International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 3, Number 3, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v3n3.14
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Cross-Language Offensive Speech Detection Using the mBERT Model

Yan Liu, Jiating Chen *

University of the East, Manila, Philippines
*Corresponding Author: 1720336874@qq.com

ABSTRACT

Aggressive speech can be detrimental to social stability. However, due to a lack of annotated data for aggressive speech, current automatic detection of aggressive speech focuses primarily on a few high-resource languages, making detection for low-resource languages difficult. We propose a cross-language, unsupervised, aggressive transfer detection method to address this. Firstly, we utilize a multilingual BERT (mBERT) model to learn aggressive features on a high-resource English dataset, resulting in an original model. Following that, by analyzing the linguistic similarity between English and low-resource languages like Danish, Arabic, Turkish, and Greek, we transfer the original model to these four low-resource languages, enabling automatic detection of aggressive speech. Experiment results show that, when compared to four other methods—BERT, linear regression (LR), multilayer perceptron (MLP), and support vector machine (SVM), our proposed method improves accuracy and F1 score by nearly 2 percentage points in Danish, Arabic, Turkish, and Greek. This method, which combines cross-language model transfer learning and transfer detection, shows promise in achieving unsupervised aggressive detection in low-resource languages, with performance comparable to current supervised detection methods.

KEYWORDS

Cross-lingual model; Offensive speech detection; BERT; Unsupervised; Transfer learning

1. INTRODUCTION

On social media, there is frequently a significant amount of aggressive speech, such as online bullying, cyber attacks, and hate speech. Aggressive speech on social media can disrupt normal communication and, in some cases, elicit emotional responses from the public, resulting in negative effects on social stability. As a result, detecting and filtering aggressive speech on the internet has grown in importance, and it has emerged as a hot research topic in natural language processing.

Due to the abundance of data resources, monolingual dictionaries, and pretrained language models, research on aggressive speech detection currently focuses primarily on high-resource languages such as English (Rosenthal et al., 2012). However, on social media platforms, a variety of languages, including those from different countries, ethnic groups, and regions, are frequently used for aggressive speech. Because of the scarcity of datasets, research on aggressive speech detection in low-resource languages faces significant challenges.

Offensive language detection is a subset of the classification task that is frequently divided into two stages: upstream language modeling and downstream classification feature learning. The Neural Network Language Model (NNLM) investigates and models natural language's intrinsic dependencies by constructing neural networks capable of representing a word or sentence with vectors. A robust representation improves the generalization ability of the downstream model. The

ability of detection methods to generalize is frequently built on a foundation of extensive data resources. As a result, when modeling low-resource languages, the limited resources available impede learning of intrinsic dependencies within the language object. As a result, learning effective semantic encoding for textual semantic features in low-resource languages becomes difficult.

Research indicates that semantic encoding of textual content in low-resource languages can be achieved through cross-language word vectors combined with transfer learning (TL). Furthermore, the downstream stage model's generalization ability determines the quality of classification performance. The amount of available data resources, however, also determines the downstream stage's generalization ability, resulting in the downstream classification model's inability to rely on these low-resource data for an effective offensive language detection model.

Low-resource offensive language detection faces two major challenges: first, due to limited resources, it is unable to effectively encode the semantic meaning of text in low-resource languages on its own; second, it cannot effectively train on the offensive features present in low-resource languages. Based on the results, this work applies a transfer learning architecture based on the BERT (Bidirectional Encoder Representation from Transformers) model, utilizing the multilingual pre-trained language model, multilingual BERT (mBERT), for transfer learning in low-resource languages. The model can now encode the semantic meaning of text in low-resource languages using this way.

Furthermore, by studying linguistic similarities between different languages, a further transfer is carried out to enable cross-language identification of objectionable language in low-resource languages. This improves the model's ability to recognize offensive words in languages with limited resources.

2. REVIEW OF RELATED LITERATURE

Many researchers, both domestic and international, have conducted extensive research on the topic of offensive speech.

Early efforts in offensive speech recognition depended mainly on manually extracting various types of features, knowledge-based functions, and multimodal information (Pamungkas E W, Patti V, 2019). Saroj et al. (2020) are an example. To identify abusive speech in Hindi on social media, four machine learning classifiers were used: stochastic gradient descent (SGD), linear support vector machine (LSVM), multinomial naive bayes (MNB), and linear regression (LR).

