Social Network Analysis and Mining (2024) 14:136
https://doi.org/10.1007/s13278-024-01291-0

ORIGINAL ARTICLE

Improving automatic cyberbullying detection in social network environments by fine-tuning a pre-trained sentence transformer language model

Karel Gutiérrez-Batista[1] · Jesica Gómez-Sánchez[2,3] · Carlos Fernandez-Basso[1,3]

Received: 15 January 2024 / Revised: 22 June 2024 / Accepted: 24 June 2024 / Published online: 15 July 2024

Abstract
The internet use among children and adolescents has increased massively recently. This situation has promoted harmful situations such as cyberbullying, which is becoming a worldwide problem that entails serious consequences for well-being. The detection of these attitudes is essential to prevent and act accordingly. Groundbreaking techniques based on deep learning, like pre-trained language models, have achieved state-of-the-art results in many downstream Natural Language Processing tasks. This paper presents a simple but effective approach to improve the detection of cyberbullying situations by fine-tuning a pre-trained sentence transformer language model. We experimented on three datasets, and the results surpassed the state-of-the-art results. The approach could help prevent cyberbullying, filter those messages, and detect those children involved in the situation, providing opportunities to develop intervention programs to address this problem.

Keywords Cyberbullying detection · Deep learning · Interdisciplinary projects · Sentence embeddings · NLP · Media in education

1 Introduction

Nowadays, society communicates through technology, thanks mainly to the widespread use of social networks. These new ways of relating to each other have changed the habits and characteristics of communication, such as the ease of information exchange, the existence of global receivers, and the accessibility of these technologies to all social sectors.

In these environments, violent and dangerous attitudes constitute a field of study for several disciplines as anyone has the opportunity to harm and bully people, being supported by anonymity easily.

Currently, cyberbullying is not only a social problem but also a health problem as it is associated with many negative consequences for victims, bullies and witnesses (Evangelio et al. 2022). Some of them go from mental health problems such as depression or anxiety (Cole et al. 2016; Wang et al. 2019), to cigarette smoking (Cassiani-Miranda et al. 2022; Kim et al. 2019), post-traumatic stress disorder (Cassiani-Miranda et al. 2022; Liu et al. 2020; Ranney et al. 2016), and even self-harm and suicidal thoughts (Kim et al. 2019), with some of them pursuing in long term (Görzig et al. 2017).

This becomes even more disturbing, given the prevalence of these attitudes. Whether this prevalence had already been increasing in recent years, the closure of educational centres during COVID-19 has made the worldwide situation worse, with a recent study conducted by the European Commission emphasising that almost half of children in 11 European countries have experienced cyberbullying situations (Lobe et al. 2021).

Jesica Gómez-Sánchez and Carlos Fernandez-Basso have contributed equally to this work.

✉ Jesica Gómez-Sánchez
gomezjs@ugr.es; jesica.gomez@ucl.ac.uk

Karel Gutiérrez-Batista
karel@decsai.ugr.es

Carlos Fernandez-Basso
cjferba@decsai.ugr.es; carlos.basso@ucl.ac.uk

[1] Department of Computer Science and Artificial Intelligence, University of Granada, Granada 18014, Andalucia, Spain

[2] Department of Developmental and Educational Psychology, University of Granada, 18014 Granada, Andalucia, Spain

[3] Division of Psychology and Language Sciences, Department of Experimental Psychology, University College London, 26 Bedford Way, London WC1H0NS, UK

Early detection of these attitudes could help prevent them and their harmful consequences. However, detecting this behaviour in these environments is very complex since many messages are sent every minute. Automating the detection process stands out as a viable approach for identifying these instances at an early stage.

Machine learning techniques have been widely used for cyberbullying detection (Ali and Syed 2020; Balakrishnan et al. 2020; Muneer and Fati 2020; Ratnayaka et al. 2020; Salawu et al. 2017. Recent studies have used deep learning models to address the cyberbullying detection challenge (Dadvar and Eckert 2020; Kargutkar and Chitre 2020; Pericherla and Ilavarasan 2021; Zhao and Mao 2017), as they can outperform the classical machine learning models in terms of performance. Considering the above-mentioned and the fact that deep learning techniques like pre-trained language models achieve outstanding results in downstream Natural Language Processing (NLP) tasks, these techniques are the most suitable for addressing these problems.

This paper proposes a feasible approach to detect cyberbullying by fine-tuning a pre-trained sentence transformer language model. Our proposal is based on pre-trained language models. Precisely, we fine-tune Sentence-BERT (Reimers and Gurevych 2019), a pre-trained language model for text representation at the sentence level. To fine-tune the model, we generate multiple *training datasets*. Each dataset instance consists of a pair of sentences, with the possibility that the sentences belong to the same class or different classes (Bullying or Not Bullying). The approach allows us to improve the performance of any classification algorithm in cyberbullying detection, as our main contribution comprises creating a training set for fine-tuning the model in order to obtain more meaningful sentence embeddings. Our approach is compared with the results presented in two state-of-the-art studies, as can be seen in Sect. 4, and our approach surpasses the state-of-the-art results. We must remark that our proposal is entirely automatic and supervised.

1.1 Contributions

The main contributions of this work are detailed as follows:

- Our primary contribution is improving cyberbullying detection in social networks by fine-tuning a pre-trained language model. Specifically, we fine-tune Sentence-BERT, a pre-trained language model for text representation at the sentence level. To do that, we create several *training datasets* where each instance comprises a pair of sentences, which ones might be of the same class or different classes (Bullying or Not Bullying).
- By creating meaningful training sets for fine-tuning the model, we improve the ability of the model to capture nuanced features of cyberbullying-related text. Furthermore, we conduct experiments using training sets of varying sizes to assess the model's performance under different conditions.
- Our approach is benchmarked against results from two state-of-the-art studies, comprehensively evaluating its effectiveness. The comparisons, detailed in Sect. 4, demonstrate that our automatic and supervised approach surpasses the state-of-the-art results. This validation reinforces the credibility and superiority of our proposed cyberbullying detection model.
- In addition to evaluating the proposed approach from the classification point of view, we also evaluate our approach in terms of semantics. We use the silhouette coefficient to evaluate the class distribution of the different datasets used.
- Our proposal is fully automated and operates supervised, making it accessible and efficient for practical use. This aspect is crucial for real-time monitoring and early intervention. The proposal also leverages the power of pre-trained language models, ensuring that it remains adaptable to evolving cyberbullying patterns and maintains relevance in dynamic online environments.

These three contributions collectively form a robust and effective strategy for addressing the complex challenge of cyberbullying detection in online communication platforms. The fine-tuned language model, benchmarking against state-of-the-art studies, and the automated, supervised framework collectively contribute to advancing the field of cyberbullying detection.

The rest of the paper is organised as follows: Sect. 2 outlines the key work previously conducted in this research area. Section 3 introduces the proposed approach for cyberbullying detection. Section 4 and 5 present the details of the experimentation and discussion of the yielded results, respectively. Finally, Sect. 6 provides the conclusions from the results.

2 Related work

This paper presents a system based on deep learning techniques to automatically detect text messages associated with bullying. Various tools and techniques have been presented in the literature to detect which texts correspond to cyberbullying and which do not.

2.1 Cyberbullying detection

It is undeniable that the popularity of social networks has grown significantly in the last few years. This has caused the increase of cyberbullying as a severe problem as well. Some studies of cyberbullying focus on the psychological

effects on victims and were mainly conducted by social scientists and psychologists using extensive surveys Langos (2012); Pericherla and Ilavarasan (2021); Sui (2015); Zhao and Mao (2017). These efforts facilitate our understanding of cyberbullying. However, this approach is time-consuming and may need to be revised to detect cyberbullying automatically. As machine learning has been increasing its popularity in the last years, the computational study of cyberbullying has enlarged the interest of researchers.

Many studies have addressed cyberbullying detection using classical machine-learning techniques. Ali and Syed (2020) propose an approach for cyberbullying detection, taking into account the sarcasm in the utterances. The results prove that the Support Vector Machine classifier performed better than other classifiers. Muneer and Fati (2020) compare different machine learning techniques for cyberbullying detection on Twitter. The authors used seven machine learning classifiers on a dataset comprising 37,373 tweets from Twitter to perform a benchmarking. The comparison concluded that Logistic Regression performed better than the others regarding accuracy and f1-score.

Other researchers have considered enhancing automatic cyberbullying detection using user features like personality, sentiment, emotion, role, etc. That is the case in Balakrishnan et al. (2020), where the authors propose to use users' psychological features and machine learning to improve cyberbullying detection. They used Big Five and Dark Triad models to determine user personalities and machine learning classifiers to classify the texts into one of four categories: bully, aggressor, spammer and normal. The experiments showcase that cyberbullying detection can be improved using personalities and sentiments, unlike when emotion was used. Ratnayaka et al. (2020) propose a novel approach to enhancing cyberbullying detection through role modelling. The authors perform multi-class classification to detect different participant roles (e.g. victim, harasser), using it later to improve cyberbullying and user role classification.

2.2 Deep learning-based cyberbullying detection

As was stated before, deep learning techniques have achieved state-of-the-art results in downstream NLP tasks. Since cyberbullying analysis is performed on textual data, several studies have used deep learning-based models to tackle this problem. That is the case in Dadvar and Eckert (2020), in which the authors reproduce and validate the deep learning-based models proposed by Agrawal and Awekar (2018) and on three datasets. The authors also apply the models to a new dataset where the deep learning-based models surpass the classical machine learning models when applied to the same dataset. Kargutkar and Chitre (2020) use machine learning techniques to study cyberbullying. They propose a system that uses a Convolutional Neural Network to mitigate cyberbullying episodes in social networks.

Many deep learning-based models in the state-of-the-art have approached cyberbullying, improving text representation. These kinds of works mainly focus on generating embeddings to obtain a better representation of the texts.

Zhao and Mao (2017), propose a representation learning method called semantic-enhanced marginalised denoising auto-encoder (smSDA). They perform a semantic extension of the stacked denoising autoencoder (SDA) model based on domain knowledge and word embeddings. This method allows learning a robust and discriminative representation of text. They conducted experiments on two datasets (Twitter and MySpace) to prove the proposal's feasibility, surpassing the existing baseline results.

In the same way, Pericherla and Ilavarasan (2021) detect cyberbullying through an approach focused on transforming network-based word embeddings. The approach consists of a pre-trained language model (RoBERTa Liu et al. (2019)) to yield word embeddings and a classifier (Light Gradient Boosting Machine) to classify the texts. The authors demonstrated the viability of the proposal on a dataset from Twitter. The results outperform classical machine learning techniques like Logistic Regression and Support Vector Machine and deep learning-based models like Convolutional Neural Networks.

