STM JOURNALS

Recent Trends in
Programming Languages

ISSN: 2455-1821
Volume 11, Issue 2, 2024
May–August
DOI (Journal): 10.37591/RTPL

https://journals.stmjournals.com/rtpl

Review RTPL

Enhancing Profanity Detection in Dravidian Languages: Leveraging Language Models for Optimization and Improvement

Jyoti Jayesh Chavhan*

Abstract

Detecting and documenting instances of abusive behaviour can significantly improve the quality of virtual environments. Given the vast amount of content published daily on social media, it is impractical for human annotators to manually identify potentially harmful content. Recent algorithmic initiatives, especially on platforms like Twitter, have advanced in abuse detection. However, for Dravidian texts, there remains a need to understand the context better and build robust language models for classification. In our study, we employed the XLM-Roberta language model alongside various optimizers to train our model, achieving state-of-the-art results. The exceptional outcomes can be attributed to the meticulous integration of these optimizers and activation functions into pre-existing language models. Our proposed technique significantly enhances overall accuracy by 19% across multiple domains. This improvement stems from incorporating advanced optimization methods into the language models. Our model demonstrated remarkable accuracy across different Dravidian languages: 74.13% for Kannada, 96.25% for Malayalam, and 79.72% for Tamil. These results highlight the efficacy of our approach in enhancing the detection of abusive content in these languages. The combination of advanced language models and tailored optimizers/activation functions has led to substantial performance gains, setting a new benchmark in the field of abuse detection for Dravidian texts.

Keywords: Detection of hate speech, Dravidian code-mixed data, language models, deep learning, natural language processing

INTRODUCTION

NLP's grasp of online abbreviations, slang, code-switching, emoticons and emojis, and hashtags are unique for social media listening. With NLP, you may collect data in any client-required language and then prepare it for consumption by an ML model. Sentiment analysis provides further information into your brand's success based on the positive, negative, and neutral emotions detected in social media discussions. And in doing so, it offers you practical, helpful information [1]. As part of your marketing strategy, you may change your advertising campaign, build your brand's reputation, enhance the characteristics of your product, or reach out to influencers based on the client sentiment found through social media monitoring. Numerous sectors, including business and marketing, politics, health care, and social policy, utilise sentiment analysis. Many applications of sentiment analysis can improve decision-making in a vast array of sectors [2]. Using sentiment analysis, global events, such as a disaster, activity, sport, or event, can be analysed. Studies juxtaposing how western and eastern nations perceive ISIS are examples. As evidenced by the data, ISIS is considered a global

***Author for Correspondence**
Jyoti Jayesh Chavhan
E-mail: jyotichavhan293@gmail.com

Assistant Professor, Department of Computer Science and Engineering, South Indian Educational Society Graduate School of Technology, Nerul, Navi Mumbai, Maharashtra, India

Received Date: June 27, 2024
Accepted Date: July 29, 2024
Published Date: August 07, 2024

Citation: Jyoti Jayesh Chavhan. Enhancing Profanity Detection in Dravidian Languages: Leveraging Language Models for Optimization and Improvement. Recent Trends in Programming Languages. 2024; 11(2): 17–23p.

terrorist from a variety of views. In addition, sentiment analysis heightens public awareness of data security and the risk of security breaches.

There are further venues where information can be found, although not everyone there is bilingual. Hence, internet translations are typically beneficial, especially for researchers. Moreover, we would not be able to watch the countless foreign films and documentaries with subtitles that are available on our video streaming channels without the NLP technology's capacity to translate audio to text fast and effectively at scale [3]. Morphology, anthropological linguistics, philology, syntax, and phonology of languages excite linguists because they are so fascinating, unique, and complex. The continuous acquisition of new knowledge by data scientists allows them to develop AI/ML models that can comprehend language [4]. At the beginning of the 20th century, the first software for machine translation was made. The main goal of MT research was to find a method that could substitute sentient translation because of the problems with clearly intentional interpretation and the high costs of translation. The primary goal of language processing is to create and improve automatic translation from one language to another. The most common way to do machine translation is with a corpus-based approach, which matches words or chunks of text from the source language with instances from other languages that have been collected in a parallel corpus or database [5]. Using a statistical method to choose translations cuts down on the number of variables and makes the translation process more accurate. Google Translate (GT), which was released in 2006, has become the most widely known machine translation programme because it uses a large database and translates more accurately than other programmes. These models are used to identify offensive language. Although the purpose of the MLM is comparable to that of the Bert model, which came before it, continuous text streams have been substituted for phrase pairs in this model. The objective of TLM is to produce parallel sentences, but the scope of MLM expands to include sentence pairings. The model can focus on the English text in its original form as well as the English text with its French translation at the same time to identify a hidden term written in English. In addition to this, it is suggested that the English and French representations be brought into alignment with one another.