Pathak et al. (2021). extracted n-gram features from text language and employed machine learning classification and regression methods to learn the features of such aggressive speech. However, feature-based methods tend to have relatively weak capabilities in text representation, often necessitating the construction of high-dimensional features for complex text learning. This process consumes significant resources during relevant computations, and feature redundancy can impact the practical effectiveness of classification.

Howard et al. (2020), inspired by Zampieri et al. (2018), used BERT to successfully apply transfer learning to aggressive speech recognition using the Universal Language Model Fine-Tuning for Text Classification (ULMFiT) technique. These methods have become typical for dealing with tough voice recognition jobs because to their better performance. For example, seven of the top ten teams in Task A in the 2019 OffensEval competition (Liu P., Li W., Zou L., 2019) used BERT, with variances mostly in parameter values and preprocessing processes. Currently, the use of cross-language pretraining models for aggressive speech detection is primarily based on cross-language model foundations that have been pretrained. The primary advantage of these methods is the use of unsupervised cross-language pretraining models, which allows for the detection of aggressive speech in languages with limited resources.

Ayo et al. (2021). proposed a method for building cross-language models for detecting aggressive and misogynistic speech using a Support Vector Machine (SVM) and BERT.

By transferring knowledge from resource-rich aggressive speech detection tasks to low-resource languages, Kapil et al. (2021) effectively improved the accuracy of aggressive speech detection in low-resource languages.

However, the aforementioned approaches' detection performance is far from ideal.

3. THEORETICAL FRAMEWORK

The research on the relevant theories above forms the basis for the subsequent investigation. Prior research on low-resource aggressive speech detection encountered two major challenges: first, it is difficult to effectively encode the semantic content of text in low-resource languages due to limited resources; and second, effective training of aggressive features in low-resource languages is difficult.

This study's method includes two components: first, monolingual aggressive speech detection learning, and second, cross-language transfer detection. Initially, the mBERT model is used for transfer learning and encoding in a high-resource monolingual (English) aggressive speech sample set, resulting in the training of a monolingual aggressive speech detector for a given set of monolingual aggressive speech samples. The monolingual aggressive speech detector is then transferred to low-resource languages to detect aggressive speech in given low-resource language texts, resulting in automatic detection of aggressive speech in low-resource language texts. This is described in the theoretical framework depicted in Figure 1 below.

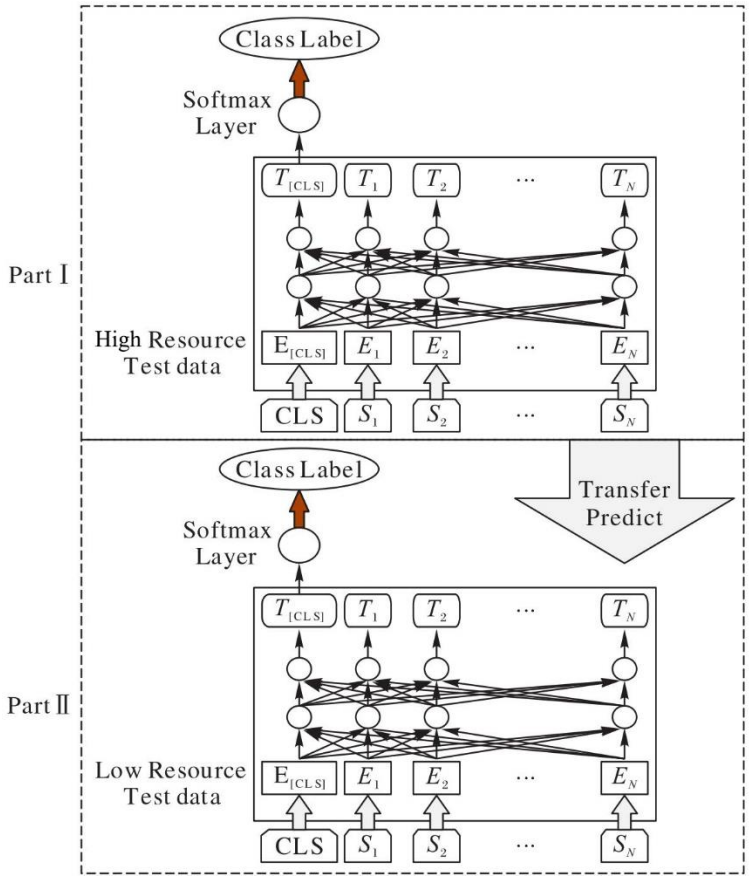

Figure 1. Research Framework of the Study

4. METHODOLOGY/RESEARCH BDESIGN

4.1. Research Instrument

The descriptive study employs the research instrument known as mBERT (Multilingual BERT), which is a multilingual pretrained model used for various natural language text processing tasks.