According to the study depicted in Paul and Saha (2020), the authors propose a novel BERT-based approach for cyberbullying identification (CyberBERT). Specifically, the main findings of the proposal include fine-tuning BERT, using BERT as a simpler neural model that provides competitive accuracy at lower computational cost and testing the feasibility of the model on three real-world cyberbullying datasets. The former research is one of the most similar to the approach developed in this paper. The main difference relies on in Paul and Saha (2020), the authors fine-tune BERT Devlin et al. (2019), while we fine-tune SBERT Reimers and Gurevych (2019). Table 1 showcases the critical differences between both models.

The approach presented in this paper improves cyberbullying detection by fine-tuning a pre-trained language model like SBERT tailored for tasks requiring efficient and accurate sentence comparisons. This process enables obtaining sentence embeddings that contain far more accurate meanings from the semantic point of view, allowing the classifiers to improve automatic and semantic evaluation performance. As can be seen, the studies presented in Pericherla and Ilavarasan (2021) and Zhao and Mao (2017) address the cyberbullying problem as we do. Therefore, we have taken these studies as a reference for benchmarking, as their results outperform the state-of-the-art.

Table 1 Differences between BERT and SBERT

Model	goal	Usage	Type of learning
BERT	General-purpose language representation	Generates contextual word embeddings	Self-supervised learning (masked language modeling and next sentence prediction)
SBERT	Sentence embeddings optimised for semantic similarity tasks	Generates fixed-size sentence embeddings	Contrastive learning on sentence pairs

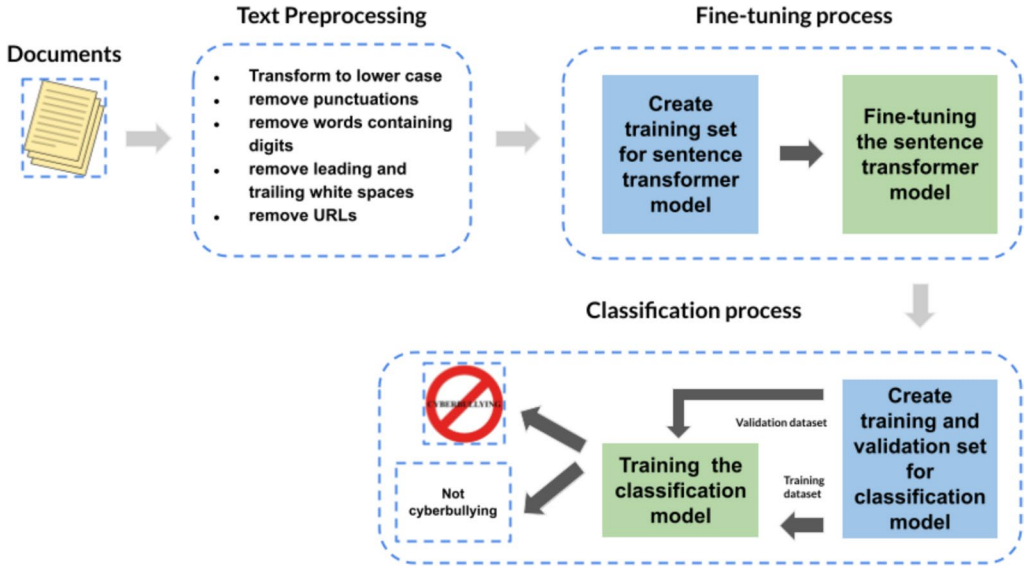

Fig. 1 General overview of the proposed approach

3 Cyberbullying detection approach

The present work aims to improve cyberbullying detection, focusing on fine-tuning a sentence-level embedding language model. This section details all the features and steps of our proposal for automatic cyberbullying detection. Figure 1 shows a general overview of our approach. For a better understanding, we have divided our proposal into three steps: *Textual data preprocessing*, *sentence transformer model Fine-tuning*, and *Cyberbullying classification*. In the following sections, we will dive deep into each step.

3.1 Textual data preprocessing

We clean and transform the textual data during the preprocessing step, which is a decisive phase for nearly all downstream NLP tasks. To do this, we applied some filters to represent data properly for further analysis. We applied the filters stated below:

- Convert all text to lowercase,
- delete any punctuation mark,
- remove any words containing digits,
- remove leading and trailing white spaces,
- delete URLs.

No advanced preprocessing techniques, such as lemma stemming, n-gram analysis, etc., are used. This is because we need the text to be as original as possible, considering that, as we will explain in the next section, we use a pre-trained language model to encode the texts.

3.2 Sentence transformer model fine-tuning

The explosion of deep learning-based algorithms in recent years has dramatically impacted the performance of most downstream NLP tasks. One of the breakthroughs has been using deep learning techniques to obtain pre-trained language models that properly represent textual data for later analysis.

3.2.1 Pre-trained language model

In the state-of-the-art, several pre-trained models can be found. These models represent the texts through featured dense vectors of fixed length called embeddings. The semantic meaning of the textual data is contained in each vector.

Pre-trained models can be classified by taking into account the fine-grained level used for representing the texts (character embeddings Bojanowski et al. (2017), word embeddings Devlin et al. (2019); Mikolov et al. (2013); Pennington et al. (2014); Peters et al. (2018); Radford et al. (2018) and sentence embeddings Reimers and Gurevych (2019)). It must be highlighted that word embeddings representation is the most widely used.

Word embeddings (WE) can be defined as contextualised and static. Static WE are the oldest sort of WE, where the embeddings are generated from a wide corpus, but a word used in different contexts always has the same representation. Examples of static WE are word2vec Mikolov et al. (2013), GloVe Pennington et al. (2014), and fastText Bojanowski et al. (2017). On the contrary, contextualised word embeddings are context-aware embeddings that capture the different meanings of a word considering the context. BERT Devlin et al. (2019), RoBERTa Liu et al. (2019), ELMo Peters et al. (2018) and OpenAI GPT Radford et al. (2018) are examples of contextualised word embeddings.

In particular, we use Sentence-BERT (SBERT) Reimers and Gurevych (2019) for text representation. SBERT (based on BERT) enables the representation of sentences, paragraphs, and images through dense vectors. SBERT aims to generate meaningful sentence embeddings from the semantic point of view in order to compare those embeddings using cosine similarity. SBERT offers an extensive list of pre-trained models for more than 100 languages. The models are based on transformer networks such as BERT and others. For our approach, we have used this sentence-level pre-trained model as it is suitable to accomplish the objective of this work.

3.2.2 Fine-tuning process

Pre-trained language models are built from humongous corpus like Wikipedia, Google, etc. Although the straightforward use of pre-trained models can yield acceptable results in many downstream NLP tasks, fine-tuning is an excellent technique to improve the results. By fine-tuning the model, we transfer the knowledge learned from the general corpus to a specific domain. In this way, the model can achieve better results.

To do that, we create a training set for fine-tuning the model, where each instance comprises a pair of sentences and real numbers indicating the similarity between the pair of sentences. This training set for fine-tuning the model is created from the training set obtained when the whole dataset is split for classification purposes (training and validation dataset). For better understanding, the pseudocode for generating the training set for fine-tuning the language model is described in the Algorithm 1.

Algorithm 1 Creating the training set for fine-tuning the pre-trained model

Require: x_train //Documents of the training dataset.
$labels$ //Labels for each document of the training dataset.
Ensure: Create the $train\_examples$ for fine-tuning the model
$train\_examples \leftarrow \emptyset$
$numClassesList \leftarrow UNIQUE(labels)$
$idxLabels \leftarrow \emptyset$
for i in $numClassesList$ **do**
$idxLabels.APPEND(INDEX(labels = i))$
end for
for idx in $RANGE(LENGTH(x\_train))$ **do**
$currentSentence \leftarrow x\_traing[idx]$
$currentLabel \leftarrow labels[idx]$
$posIdx \leftarrow RANDOM(idxLabels[numClassesList = label])$
$posSentence \leftarrow x\_train[posIdx]$
$train\_examples.APPEND(PAIR([currentSentence, posSentence], 1.0))$
$negIdx \leftarrow (labels \neq label)$
$negSentence \leftarrow x\_train[RANDOM(negIdx)]$
$train\_examples.APPEND(PAIR([currentSentence, negSentence], 0.0))$
end for
return $train\_examples$

The above algorithm shows that for every sentence on the original dataset (*x_train*), we randomly select a sentence of the same class and a sentence of a different class. Then, we append one pair per sentence to the training set for fine-tuning (*train_examples*). The only difference between these pairs is that when the sentence belongs to the same class, the degree of similarity is 1; when the sentence belongs to a different class, the degree of similarity is 0. Thus, the model learns the semantic representation of the documents, considering the degree of similarity passed. We must highlight that this process can be carried out several times (called **Iterations**), thus generating a more extensive training set to train the model. We have tried this during the experimentation to demonstrate whether the number of iterations affects the final results.

As was mentioned above, SBERT provides a broad list of pre-trained models. We will use *all-mpnet-base-v2*,[1] which is a pre-trained model that maps sentences and paragraphs to a 768-dimensional dense vector space. Clustering or semantic search are some tasks in which this representation can be used.

3.3 Cyberbullying classification

Finally, once the sentence transformer model has been fine-tuned, we classify the texts into bullying or not bullying using its vector representation. In this phase, we could use any classification algorithm. In our case, we have used three

[1] https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

Table 2 Datasets description

Dataset	Documents	Classes	Class distribution	
			Bullying	Not bullying
`bullyingV3.0`	3889	2	1008	2881
`myspace`	2029	2	426	1603
`hate-speech`	10360	2	2723	7637

Table 3 Description of the configurations for the fine-tuning process

Dataset	One iteration (training dataset)	Two iterations (training dataset)	Three iterations (training dataset)
bullyingV3.0	5444	10888	16332
MySpace	2840	5680	8520
hate-speech	14462	28924	43386

algorithms: Support Vector Machine (SVM) Chang and Lin (2011), Light Gradient Boosting Model (LGBM) Ke et al. (2017), and Logistic Regression (LR) McCullagh and Nelder (1989). As detailed in Sect. 4.2, we have chosen these algorithms based on the literature we will use for comparison.

4 Experimental details

In this section, we delve into the specifics of our experimentation. Initially, we introduce the datasets utilised, followed by a comprehensive overview of the procedural steps taken throughout the experiment.

