitter," in *Proceedings of the NAACL Student Research Workshop*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 88–93. [Online]. Available: https://aclanthology.org/N16-2013

[5] Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, and M. Guerini, "CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2819–2829. [Online]. Available: https://aclanthology.org/P19-1271

[6] M. Obermaier, D. Schmuck, and M. Saleem, "I'll be there for you? effects of islamophobic online hate speech and counter speech on muslim in-group bystanders' intention to intervene," *New Media & Society*, vol. 0, no. 0, p. 14614448211017527, 0. [Online]. Available: https://doi.org/10.1177/14614448211017527

[7] N. Prasad, S. Saha, and P. Bhattacharyya, "A multimodal classification of noisy hate speech using character level embedding and attention," in *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*. IEEE, 2021, pp. 1–8. [Online]. Available: https://doi.org/10.1109/IJCNN52387.2021.9533371

[8] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on bert model," *PloS one*, vol. 15, no. 8, p. e0237861, 2020.

[9] E. W. Pamungkas, V. Basile, and V. Patti, "Towards Multidomain and Multilingual Abusive Language Detection: A Survey," *Personal and Ubiquitous Computing*, 2021, published online: 11 August 2021. [Online]. Available: https://link.springer.com/article/10.1007/s00779-021-01609-1

[10] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and 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*. New Orleans, Louisiana, USA: Association for Computational Linguistics, Jun. 2018, pp. 36–41. [Online]. Available: https://aclanthology.org/W18-1105

[11] B. R. Chakravarthi, A. K. M, J. P. McCrae, B. Premjith, K. P. Soman, and T. Mandl, "Overview of the track on hasoc-offensive language identification-dravidiancodemix," in *Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020*, ser. CEUR Workshop Proceedings, P. Mehta, T. Mandl, P. Majumder, and M. Mitra, Eds., vol. 2826. CEUR-WS.org, 2020, pp. 112–120. [Online]. Available: http://ceur-ws.org/Vol-2826/T2-2.pdf

[12] E. Ombui, L. Muchemi, and P. Wagacha, "Hate speech detection in code-switched text messages," in *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. IEEE, 2019, pp. 1–6.

[13] A. D. Asti, I. Budi, and M. O. Ibrohim, "Multi-label classification for hate speech and abusive language in indonesian-local languages," in *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE, 2021, pp. 1–6.

[14] S. D. A. Putri, M. O. Ibrohim, and I. Budi, "Abusive language and hate speech detection for javanese and sundanese languages in tweets: Dataset and preliminary study," in *2021 11th International Workshop on Computer Science and Engineering, WCSE 2021*. International Workshop on Computer Science and Engineering (WCSE), 2021, pp. 461–465.

[15] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the indonesian language: A dataset and preliminary study," in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE, 2017, pp. 233–238.

[16] M. O. Ibrohim and I. Budi, "Multi-label hate speech and abusive language detection in Indonesian Twitter," in *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 46–57. [Online]. Available: https://aclanthology.org/W19-3506

[17] N. Aulia and I. Budi, "Hate speech detection on indonesian long text documents using machine learning approach," in *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence, ICCAI 2019, Bali, Indonesia, April 19-22, 2019*. ACM, 2019, pp. 164–169. [Online]. Available: https://doi.org/10.1145/3330482.3330491

[18] E. Sazany and I. Budi, "Hate speech identification in text written in indonesian with recurrent neural network," in *2019 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2019*, ser. 2019 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2019. United States: Institute of Electrical and Electronics Engineers Inc., Oct. 2019, pp. 211–216, 11th International Conference on Advanced Computer Science and Information Systems, ICACSIS 2019 ; Conference date: 12-10-2019 Through 13-10-2019.

[19] A. Marpaung, R. Rismala, and H. Nurrahmi, "Hate speech detection in indonesian twitter texts using bidirectional gated recurrent unit," in *2021 13th International Conference on Knowledge and Smart Technology (KST)*, 2021, pp. 186–190.

[20] M. A. Fauzi and A. Yuniarti, "Ensemble method for indonesian twitter hate speech detection," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 11, no. 1, pp. 294–299, 2018.

[21] S. D. A. Putri, M. O. Ibrohim, and I. Budi, "Abusive language and hate speech detection for indonesian-local language in social media text," in *Recent Advances in Information and Communication Technology 2021*, P. Meesad, D. S. Sodsee, W. Jitsakul, and S. Tangwannawit, Eds. Cham: Springer International Publishing, 2021, pp. 88–98.

[22] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *TACL*, vol. 7, pp. 597–610, 2019. [Online]. Available: https://transacl.org/ojs/index.php/tacl/article/view/1742

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[25] E. Bassignana, V. Basile, and V. Patti, "Hurtlex: A multilingual lexicon of words to hurt," in *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*, ser. CEUR Workshop Proceedings, E. Cabrio, A. Mazzei, and F. Tamburini, Eds., vol. 2253. CEUR-WS.org, 2018. [Online]. Available: http://ceur-ws.org/Vol-2253/paper49.pdf

[26] P. Chiril, E. W. Pamungkas, F. Benamara, V. Moriceau, and V. Patti, "Emotionally informed hate speech detection: A multi-target perspective," *Cogn. Comput.*, vol. 14, no. 1, pp. 322–352, 2022. [Online]. Available: https://doi.org/10.1007/s12559-021-09862-5

[27] E. W. Pamungkas, V. Basile, and V. Patti, "A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection," *Inf. Process. Manag.*, vol. 58, no. 4, p. 102544, 2021. [Online]. Available: https://doi.org/10.1016/j.ipm.2021.102544