Google's mBERT is a pre-trained multilingual model based on the BERT architecture. It has 12 self-attention mechanisms modules and is made up of 12 stacked Transformer layers with a hidden layer size of 768.

Pre-training the mBERT model involves connecting monolingual Wikipedia data from 104 different languages, including English, Hindi, Turkish, Malayalam, and others. It makes use of a shared vocabulary of 120,000 words. By utilizing this shared vocabulary, the model ensures that character encodings for all languages exist within a unified embedding space and encoder.

This design implies that text from different languages is mapped to the same encoding space, allowing the model to learn universal language representations in this unified space. Such a configuration allows the mBERT model to seamlessly apply to different language tasks without requiring specific training for each language.

4.2. Data Collection and Sample Size

To validate the performance of the proposed research method, publicly available aggressive speech detection datasets were utilized, as outlined in Table 1, all sourced from the official SemEval website.

For the high-resource languages in the aggressive speech detection dataset, this study employed the English labeled dataset released as part of the 2019 OffensEval shared task 6 (EN-OLID). OLID (EN-OLID) stands out as one of the most popular English language datasets. For other low-resource languages, datasets in Danish, Arabic, Turkish, and Greek were chosen from the 2020 OffensEval shared task 12. The OLID dataset comprises three subtasks: Subtask A involves detecting whether language text is aggressive or not, along with the sum of both samples. Subtask B classifies the type of aggression in aggressive language text as targeted insult (TIN), targeted threat (TTH), or non-targeted (UNT). Subtask C involves determining the target of aggression as an individual (IND), group (GRP), organization or entity (ORG), and other (OTH). However, Arabic, Danish, Greek, and Turkish datasets only include Subtask A. This study specifically focuses on exploring all the Subtask A data in the experiments. Additionally, the experimental setup maintains a training-to-testing sample ratio of 9:1.

Table 1. Sample data distribution

Language	Training set			Test set			Training and Test set union
	Adversarial	Non-adversarial	Training set union	Adversarial	Non-adversarial	Test set union	
English	4 000	7916	11916	400	924	1324	13240
Danish	344	2320	2664	40	256	296	2960
Arabic	1395	5660	7055	155	629	784	7839
Turkish	5441	22708	28149	605	2523	3128	31277
Greek	2238	5631	7869	248	626	874	8743

4.3. Parameter Settings

Word Vector Dimension: The word vector dimension for the model in this method is set to 768 dimensions.

Vocabulary: The vocabulary corresponding to the mBERT pretrained model is set as the shared vocabulary for the text experiments in this study. This vocabulary includes 104 languages with a total of 120,000 words.

Ross-Language Transfer Learning: After analyzing the sample data, it was observed that the majority of data lengths are within 120 words. Therefore, the maximum sentence length is set to 120. The Softmax layer's hidden layer units are set to the number of label categories, which is two.

Model Training: The number of epochs is set to 10, and the training batch size is set to 64.

Optimizer Options: Adam is the optimizer.

Optimization Parameter Settings: The dropout parameter for the hidden layers is set at 0.01, while the learning rate is fixed at 0.00002.

4.4. Data Detection Method

For aggressive speech detection, the researchers used a method that combines mBERT with transfer learning (TL). This method entails learning detection on a high-resource language and then applying that knowledge to a low-resource language via transfer learning, resulting in cross-language aggressive speech detection. The detection effectiveness of low-resource languages is improved by sharing knowledge across languages.

4.5. Monolingual Aggressive Speech Detection Learning

Given the small size of the provided monolingual aggressive speech sample set, building a reasonably complete model to express the semantic information in these speech texts is insufficient. Cross-language transfer learning methods, on the other hand, can benefit from larger datasets made available by other languages. Transfer learning improves model learning efficiency by transferring previously learned model parameters (or knowledge) to a new model, avoiding the drawbacks of many networks that start from scratch.