4.1 Datasets

We have selected three datasets, two from Twitter and the other comprising posts from MySpace. We must highlight that all the used datasets are related to the topic addressed in this work, *Cyberbullying* and *Hate speech* detection. Table 2 shows a detailed description of the datasets.

4.1.1 BullyingV3.0

`BullyingV3.0` Sui (2015) is a dataset for the study of bullying in social media, composed of tweets crawled by the public Twitter stream API (from August 6, 2011, to August 31, 2011). It is the third version released in June 2015 with 7,321 manually annotated tweets.[2]

The tweets were collected so that each tweet contains at least one of the following keywords: *bully*, *bullied*, *bullying*. The dataset contains many features, among them the *tweet ids*, which is helpful to retrieve tweets from Twitter using Twitter APIs, and *BullyingTraces*, which means if a tweet is annotated as a bullying trace or not.

We have used the Python script provided with the dataset to download the text of each tweet. As can be seen in Table 2, after retrieving the texts and applying some preprocessing steps, the `bullyingV3.0` dataset contains 3,889 tweets. This is mainly because many tweets have been removed from Twitter.

4.1.2 MySpace

As `BullyingV3.0`, myspace Bayzick et al. (2011) is a dataset oriented to cyberbullying and hate speech detection. The dataset comprises a small subset of data crawled from myspace.[3] The data has been manually labelled for bullying content by three independent coders.

Each post on the dataset is considered a single body of chat text posted by a user, and each body of text could contain multiple sentences or even several paragraphs. Due to the interactive nature of cyberbullying, the conversations were captured using a moving window of 10 posts to obtain context Bayzick et al. (2011). Each window has been labelled as cyberbullying if at least two of the three annotators labelled it as cyberbullying.

4.1.3 Hate-speech

Evenly, `hate-speech` is a dataset of tweets annotated for hate speech Waseem and Hovy (2016). As `bullyingV3.0`, the dataset provides the *ID* and the *annotation*, and we have used the Twitter APIs to obtain the text of the tweets.

Unlike the previous datasets, the tweets of `hate-speech` have been annotated in three classes: *none*, *sexism*, and *racism*. Taking into account the primary goal of this work, for simplicity, we have grouped the tweets annotated as *sexism* and *racism* into a more general group annotated as *cyberbullying*.

4.2 Experimental setup

For each dataset, we have conducted a set of experiments. In the following, we state the steps that comprise the experimentation process:

1. Firstly, the texts are preprocessed following the steps stated in Sect. 3.1.
2. Once the texts have been preprocessed, the sentence transformer model is fine-tuned. As noted in the fine-

[2] https://research.cs.wisc.edu/bullying/data.html.

[3] https://www.chatcoder.com/data.html.

Table 4 Results in terms of *accuracy*, *f1-score*, and *recall* (percent) for all the configurations

Dataset			One iteration			Two iterations			Three iterations		
	Algorithm	Model/ measure	Accuracy	f1-score	Recall	Accuracy	f1-score	Recall	Accuracy	f1-score	Recall
bullyingV3.0	SVM	Without fitting	83.2	83.1	83.1	82.94	83.2	83.2	82.43	82.1	83.2
		Fitted	83.8	**84.2**	**84.2**	84.91	85.1	85.1	83.97	**84.3**	**84.4**
	LGBM	Without fitting	83.03	83.1	83.1	81.74	82.1	82.1	81.83	81.2	82.3
		Fitted	83.4	83.1	83.1	84.91	85.2	85.2	**83.98**	84.1	**84.4**
	LR	Without fitting	82.94	83.2	83.2	81.31	81.3	81.3	81.41	80.2	81.3
		Fitted	**83.71**	84.1	84.1	**85.1**	**85.4**	**85.4**	83.8	84.1	84.3
myspace groups	SVM	Without fitting	**92.77**	**93.1**	**93.2**	92.77	92.33	93.4	93.1	93.25	93.3
		Fitted	91.62	92.2	92.2	95.84	96.1	96.3	95.73	95.2	**96.3**
	LGBM	Without fitting	90.96	91.02	91.1	93.26	93.24	93.33	93.1	92.6	93.2
		Fitted	91.95	92.08	92.2	96.22	**96.21**	**96.41**	**96.39**	**95.8**	96.12
	LR	Without fitting	86.37	85.13	86.2	85.05	83.3	85.2	86.37	85.3	86.4
		Fitted	90.97	91.18	91.22	**96.25**	96.09	96.3	95.24	94.6	95.5
hate-speech	SVM	Without fitting	87.06	86.02	87.2	86.47	86.12	86.3	86.61	86.15	87.3
		Fitted	**88.57**	88.1	88.2	88.03	**88.2**	**88.31**	**87.71**	**88.27**	**88.33**
	LGBM	Without fitting	85.61	84.2	85.1	84.25	83.19	84.3	84.61	84.6	85.4
		Fitted	88.12	**88.18**	88.22	**88.13**	**88.2**	88.25	87.58	87.6	87.92
	LR	Without fitting	85.22	84.2	85.32	85.05	84.3	85.2	85.01	84.13	85.8
		Fitted	88.23	88.11	**88.43**	88.1	87.9	88.04	87.51	87.3	88.15

Best results per dataset and iteration are highlighted in bold

tuning process section, we used a pre-trained model: *all-mpnet-base-v2*. For each dataset, the fine-tuning process is carried out, taking into account different-sized training datasets (see Sect. 3.2). Table 3 showcases the different configurations for fine-tuning the sentence transformer model.

3. Finally, we apply three machine learning algorithms to classify the texts (SVM, LGBM, and LR).

In all cases, we have used k-fold cross-validation ($k = 5$) to reduce the bias of the model skill. SVM was selected because one of the state-of-the-art works we will compare with used this algorithm Zhao and Mao (2017). In the same way, as for the previous case, LGBM was selected because we will establish a benchmarking with the results presented in Pericherla and Ilavarasan (2021). In addition, we have also used LG to strengthen our experimentation.

5 Results and discussion

5.1 Automatic evaluation measures

As stated, we will compare our approach with two state-of-the-art works to assess its feasibility. These works were chosen as, in both cases, the authors used approaches similar to ours (based on deep learning techniques). Besides, they performed a thorough comparison with the most relevant studies regarding cyberbullying, and we must highlight that in both cases, the obtained results surpassed the state-of-the-art results.

Before comparing our results with the state-of-the-art, we will analyse all the values obtained during the experimentation. Table 4 shows the experimentation results with the best results of each. This table can be challenging to understand due to the number of results. However, the key here is to note whether there are differences in the obtained results when the sentence transformer model is **fine-tuned** or **not** and whether differences exist, taking into account the size of the dataset used for fine-tuning the sentence transformer model (**number of iterations**).

A statistical analysis was conducted to detect any significant differences for a better understanding and a more detailed analysis supporting the obtained results. Figure 2 depicts the boxplots for accuracy (a), f1-score (b) and recall (c) regarding the two approaches (**model fitted** and **without fitting**). At first sight, it can be seen that the best results for all metrics are yielded when the sentence transformer model has been fine-tuned (model fitted). On the other hand, Fig. 3 shows the boxplots for accuracy (a), f1-score (b) and recall (c) regarding the size of the dataset used to fine-tune the sentence transformer model (1, 2 and 3 iterations). In this case, the differences can not be appreciated at first glance.

We used two non-parametric tests for the statistical analysis: Wilcoxon Wilcoxon (1945) and Kruskal–Wallis Kruskal

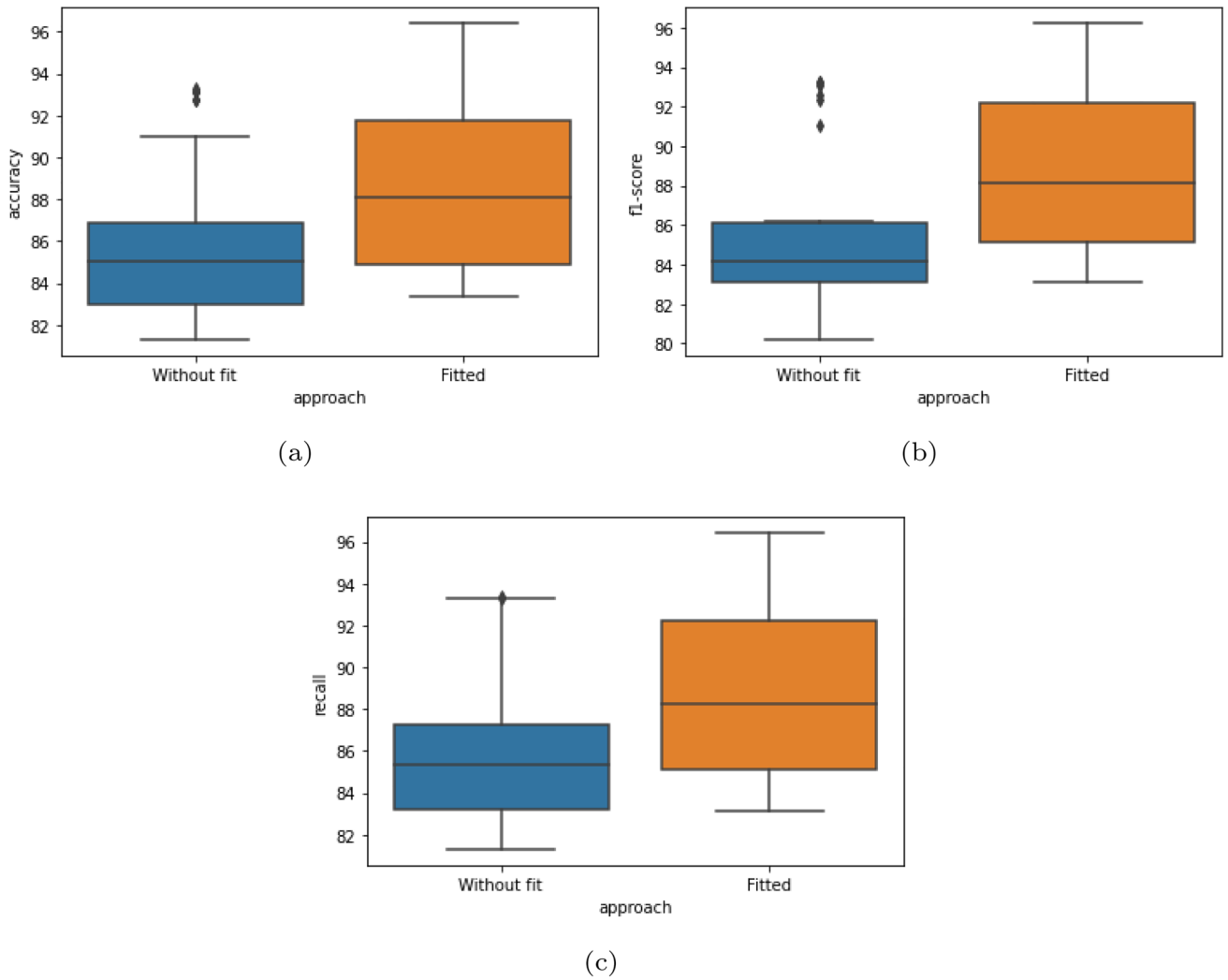

Fig. 2 Statistical analysis using the different models (fitted and without fitting) and the measures (accuracy, f1-score and recall) on all the datasets

and Wallis (1952). In the first case, we have used Wilcoxon as there are only two groups (**Fitted** and **Without fitting**). Considering the *p*-values in Table 5, we could assume that both groups are statistically different for all the metrics, where the best results are obtained when the model is fitted. In the second case, we have used Kruskal-Wallis because, in this case, there are three groups (**1**, **2** and **3 iterations**). Unlike the first case, we can see in Table 5 that no significant differences exist between groups for any metrics.

As stated before, we have compared our approach with two promising state-of-the-art works. Pericherla and Ilavarasan (2021) presented a transformer network-based word embedding approach for autonomous cyberbullying detection. Table 6 compares our approach and the results yielded in the paper above. To make a fair comparison, we have used the same dataset (`hate-speech`), the same classification algorithm (LGBM) and the same measures (precision, f1-score and recall). As seen in Table 6, our proposal considerably surpasses the results achieved in the paper for all the measures (higher values in bold).

We have also established a benchmarking with the results presented in Zhao and Mao (2017). As in the previous case, we have used the same algorithm (SVM), the same datasets (`bullyingV3.0` and `myspace`) and the same measures (accuracy and f1-score). Table 7 shows the results of our approach and the results obtained by smSDA. Our proposal outperforms the smSDA algorithm in all the cases except for the accuracy measure on the `bullyingV3.0` dataset. Although our result does not improve for this specific case, we must highlight that the results are very similar (smSDA: 84.9 and our approach: 84.22).

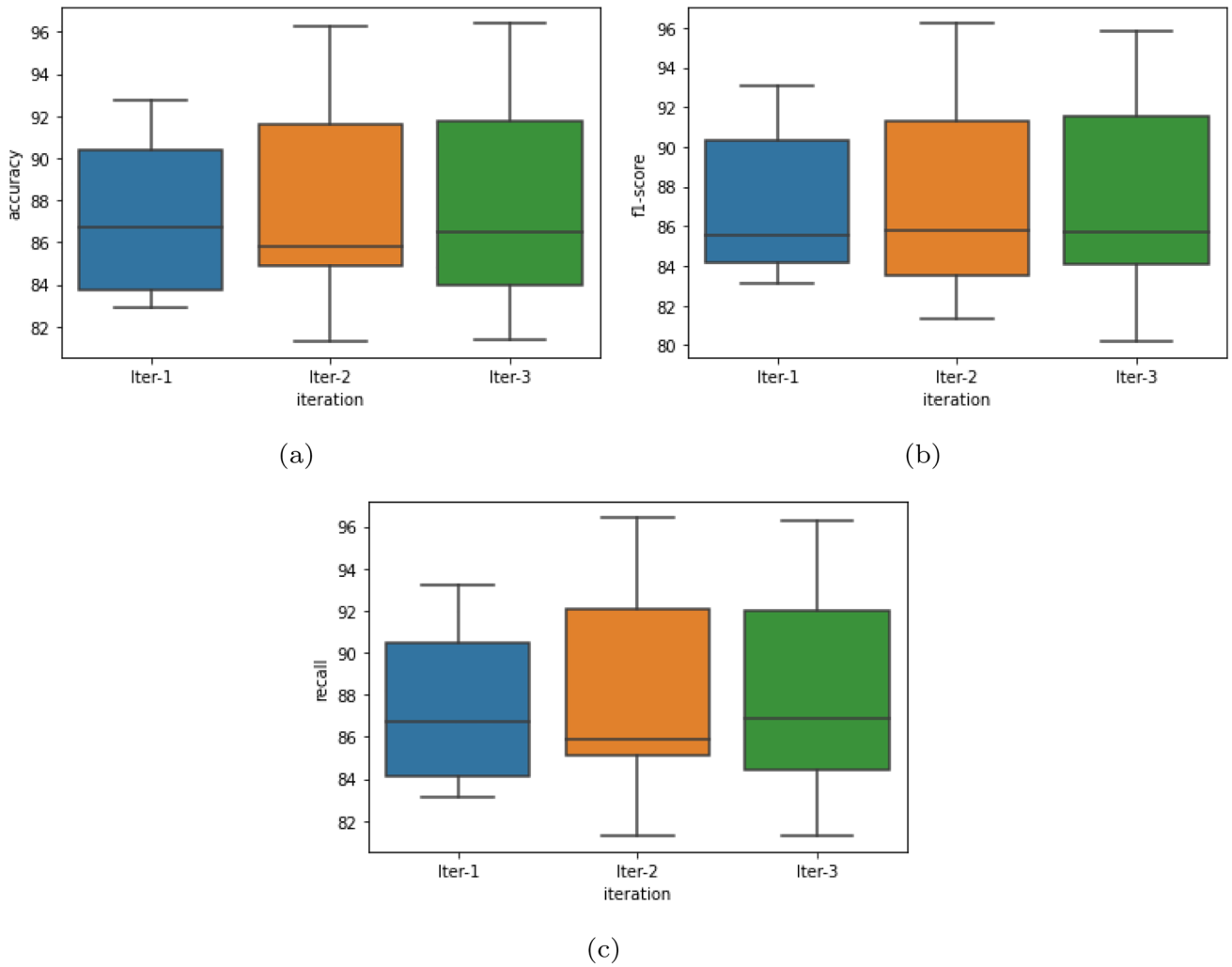

Fig. 3 Statistical analysis using the number of iterations (1,2 and 3) and the measures (accuracy, f1-score and recall) on all the datasets

Table 5 *p*-values for the statistical analysis using all metrics

Measure	Model (Fitted and without fitting)	Iterations (1,2 and 3)
Accuracy	0.00001	0.95245
F1-score	0.00001	0.85408
Recall	0.00001	0.84077

Table 6 Comparison in terms of precision, f1-score and recall with results presented in Pericherla and Ilavarasan (2021)

Measure/approach	Transformer network-based	Sentence-level embeddings
Precision	82.8	**87.67**
F1-score	80.2	**87.99**
Recall	78.4	**88.13**

Best results per dataset and iteration are highlighted in bold

Table 7 Comparison in terms of accuracy and f1-score with results presented in Zhao and Mao (2017)

Dataset/approach		smSDA	Sentence-level embeddings
`bullyingV3.0`	Accuracy	**84.9**	84.22
	F1-score	71.9	**84.5**
`myspace groups`	Accuracy	89.7	**94.4**
	F1-score	77.6	**94.5**

Best results per dataset are highlighted in bold

5.2 Semantic evaluation

As demonstrated in the previous section, our proposal for cyberbullying detection achieves excellent performance in terms of classification measures. This section evaluates our approach from the semantic point of view to ensure that the

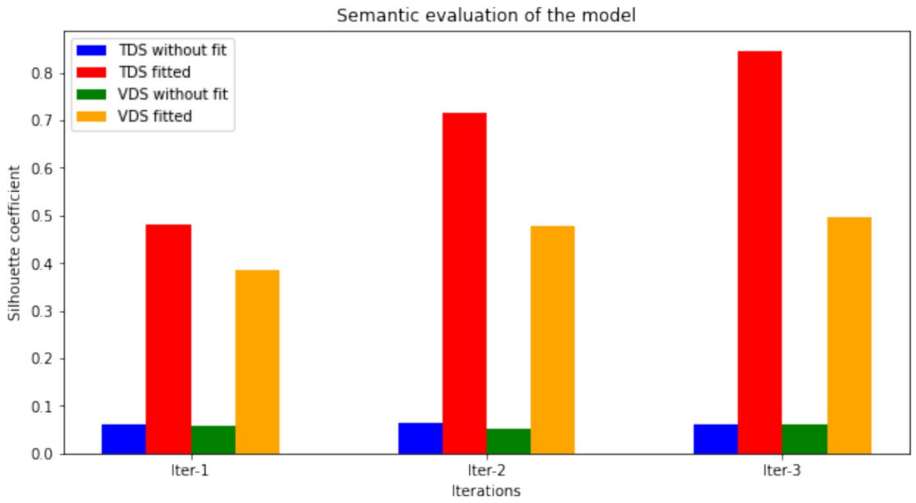

Fig. 4 Semantic evaluation for the proposed approach (silhouette coefficient) on the `bullyingV3.0` dataset

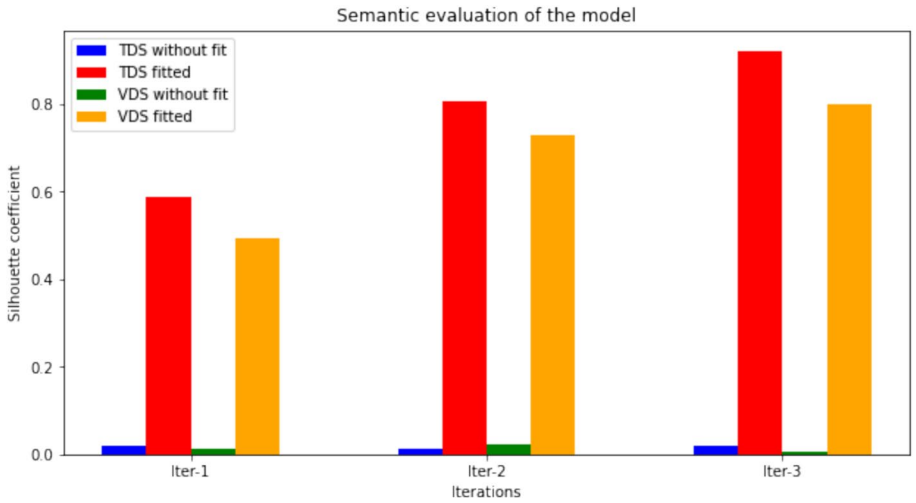

Fig. 5 Semantic evaluation for the proposed approach (silhouette coefficient) on myspace dataset

proposed approach also performs well from this perspective. This analysis proves that the model can learn better representations (sentence embeddings) of the texts during the fine-tuning process. In other words, we will compare the similarity between sentence embeddings of the same class when the model is fitted and without fit.