Kudugunta et al. (2019) and Kondratyuk et al. (2019) demonstrated that important features from cross-language pretrained models could be retrieved for downstream tasks such as named entity recognition and part-of-speech tagging, obtaining language-knowledge-based information for specific tasks.

Kumar et al. (2020). achieved significant results in German and Hindi tasks involving detecting hate speech and offensive content by fine-tuning the pre-trained member model.

Libovicky et al. (2021). demonstrated that context-based mBERT can capture language similarities and cluster languages based on linguistic characteristics. Cross-language fine-tuning has no effect on this property, which means that mBERT can encode a portion of the language information based on its position in the embedding space, concentrating the encoding of each language and achieving a certain level of cross-language capability.

This study, inspired by previous work, makes use of the mBERT model's cross-language nature, allowing the detector to capture text features in multiple languages. Through transfer learning, mBERT's knowledge and information about various languages is shared with the new model. The detector model is then trained based on the mBERT model parameters, enhancing the model's learning efficiency.

The BERT model is used as the foundational structure for building the aggressive speech detection model in the proposed research method. The mBERT model parameters serve as the initial parameters for the previously mentioned aggressive speech detection model. On this foundation, the training of the aggressive speech detector is completed by fine-tuning the BERT model.

For a given monolingual aggressive speech sample, the Tokenizer method is used to process the sample into a token sequence that includes a special token [CLS]. The special token [CLS] is added at the beginning of the token sequence. The token sequence X is represented as shown in Equation (1):

$$X = [\left[\mathrm{CLS}\right], S_1, S_2, \cdots, S_N] \tag{1}$$

After that, the token sequence X is processed and fed into the BERT model. It then passes through 12 stacked Transformers to produce an output matrix. The text sequence and the special token [CLS] are represented by multiple representation vectors in this output matrix. The representation vector for the special token [CLS] in the aggressive speech detection task frequently captures elements of the aggressive speech text. After feeding this vector into the Softmax layer for the classification task, a probability vector with a length equal to the total number of labels is produced as the output.

For the input text, the predicted label is the category that corresponds to the maximum probability value in the probability vector, as shown in Equation (2):

$$m = \delta\left(W \cdot BERT_{[\mathrm{CLS}]} + b\right) \tag{2}$$

Where N is the size of the model output corresponding to the number of possible class labels for the text, B is the dimensionality of the BERT model output, and BERT[CLS] is the vector corresponding to the special token '[CLS].' Where m is the output probability vector, δ is the Softmax function, i.e., the Softmax layer, and W ∈ RN*B and b ∈ RN are the parameters of the Softmax layer.

Ultimately, the loss function used for this assignment is the cross-entropy function. Equation (3) depicts the loss function as follows:

$$Loss = -\sum_{i} \gamma_i * \mathrm{lb}\left(\phi_i\right) \tag{3}$$

Where yi represents the true label for the i-th text.

By utilizing transfer learning, the weight parameters of the mBERT cross-language pretrained model are transferred to the monolingual aggressive speech detection model as initial parameters. On this basis, the model learns aggressive features in aggressive speech and ultimately obtains a monolingual aggressive speech detection model.

4.6. Cross-Language Aggressive Speech Detection

The aforementioned monolingual aggressive speech detection model not only explains aggressive features in high-resource aggressive speech but also inherits the semantic encoding capabilities across multiple languages from the mBERT pretrained model. The monolingual aggressive speech detection model can now detect languages that do not engage in aggressive feature learning. Cross-language detection, as used in this study, refers to the method of using the trained monolingual aggressive speech detection model to detect languages that did not participate in aggressive feature learning.

It is important to note that the method in this study did not use the aforementioned monolingual aggressive speech detection model for further transfer learning on low-resource data. As a result, this method is known as unsupervised cross-language detection.

For certain languages, such as Danish, Arabic, and Hindi, the resources for aggressive speech samples are extremely scarce. When using existing resources for cross-language transfer learning, the sample size is insufficient to train a complete aggressive speech detector for that language. Table 1 shows

that these languages have some similarities, to varying degrees. In real-life scenarios, for example, Danish and English share a significant portion; English and Turkish have similar character compositions, whereas Hindi, Arabic, and Greek have significant differences in character compositions, indicating lower similarity among these languages. As a result, a method is proposed for investigating linguistic similarity between different languages, selecting an appropriate monolingual aggressive speech detector, and employing it to detect aggressive speech in texts written in low-resource languages.

Table 2. Sample data for various languages

Language	Sample Data
English	And this from the clown that should be in prison?
Danish	Og dette fra klovnen, der burde være i fængsel?
Arabic	وهذا من المهرج الذي يجب أن يكون في السجن؟
Turkish	Ve bu da hapiste olmas1 gereken palyac;odan m1?
Greek	Και αυτό από τον κλόουν που θα έπρεπε να είναι στη φυλακή;

This study used the Gromov-Hausdorff (GH) distance proposed by Patra et al. (2019). to quantify the semantic similarity between two languages. Unlike Patra et al., who assigned words from different languages to different embedding spaces, the embeddings encoded by the cross-language pretrained model mBERT are assigned to the same space. In addition, visualization of the embeddings (as shown in Figure 2) reveals that different language encodings cluster in different regions. As a result, the equidistant distance between the embeddings of two languages in different regions is all that is needed for this study to quantify the semantic similarity between the two languages.

Specifically, the GH distance is defined as shown in equation (4):

$$\mathcal{H}(\mathcal{X},\mathcal{Y}) = \max\left\{\sup_{x\in\mathcal{X}}\inf_{y\in\mathcal{Y}} d(x,y), \inf_{y\in\mathcal{Y}}\sup_{x\in\mathcal{X}} d(x,y)\right\} \tag{4}$$

Where X and J are two given embedding spaces, d () is the distance function, and x and y are the embeddings of two languages under comparison. Consequently, the language similarity between different languages can be obtained. Based on this, a language selector can be constructed according to the language similarities between different languages. Finally, for a given text in a low-resource language, the language selector chooses the monolingual aggressive detection model with the highest language similarity for transfer detection.

5. PRESENTATION AND DISCUSSION OF RESULTS

5.1. Metrics

The metrics mainly include the model's predicted accuracy, precision, recall, and macro F1 value (referred to as F1 value hereafter).

5.2. Comparison Experiment Analysis of Different Language Model Detection Methods

As shown in Table 3, when compared to the four methods set, our proposed method improves in both accuracy and F1 value, confirming the superiority of our proposed cross-language transfer detection method over single-language model-based detection methods. The reasons can be divided into two categories:

(1) Theoretically, automatic detection of aggressive language can be achieved by fine-tuning the BERT model. However, this goal necessitates a large corpus, and due to a lack of data, the model struggles to effectively represent and learn the aggressive features in the text. Similarly, when using TF-IDF to represent text features, it fails to effectively represent the multidimensional information in the text in cases of limited datasets. As a result, while these single-language models can learn some aggressive features, their detection performance during testing is generally mediocre.

(2) In scenarios with extremely limited data resources, the composition of words in different languages can vary significantly. This results in substantial differences in the information represented by the detector when using text from other languages as input. The detector has a substantial information gap between the information it can represent for such text and the information it represents for text in the trained language, making it challenging for transfer detection. The transfer learning from mBERT equips the detector with a certain representation capability for different languages, helping narrow the information gap in the text representation stage. This facilitates the transfer of features learned by the monolingual detector to perform detection in other languages.

Table 3. Comparison of experimental results of different methods

Language detection	Model	Accuracy	F1
Danish	This research method	0.796	0.619
	BERT	0.602	0.412
	LR	0.563	0.407
	SVM	0.615	0.444
	MLP	0.592	0.441
Arabic	This research method	0.764	0.508
	BERT	0.723	0.443
	LR	0.651	0.22
	SVM	0.735	0.499
	MLP	0.672	0.478
Turkish	This research method	0.73	0.553
	BERT	0.569	0.397
	LR	0.521	0.336
	SVM	0.602	0.449
	MLP	0.551	0.401
Greek	This research method	0.69	0.525
	BERT	0.58	0.418
	LR	0.535	0.378
	SVM	0.596	0.445
	MLP	0.554	0.368

In practice, the collection and annotation of data for languages with extremely limited available resources are laborious and resource-intensive tasks. As a result, these languages end up with limited resources. However, using the cross-language transfer detection method allows for the migration of aggressive features learned from other languages to detect aggressive language in low-resource languages, thereby broadening the method's applicability.

Experiments also show that the cross-language transfer detection method can detect aggressive language in languages with limited resources.