To do that, we have used the silhouette coefficient Rousseeuw (1987) (Eq. 1). This measure has been mainly used in clustering problems because it helps to determine the obtained clusters' quality. In our case, we have used it to evaluate the class distribution of the different subsets. In particular, we have used the four subsets created for the cyberbullying classification process: training dataset (TDS) without fit, training dataset fitted, validation dataset (VDS) without fit, and validation dataset fitted. The main difference between these pairs of subsets is that the sentence embeddings representing the texts in the fitted subsets are obtained after fine-tuning the model.

$$S(i) = \frac{a(i) - b(i)}{max\{a(i), b(i)\}} \tag{1}$$

Figures 4, 5, and 6 depict the values of silhouette coefficient for each subset, with different configurations (1, 2 and 3 iterations) on `bullyingV3.0`, `myspace`, and `hate-speech` datasets respectively. It can be observed that when the sentence transformer model is fine-tuned, the values of the coefficient improve considerably (**TDS fitted** and **VDS fitted**). Note that the values of the silhouette coefficient are in the range [-1,1], with those values closer to 1 being the best results.

Another important thing we must remark is that for all the cases, the values of silhouette coefficient for **TDS fitted** subset are better than the values for **VDS fitted**. This is mainly because the texts used to fine-tune the model belong to the **TDS fitted** subset.

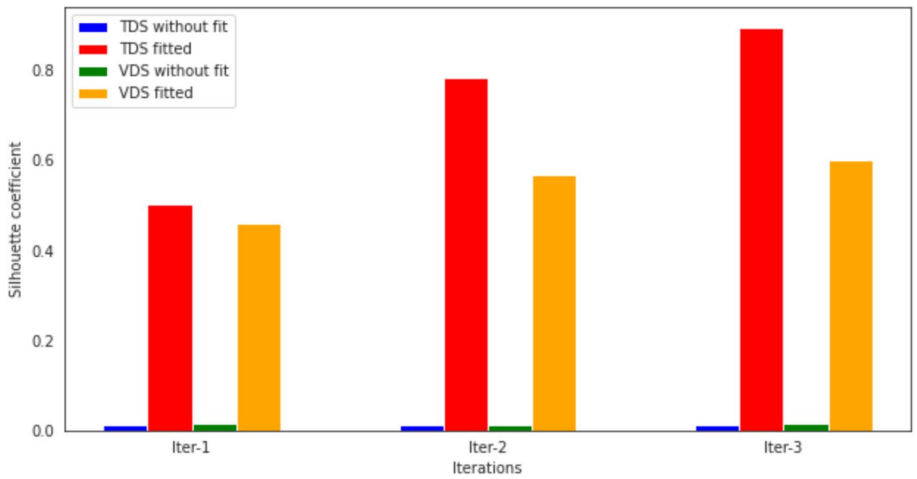

Fig. 6 Semantic evaluation for the proposed approach (silhouette coefficient) on the `hate-speech` dataset

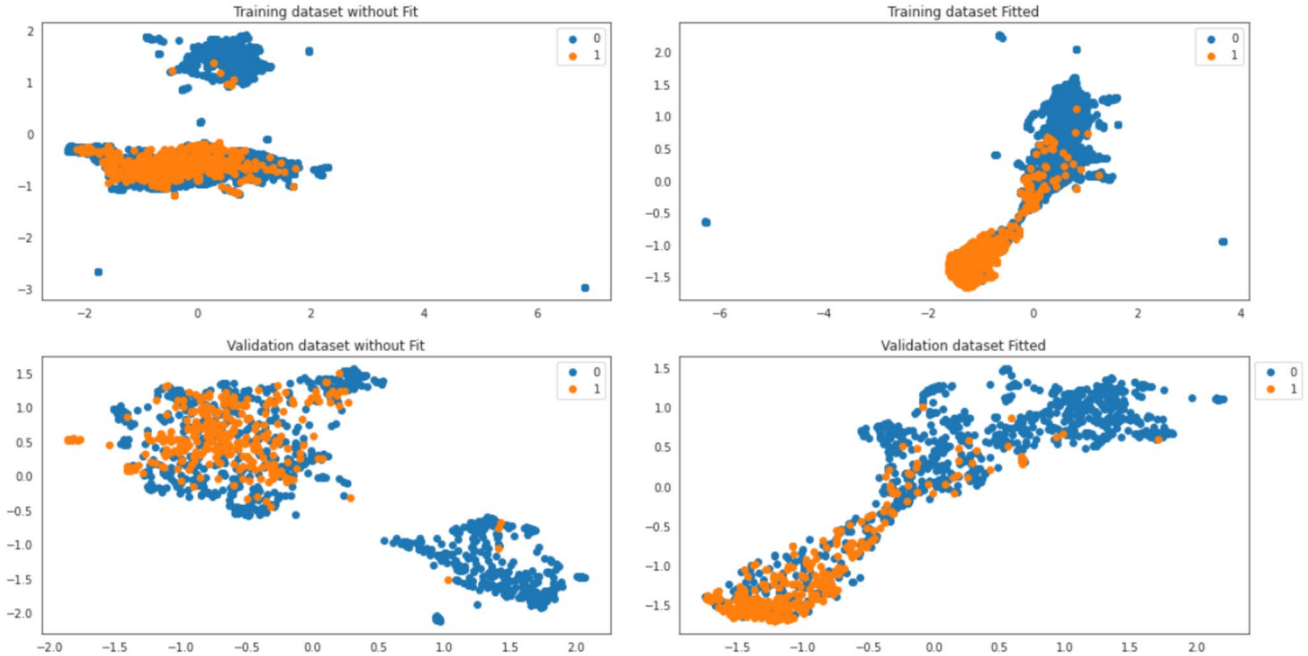

Fig. 7 Embedding representation of texts for one iteration on the `bullyingV3.0` dataset

5.3 Impact of preprocessing on model performance

In addition to the automatic and semantic evaluation, in this section, we analyse how the preprocessing pipeline affects the model performance. Although in Sect. 5.1, specifically in Table 4 and Fig. 3, we analyse and discuss the results taking into account the size of the dataset used for fine-tuning the model, we consider the importance of providing more insights into this aspect. This preprocessing step is one of the main contributions of our proposal and, thus, deserves more attention.

In order to provide more wisdom, we visually analyse the embedding representation of the documents for each iteration and each dataset. Figures 7, 8, and 9 showcase the embedding representation of the texts in `bullyingV3.0` after applying principal component analysis. These figures make it easier to conclude that as more data is used to fine-tune the model, the embeddings in the plots are better separated. Another important conclusion is that the embedding representation is better when fine-tuning (the embeddings are better separated, too), demonstrating the model's ability to learn nuanced cyberbullying. The same conclusion is derived from Figs. 10, 11, and 12, and Figs. 13, 14, and 15 for `myspace`, and `hate-speech` datasets respectively.

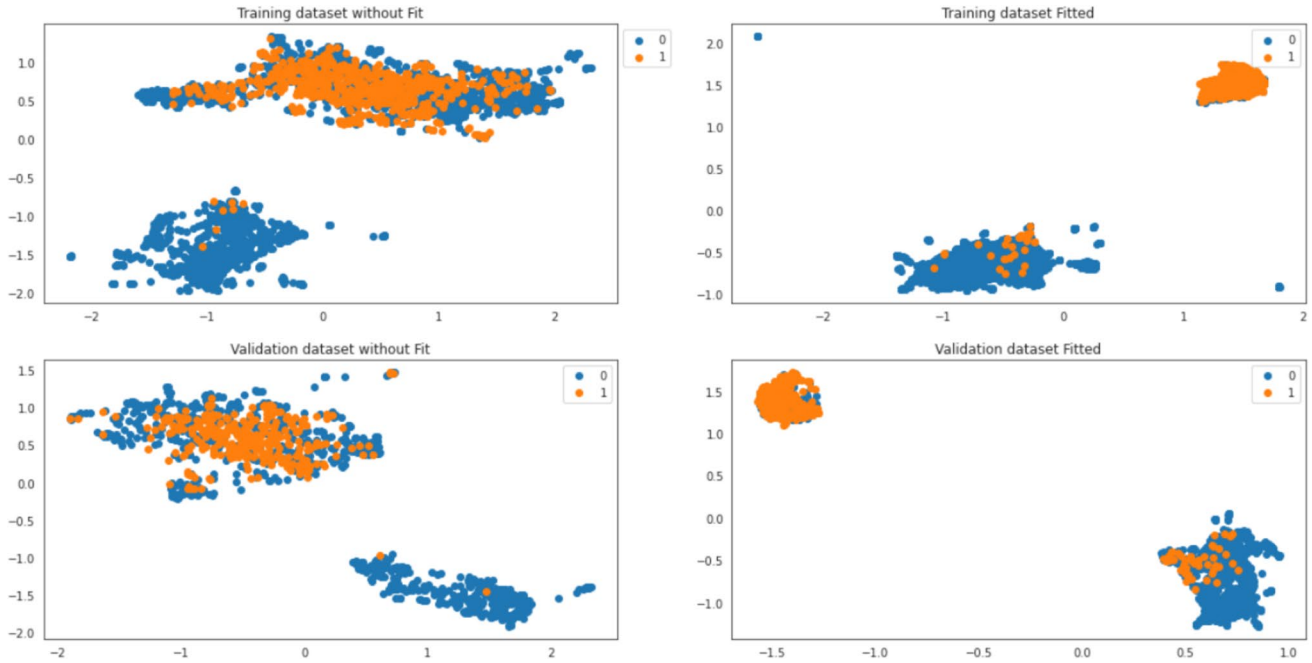

Fig. 8 Embedding representation of texts for two iterations on the `bullyingV3.0` dataset

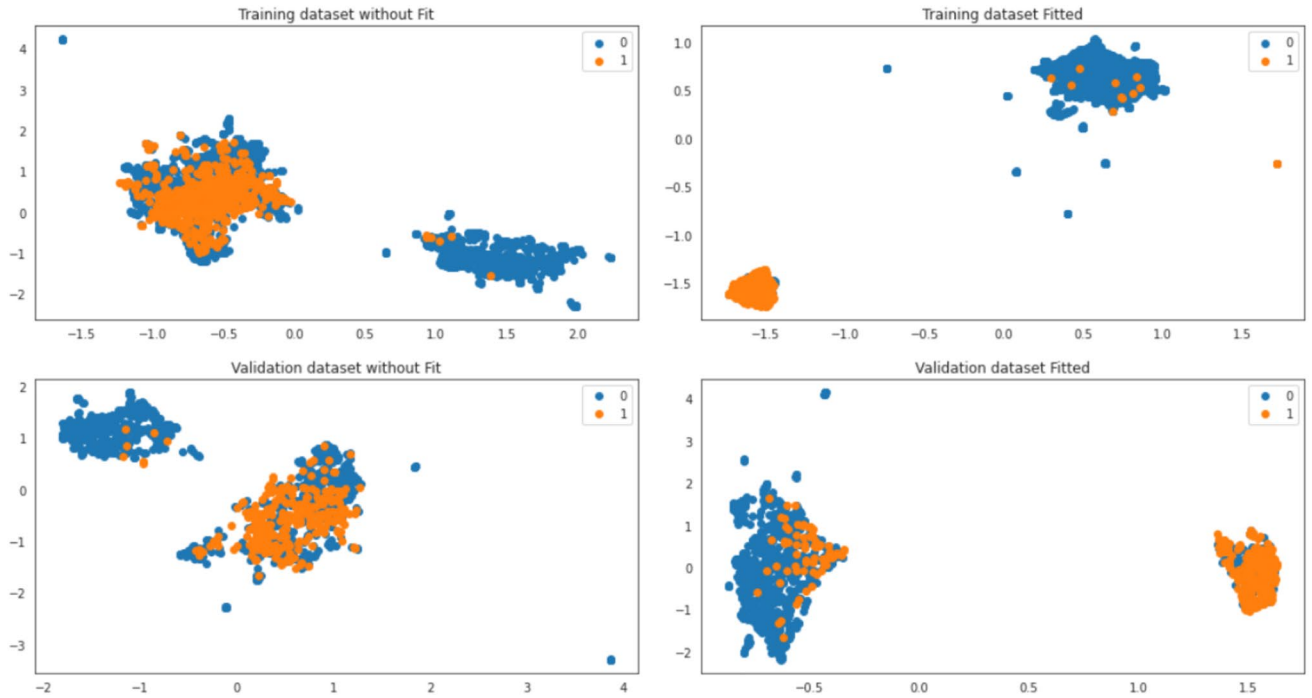

Fig. 9 Embedding representation of texts for three iterations on the `bullyingV3.0` dataset

Notice that for all the figures, the orange points represent texts classified as bullying statements. On the other hand, the blue ones represent texts classified as not bullying.

5.4 Analysis of model generalizability

As stated before, cyberbullying can be found in different systems and platforms regardless of language and culture.

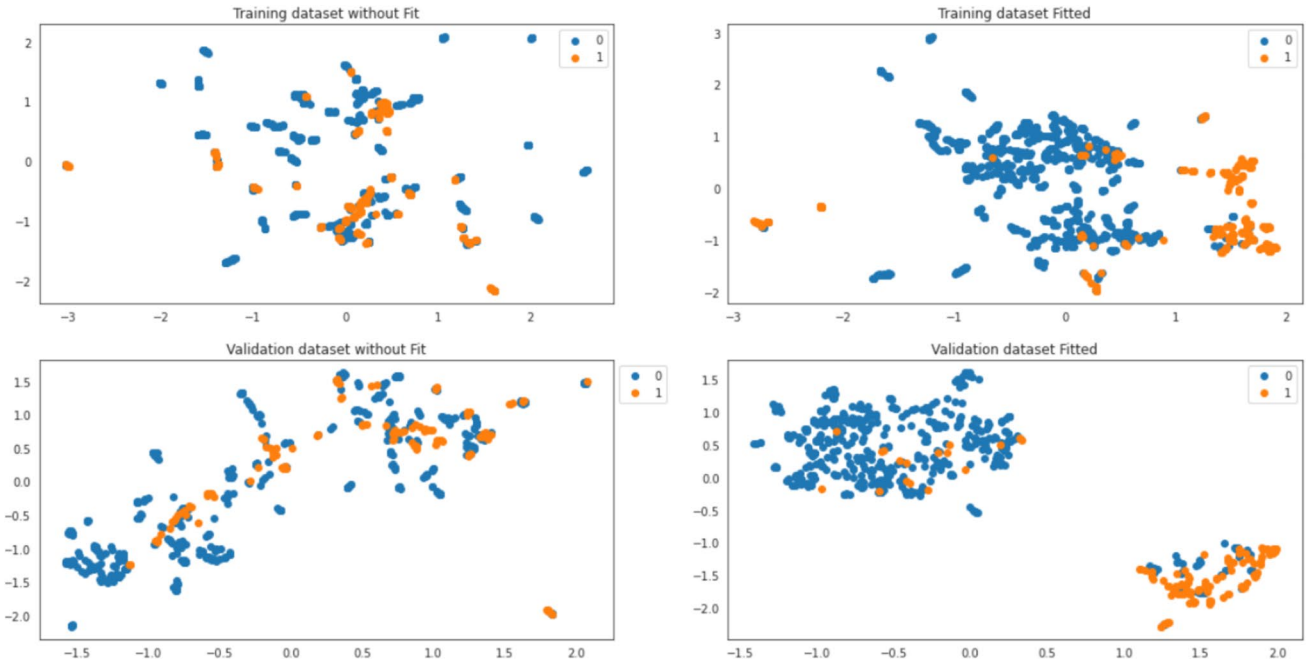

Fig. 10 Embedding representation of texts for one iteration on the `myspace` dataset

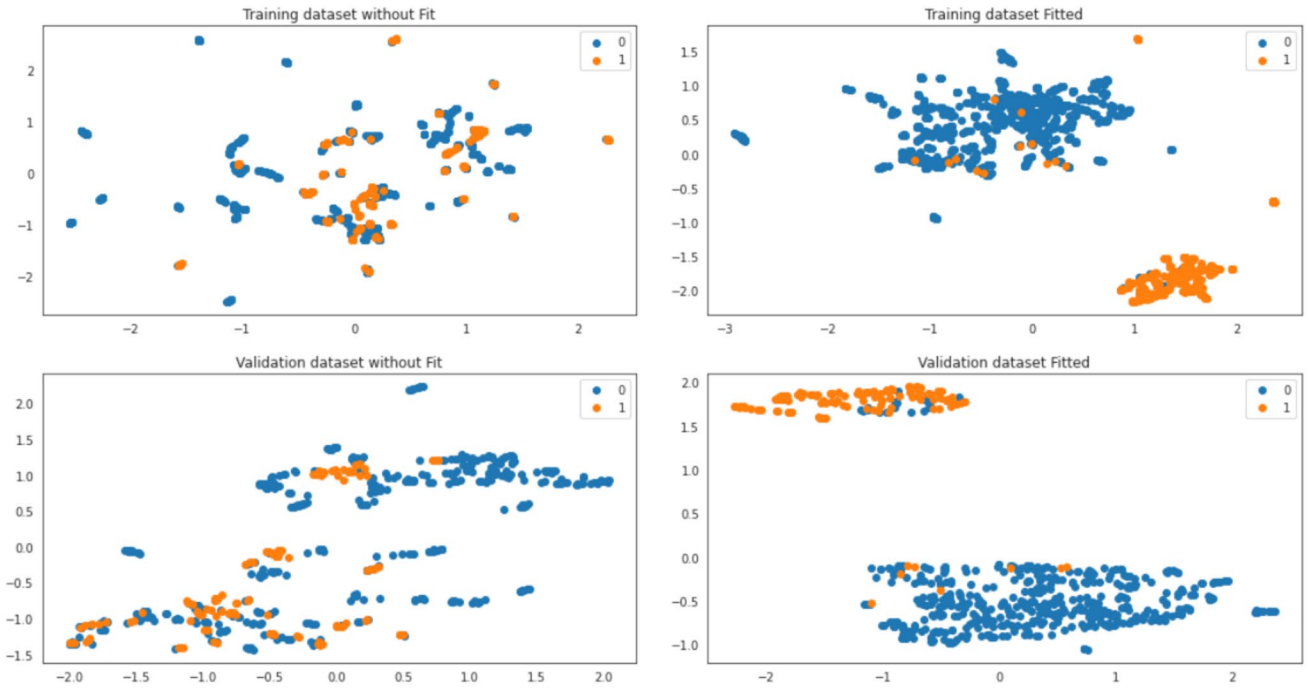

Fig. 11 Embedding representation of texts for two iterations on the `myspace` dataset

For this reason, it is crucial to create models that can capture the nuances of every platform, culture, and language.

The methodology proposed in this paper ensures good performance across different social media platforms and languages. This is mainly due to how the methodology has been implemented and the tools used. In order to achieve the generalizability and flexibility of the proposal presented in this research, we have taken into account several aspects:

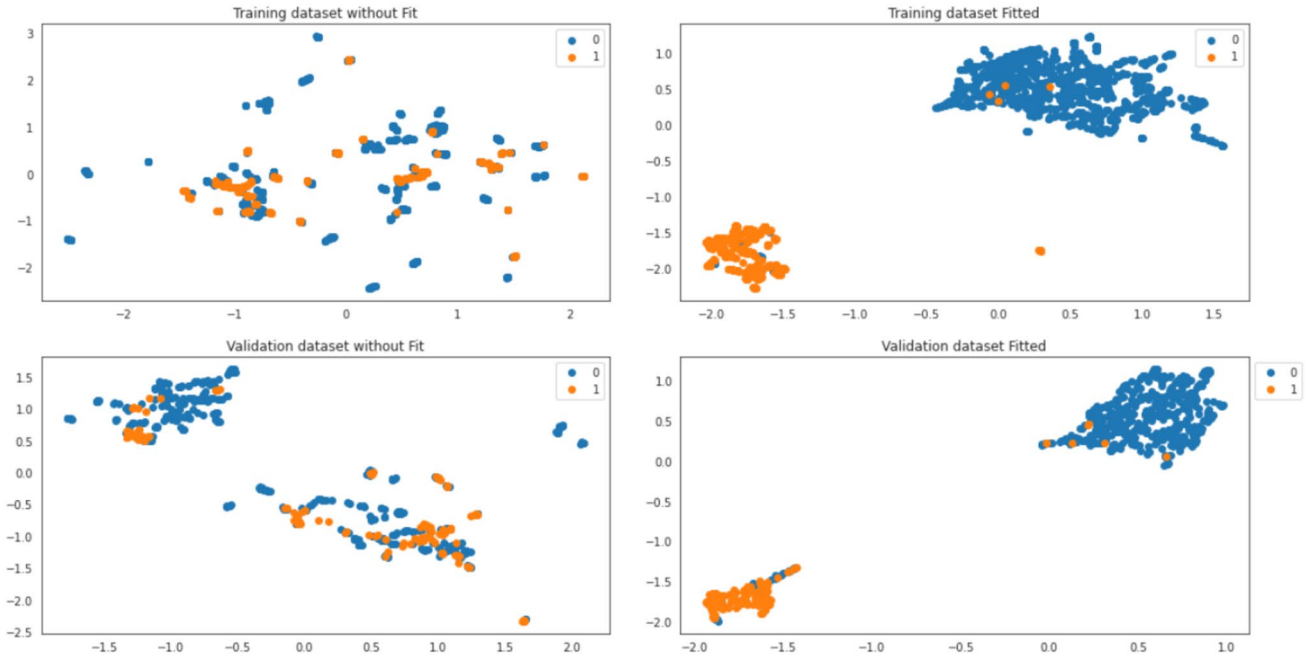

Fig. 12 Embedding representation of texts for three iterations on the `myspace` dataset