Table 3 illustrates how different languages have different transfer detection outcomes from the aggressive language detector using the English dataset. When compared to other languages, Danish has the greatest results for transfer detection.

The original data in Table 3 show that the transfer detection effects of different monolingual detection models vary across languages.

The results show that the difficulty of recognizing hostile language in low-resource circumstances can be addressed by transfer detection across two semantically comparable languages. To further evaluate the efficacy of our study methodology, we compute the GH distance between two languages using Equation (4) as a measure to assess the performance of the optimal transfer detection model, as indicated in Table 4. In order to assess the degree of linguistic similarity between various language pairings, Table 4 computes the GH distance of word vectors between the three languages (English, Turkish, and Greek) with the largest sample sizes and other languages. A higher degree of similarity between two corresponding languages is indicated by a smaller value.

From Table 4, it can be observed that English has a higher similarity with Danish, Greek has a higher similarity with Danish, and Turkish has a higher similarity with Arabic, aligning with common linguistic observations.

Table 4. GH distance

Language	Danish	Arabic	Turkish	Greek
English	0.31	0.38	0.39	0.47
Greek	0.25	0.36	0.29	
Turkish	0.22	0.2		0.39

5.3. Analysis of the Impact of Language Similarity on Transfer Effects

To investigate the effect of semantic similarity on transfer effects, monolingual detection models were trained in the three languages with the largest sample sizes and then transferred to other languages for transfer detection experiments, evaluating the transfer detection performance of different monolingual detection models on different languages.

First, offensive language detection models were trained separately on datasets in three languages: English, Turkish, and Greek. This resulted in the development of detection models for English, Turkish, and Greek. Following that, these three detection models were individually tested on datasets in other languages, as shown in Figure 2.

And the letters en, da, ar, tr, and el stand for English, Danish, Arabic, Turkish, and Greek, respectively.

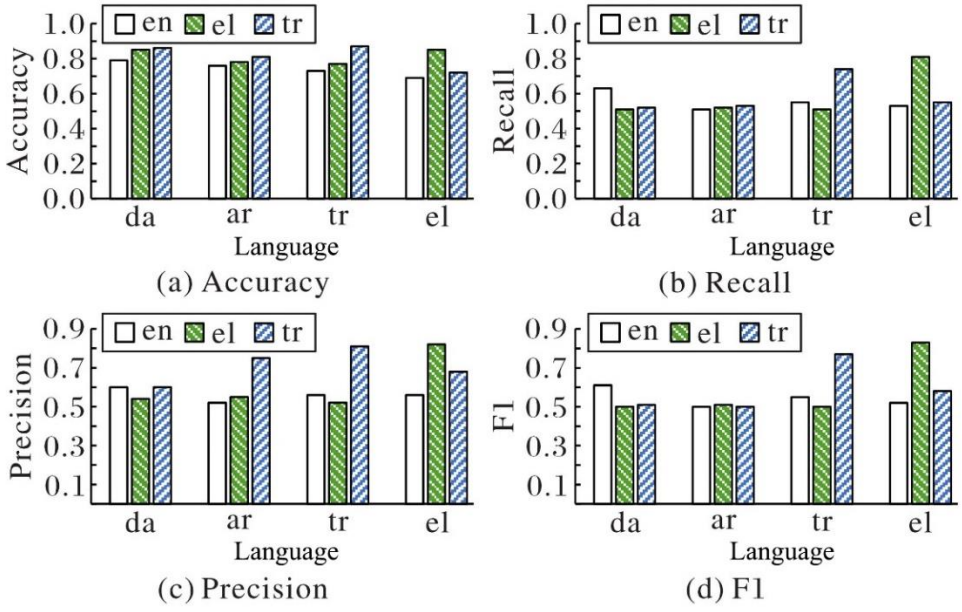

Figure 2. Comparison of accuracy, recall, precision and F1 performance of five language models

From Figure 2, it can be observed that, for F1 and Recall in Danish, Arabic, and Turkish, the evaluation metrics corresponding to the English detector are generally superior to the other two classifiers (in this analysis, the results of using the Turkish detector for Turkish and the Greek detector for Greek are not included). The Gromov-Hausdorff (GH) distance between these three languages and English is also significantly smaller than the GH distance between the other two.

For Accuracy and Precision, the bar chart for Turkish is higher than the one for English, indicating that this method takes advantage of language similarity to some extent. This reduces information loss when transferring between semantically similar languages, emphasizing the significance of semantic similarity in cross-language detection.

Using this feature, one can select high-resource data with the highest semantic similarity to the low-resource language for training a monolingual offensive language detection model. This method is more effective in detecting offensive content in low-resource languages.

5.4. Analysis of the Impact of Training Resource Quantity on Transfer Effects

Different training sample sizes were set to analyze the changes in transfer detection performance during the variation of sample size, as shown in Figure 3.

It is evident that transfer detection performance improves with increasing training sample size. Figure 3 shows that the model is now in a less-than-ideal state when the training sample size is less than 3000. The performance metrics for transfer identification in different languages are all less than 0.35. As the training sample size increases, the performance metrics for transfer detection also rise, and when the training sample size reaches 12,000, the various metrics in the chart are in a slow-growing or even stable state. Furthermore, when compared to other languages, Danish exhibits the fastest growth in detection performance indicators, owing to its highest semantic similarity with English. As a result, the greater the number of training samples, the better the model performs when transferred to other languages. The impact of the amount of training resources, however, decreases when sample size rises above a particular threshold. Moreover, the change in detection performance for the low-resource language with the highest semantic similarity to the high-resource language is most noticeable.

Therefore, according to this study, language similarity is the primary factor leading to superior transfer effectiveness, with the more similar two languages being, the better the transfer detection effect.

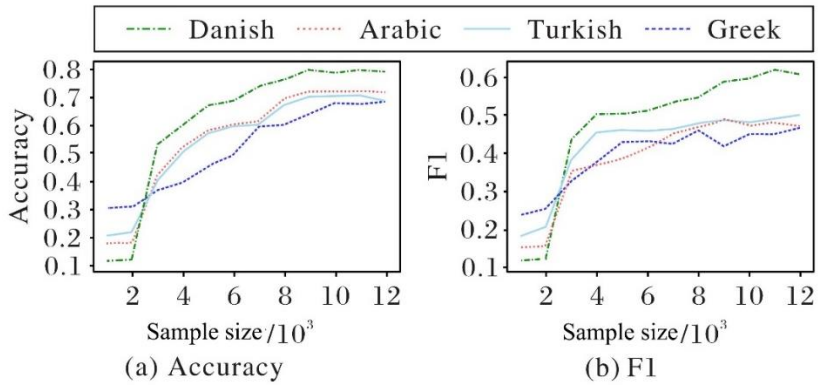

Figure 3. Comparison of model transfer detection with different training sample sizes

5.5. Comparison Experiment Analysis between Unsupervised and Supervised Methods

This study relies heavily on an unsupervised detection method based on mBERT. Unsupervised methods were compared to a set of supervised methods to further investigate the impact of language

similarity on low-resource language tasks. The specific implementation entails creating a detection model for a high-resource language, such as English, and then performing additional transfer learning on a small set of low-resource languages.

Table 5 shows the experimental results. Table 5 shows that the unsupervised method in this study performs similarly to supervised methods, with some variation in proximity between languages. In Accuracy and F1, the supervised method outperforms the unsupervised method by 0.029 and 0.090, respectively.

Additionally, it is observed that Danish, which has a higher similarity to English, outperforms other languages in terms of various metrics and proximity. This adds to the evidence that language similarity has a consistent impact on tasks involving low-resource languages.

Table 5. Compari son of supervised method and proposed unsupervised method

Method	Danish		Arabic		Turkish		Greek	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
Supervised	0.825	0.709	0.801	0.692	0.768	0.672	0.791	0.683
This research method (Unsupervised)	0.796	0.619	0.764	0.508	0.73	0.553	0.69	0.525

6. CONCLUSIONS

(1) This study was successful in developing an innovative offensive language detection system that addresses the problem of detecting offensive language in low-resource languages. The comprehensive approach, which combines cross-language model transfer learning with transfer detection, provides a feasible and efficient solution for offensive detection in low-resource languages.

(2) This study ensured the model's semantic encoding capability for multiple languages by transferring the mBERT pre-trained language model within the BERT framework. Furthermore, this transfer learning strategy reduced resource consumption during multilingual model training, improving overall resource efficiency.

(3) This study successfully detected offensive language in specific languages by fine-tuning the BERT model. The process of fine-tuning allowed the model to better adapt to the linguistic context and characteristics of the target language, improving the precision and accuracy of offensive language detection.