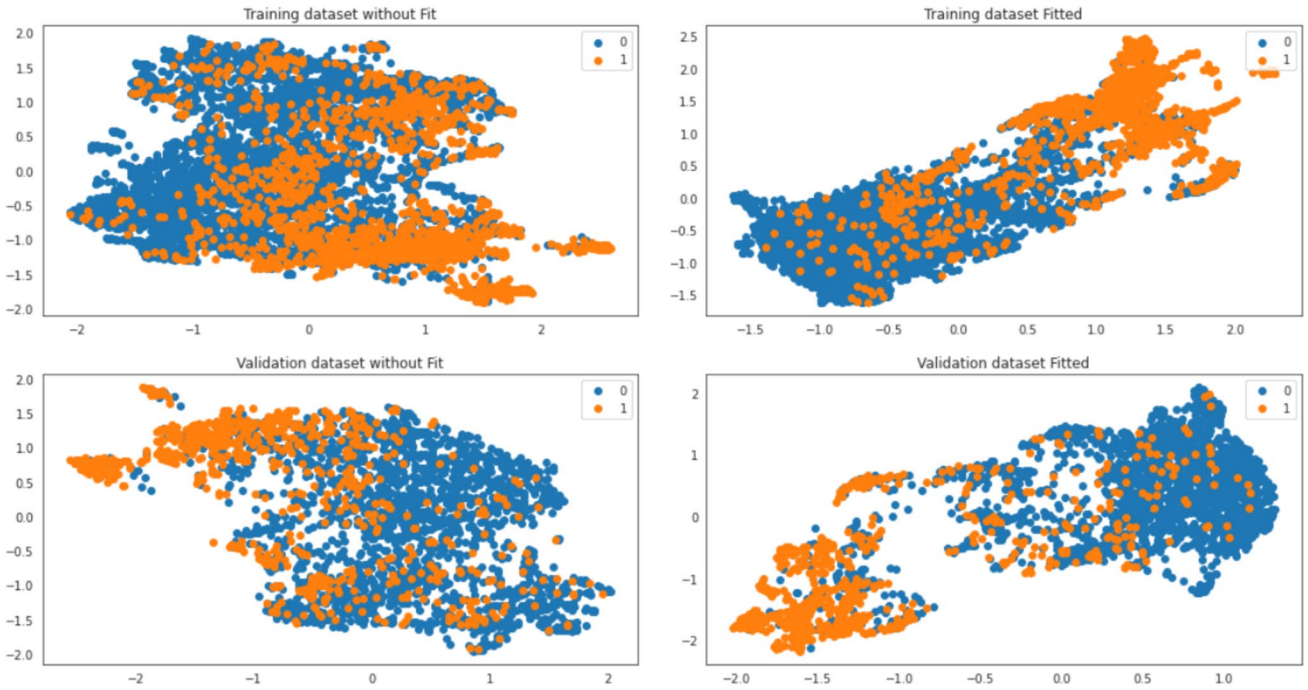

Fig. 13 Embedding representation of texts for one iteration on the `hate-speech` dataset

- *Data from different platforms* As explained in Sect. 4.1, we collected data from two social media platforms to ensure our model learns from diverse sources. This includes platforms known for their distinct cyberbullying dynamics, such as Twitter and MySpace.
- *Multilingual* Despite our proposal not including content from multiple languages, taking into account the steps in the proposed methodology, it is pretty simple to analyse textual data written in any language to capture the linguistic nuances of cyberbullying across

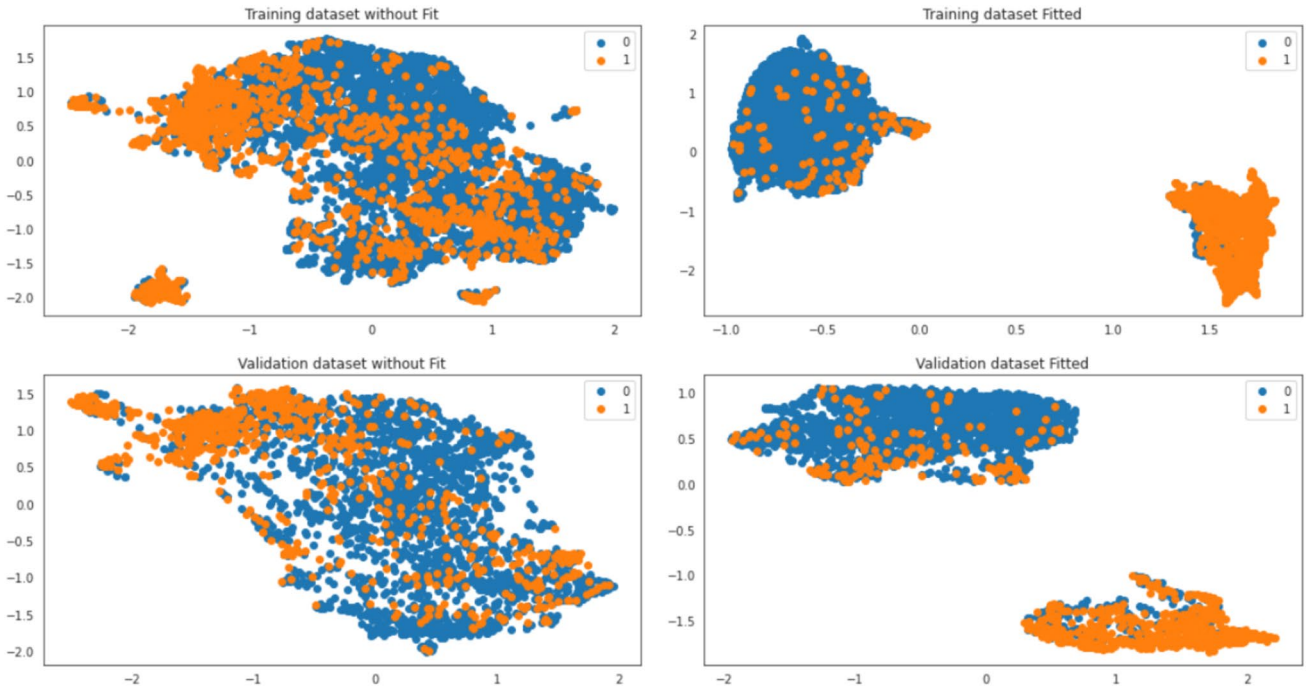

Fig. 14 Embedding representation of texts for two iterations on the `hate-speech` dataset

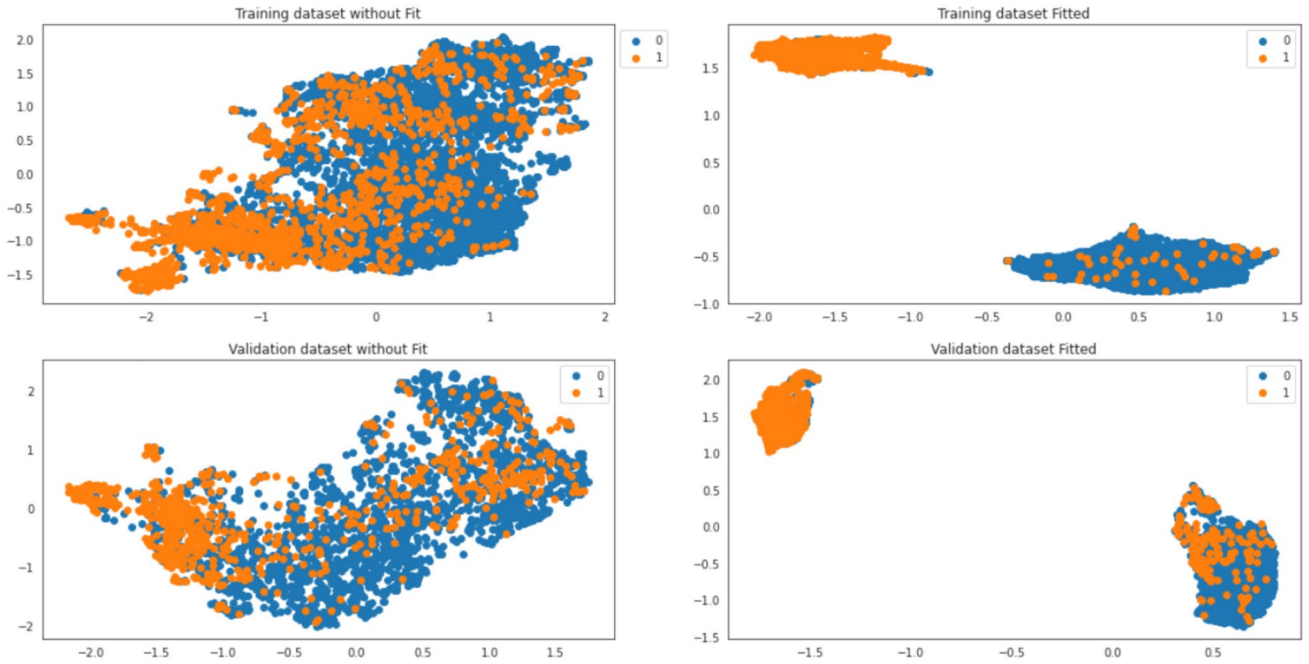

Fig. 15 Embedding representation of texts for three iterations on the `hate-speech` dataset

different cultures and regions. In our case, as posed in Sect. 4.2, we have used the *all-mpnet-base-v2* model, but instead, it is possible to use multilingual models like *paraphrase-multilingual-mpnet-base-v2* in order to support texts of different languages.