(4) Investigating linguistic similarities between languages was the aim of this study, which aimed to increase the efficacy of offensive language transfer detection in low-resource languages. The innovation is in leveraging language similarities to ensure greater consistency in semantic representation.

(5) The experimental results show that the method used in this study improves transfer detection significantly in low-resource languages. The results provide strong support for the proposed approach's effectiveness in addressing offensive language detection issues in low-resource language scenarios.

7. RECOMMENDATIONS

(1) Expand offensive speech corpora collection and annotation efforts in low-resource languages, add more samples to improve model adaptability, and better accommodate the linguistic contexts of different languages.

(2) Expand the study method's potential applications in other domains of natural language processing by investigating its use in a variety of tasks, including text production and machine translation.

(3) Given the study's relatively small dataset size, future work could include experimenting with and analyzing larger, publicly available datasets to validate the model's performance under different conditions.

(4) In the future, collaboration with other research teams and institutions to share datasets and models will be encouraged, fostering the development of relevant research fields and advancing the study of unsupervised cross-language models.

REFERENCES

[1] ROSENTHAL S, ATANASOVA P, KARADZHOV G, et al. (2021). SOLID: a large-scale semi-supervised dataset for offensive language identification. Findings of the Association for Computational. Linguistics: ACL-IJCNLP 2021. Stroudsburg, PA: Association for Computational Linguistics, 915-928.

[2] PAMUNGKAS E W, PATTI V. (2019). Cross-domain and cross-lingual abusive language detection: a hybrid approach with deep learning and a multilingual lexicon. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research W orkshop. Stroudsburg, PA: Association for Computational Linguistics, 363-370.

[3] SAROJ A, PAL S. (2020). An Indian language social media collection for hate and offensive speech. Proceedings of the lst W orkshop on Resources and Techniques for User and Author Profiling in Abusive Language. Paris: European Language Resources Association, 2-8.

[4] PATHAK V, JOSHI M, JOSHI P A, et al. (2021). KBCNMUJAL@ HASOC-Dravidian-CodeMix-FIRE2020: using machine learning for detection of hate speech and offensive code-mixed social media text. https /arxiv. org/ftp/ arxiv/papers/2102/2102. 09866. pdf.

[5] ZAMPIERI M, NAKOV P, ROSENTHALS, et al. (2018). SemEval-2018 Task 12: multilingual offensive language identification in social media. Proceedings of the 14th W orkshop on Semantic Evaluation. International Committee fo Computational Linguistics, 1425-1447.

[6] HOW ARD J, RUDER s. (2020). Universal language model fine tuning for text classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: Association for Computational Linguistics, 328-339.

[7] LIU P, LI W, Z0U L. (2019). NULI at SemEval-2019 Task 6: transfer learning for offensive language detection using bidirectional transformers. Proceedings of the 13th International Workshop on Semantic Evaluation. Stroudsburg, PA: Association for Computational Linguistics, 87-91.

[8] AYOF E, FOLORUNSO 0, IBHARALU F T, et al. (2021). Hate speech detection in Twitter using hybrid embeddings and improved cuckoo search-based neural networks. International Journal of Intelligent Computing and Cybernetics, 13(4) :485-525.

[9] KAPIL P, EKBAL A. (2021). A deep neural network based multi-task learning approach to hate speech detection [J]. Knowledge-Based Systems, 210: No. 106458.

[10] KUDUGUNTA S, BAPNA A, CASWELL I, et al. (2019). Investigating multilingual NMT representations at scale. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 1565-1575.

[11] KONDRATYUK D, STRAKA M. (2019)75 languages, 1 model: parsing universal dependencies universally. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2779-2795.

[12] KUMAR A, SAUMYA S, SINGH J P. (2020). NITP-AI-NLP@HASOC- FIRE2020: fine tuned BERT for the hate speech and offensive content identification from social media [C]. Proceedings of the 12th Meeting of Forum for Information Retrieval Evaluation. Aachen: CEUR-WS. org, 266-273.

[13] LIBOVICKY J, ROSA R, FRASER A. (2021). How language-neutral is multilingual BERT? https:/arxiv. org/pdf/1911. 03310. pdf.

[14] PATRA B, MONIZ J R A, GARG S, et al. (2019) Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA: Association for Computational Linguistics, 184-193.