- *Fine-tuning process* We employ a fine-tuning technique to adapt our model to different social media platforms and languages. We ensure our model can learn from diverse data effectively by leveraging pre-trained

models such as *all-mpnet-base-v2* and domain adaptation methods.

Our approach can not handle other kinds of data, such as images, videos, and emojis, which are also commonly used for harassers across various social media platforms. Considering the above issue, multimodal models are obtaining more attention. These sorts of models incorporate multimodal information, allowing the capture of cyberbullying cues expressed using different types of data.

5.5 Prevention and intervention

The rise of cyberbullying in online platforms requires proactive measures to prevent and intervene, and the model in this paper could be a valuable tool for such actions.

Traditional methods of combating cyberbullying, such as reactive moderation and user reporting mechanisms, are often insufficient in addressing the scale and complexity of online harassment. Upon detection of potentially harmful content, the model can identify patterns of cyberbullying behaviour as soon as they occur on the online platform, allowing quick intervention. When the model identifies potentially harmful content, it can alert platform moderators automatically. Similarly, the model can implement automatic measures in severe and obvious cases of cyberbullying. Depending on the severity, intervention measures may range from issuing warnings and hiding or removing inappropriate content to suspending user accounts.

However, in addition to reactive measures, automatic detection models can contribute to proactive prevention efforts by educating users about the consequences of cyberbullying. By integrating educational materials, warning messages, or links to support resources for those affected by online harassment, users are empowered to make informed decisions and cultivate a culture of respect and empathy in online communities.

In summary, this model that automatically detects cyberbullying can be a powerful tool for proactively preventing and addressing online harassment, thereby protecting the safety and well-being of users on digital platforms.

5.6 Ethical implications

The early detection of cyberbullying can be of great value, given the current issues. However, concerns can be raised regarding the ethical implications, such as users' privacy, the probability of false positives or the opacity of these systems to users.

6 Conclusions

Although the use of new technologies in education shows a positive impact, the increase in their use can also lead to harmful situations. Cyberbullying is proliferating among children and hurts their well-being. However, it is possible to detect text containing cyberbullying in time through machine learning techniques, which could help to prevent these attitudes in an early stage. We have developed a cyberbullying detection tool and conducted the experiments on three datasets (Table 2). Our results entail an essential improvement over the other two models tested Pericherla and Ilavarasan (2021); Zhao and Mao (2017). The precision of the implemented tool is high, detecting cyberbullying with an accuracy between 83 and 97%. This entails an improvement of around 10% of the previously implemented tools. This study contributes to the early detection of cyberbullying and, therefore, to its prevention. Moreover, it involves critical practical implications for teachers, educational institutes and social institutions. For example, it could help teachers filter those messages in a blog and detect people involved in those situations. This could also provide opportunities to intervene in educational or social terms.

However, further research would be necessary to obtain not only a prediction of cyberbullying but also a more explanatory model that allows end-users to understand how bullying occurs. In future endeavours, we propose exploring and implementing Explainable AI models (XAI) to enhance the explainability of the outcomes.

Acknowledgements This work was supported by the Administration of the Junta de Andalucía and the Spanish Ministry of Universities (NextGeneration EU). It is also supported by the project PPJIA2022-47 of the University of Granada's own research plan. We thank María Dolores Ruiz and David Lagnado for their comments and helpful suggestions.

Author Contributions K.G. designed and performed the experiments, derived the models and analysed the data. C.F. helped develop the algorithm and select the test to validate the results and J.G. helped carry out the validation of the results. K.G. and C.F. wrote the manuscript in collaboration with J.G.

Funding Funding for open access publishing: Universidad de Granada/CBUA.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Agrawal S, Awekar A (2018) Deep learning for detecting cyberbullying across multiple social media platforms. In: European conference on information retrieval, pp 141–153. https://doi.org/10.1007/978-3-319-76941-7_11 . Springer

Ali A, Syed AM (2020) Cyberbullying detection using machine learning. Pak J Eng Technol 3(2):45–50. https://doi.org/10.51846/vol3iss2pp45-50

Balakrishnan V, Khan S, Arabnia HR (2020) Improving cyberbullying detection using twitter users' psychological features and machine learning. Comput Secur 90:101710. https://doi.org/10.1016/j.cose.2019.101710

Bayzick J, Kontostathis A, Edwards L (2011) Detecting the presence of cyberbullying using computer software

Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146. https://doi.org/10.1162/tacl_a_00051

Cassiani-Miranda CA, Campo-Arias A, Caballero-Domínguez CC (2022) Factors associated with cyberbullying victimisation among colombian high-school adolescents. J Child Adolesc Trauma 15(1):27–36. https://doi.org/10.1007/s40653-021-00355-z

Chang C-C, Lin C-J (2011) Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol 2(3):1–27. https://doi.org/10.1145/1961189.1961199

Cole DA, Zelkowitz RL, Nick E, Martin NC, Roeder KM, Sinclair-McBride K, Spinelli T (2016) Longitudinal and incremental relation of cybervictimization to negative self-cognitions and depressive symptoms in young adolescents. J Abnorm Child Psychol 44(7):1321–1332. https://doi.org/10.1007/s10802-015-0123-7

Dadvar M, Eckert K (2020) Cyberbullying detection in social networks using deep learning based models. In: international conference on big data analytics and knowledge discovery, pp 245–255. https://doi.org/10.1007/978-3-030-59065-9_20 . Springer

Devlin J, Chang M-W, Lee K, Toutanova K (2019) 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). Association for computational linguistics, minneapolis, Minnesota

Evangelio C, Rodriguez-Gonzalez P, Fernandez-Rio J, Gonzalez-Villora S (2022) Cyberbullying in elementary and middle school students: a systematic review. Comput Educ 176:104356. https://doi.org/10.1016/j.compedu.2021.104356

Görzig A, Milosevic T, Staksrud E (2017) Cyberbullying victimization in context: the role of social inequalities in countries and regions. J Cross-Cult Psychol 48(8):1198–1215. https://doi.org/10.1177/0022022116686186

Kargutkar SM, Chitre V (2020) A study of cyberbullying detection using machine learning techniques. In: 2020 fourth international conference on computing methodologies and communication (ICCMC), pp 734–739. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-000137

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 30:3146–3154

Kim S, Kimber M, Boyle MH, Georgiades K (2019) Sex differences in the association between cyberbullying victimization and mental health, substance use, and suicidal ideation in adolescents. Canad J Psychiatr 64(2):126–135. https://doi.org/10.1177/0706743718777397

Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. J Am Stat Assoc 47(260):583–621

Langos C (2012) Cyberbullying: the challenge to define. Cyberpsychol Behav Soc Netw 15(6):285–289. https://doi.org/10.1089/cyber.2011.0588

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach. ArXiv abs/1907.11692https://doi.org/10.48550/arXiv.1907.11692

Liu C, Liu Z, Yuan G (2020) The longitudinal influence of cyberbullying victimization on depression and posttraumatic stress symptoms: the mediation role of rumination. Archiv Psychiatr Nurs 34(4):206–210. https://doi.org/10.1016/j.apnu.2020.05.002

Lobe B, Velicu A, Staksrud E, Chaudron S, Di Gioia R (2021) How children (10-18) experienced online risks during the covid-19 lockdown-spring 2020. Key findings from surveying families in 11 European countries https://doi.org/10.2760/562534

McCullagh P, Nelder JA (1989) Generalized linear models. Springer, Boston

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: Bengio, Y., LeCun, Y. (eds.) 1st international conference on learning representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop track proceedings. https://doi.org/10.48550/arXiv.1301.3781

Muneer A, Fati SM (2020) A comparative analysis of machine learning techniques for cyberbullying detection on twitter. Future Internet 12(11):187. https://doi.org/10.3390/fi12110187

Paul S, Saha S (2020) Cyberbert: Bert for cyberbullying identification. Multimed Syst 28:1897–1904

Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543

Pericherla S, Ilavarasan E (2021) Transformer network-based word embeddings approach for autonomous cyberbullying detection. Int J Intell Unmanned Syst. https://doi.org/10.1108/IJIUS-02-2021-0011

Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, Volume 1 (long papers), pp 2227–2237. Association for computational linguistics, New Orleans, Louisiana

Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training

Ranney ML, Patena JV, Nugent N, Spirito A, Boyer E, Zatzick D, Cunningham R (2016) Ptsd, cyberbullying and peer violence: prevalence and correlates among adolescent emergency department patients. Gen Hosp Psychiatr 39:32–38. https://doi.org/10.1016/j.genhosppsych.2015.12.002

Ratnayaka G, Atapattu T, Herath M, Zhang G, Falkner K (2020) Enhancing the identification of cyberbullying through participant roles. CoRR abs/2010.06640https://doi.org/10.48550/arXiv.2010.06640

Reimers N, Gurevych I (2019) Sentence-bert: sentence embeddings using siamese bert-networks. In: proceedings of the 2019

conference on empirical methods in natural language processing. Association for computational linguistics. https://doi.org/10.48550/arXiv.1908.10084

Rousseeuw P (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20(1):53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Salawu S, He Y, Lumsden J (2017) Approaches to automated detection of cyberbullying: a survey. IEEE Trans Affect Comput 11(1):3–24. https://doi.org/10.1109/TAFFC.2017.2761757

Sui J (2015) Understanding and fighting bullying with machine learning. PhD thesis, The University of Wisconsin-Madison

Wang W, Xie X, Wang X, Lei L, Hu Q, Jiang S (2019) Cyberbullying and depression among chinese college students: a moderated mediation model of social anxiety and neuroticism. J Affect Disord 256:54–61. https://doi.org/10.1016/j.jad.2019.05.061

Waseem Z, Hovy D (2016) Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In: proceedings of the NAACL student research workshop, pp 88–93. Association for computational linguistics, San Diego, California

Wilcoxon F (1945) Individual comparisons by ranking methods. Biom Bull 1(6):80–83. https://doi.org/10.1007/978-1-4612-4380-9_16

Zhao R, Mao K (2017) Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. IEEE Trans Affect Comput 8(3):328–339. https://doi.org/10.1109/TAFFC.2016.2531682