Due to the fact that the bulk of traditional translation work is completed manually, the accuracy and quality of the final product are always guaranteed. In the case of frequent international interactions, the efficacy and cost of human translation fall far short of the required requirements [6]. Due to the rapid expansion of the Internet and the vast processing power of computers, machine translation has advanced significantly in terms of speed and cost, but at the sacrifice of translation quality. To complete the merger of statistical machine translation vocabulary knowledge, this component builds a recommendation module for machine translation and statistical word recognition. The framework's foundation is the neural machine translation “Encoder-Decoder” system [7]. The statistical machine translation vocabulary recommendation module analyses target language and attention indicators to give historical data for word suggestions.

Websites provide users with a conversational environment, with internet users on hand to maintain a courteous tone and encourage meaningful dialogue. The panellists evaluate compliance with the platform's conversational norms, such as the restriction of offensive language. They adhere to these standards by eliminating half or the entire message of a participant [8]. Whoever utilises social media, message boards, or other online forums risks being taunted or even harassed. People often say online that they should all go to pieces for much of what they did or something similar. These can ruin the experience for users and make the community less nice. In addition to software that employs regular expressions and blacklists to detect offensive language and delete communications, numerous online organisations have rules and laws that users must adhere to in order to avoid the use of abusive language.

As a result of this harassment, Twitter modified its policies addressing hate speech. Although automatically detecting abusive language online is a significant topic and endeavour, earlier work has been incoherent, impeding progress. Abusive language is probably very grammatical and flexible.

Although there are occasional instances of offensive language being used online, such as “Add another Jew fined a billion dollars for stealing like a cockroach”, the use of such language is uncommon. Put them all on wait. Using this automatic process indicator can be beneficial. The usage of social networking sites for a variety of purposes, including product advertising, news dissemination, and achievement celebration, has increased significantly. But it is also used to verbally abuse, threaten, and degrade particular racial or ethnic groups [9]. It is crucial to discover and remove harmful posts from social media platforms as quickly as possible, as they frequently have detrimental effects on people and spread rapidly. There has been a rise in demand for controversial and unpopular content in recent years. Code-mixed language is one of the various obstacles associated with identifying violent speech [10].

One of the recurring problems with online social media platforms is the use of offensive comments or statements that may be directed at a particular individual or group [11]. In recent years, the detection and recognition of such detrimental comments and posts has developed into a crucial subfield of natural language processing. Transformer models have already been used to teach Tamil, Kannada, and Malayalam speakers how to recognise abusive language. We restrict text preparation techniques such as lemmatization, stemming, and phrase-based model removal to retain the context of the user's intent [12]. Although transformer representations are contextual models (e.g., BERT, XLM-RoBERTa, etc.), it can be observed that stop words and non-stop words receive nearly the same amount of attention when they are deployed. In settings involving many languages, code-mixing is common, and the resulting writings may utilise scripts other than those of the local language.

Users can upload information casually to social networking and product review websites [10]. In addition, to increase customer satisfaction, these platforms ensure that users may express themselves naturally, for as by speaking their native tongue or switching between several or perhaps more varieties during a single discussion. Nonetheless, as the majority of natural language writings are based on formal languages with rigid syntax, analysing “user-generated” comments creates difficulties. While user-generated information in low-resource environments is frequently combined with English or other somewhat elevated languages, the majority of advances in emotion classification and abusive phrase identification algorithms are based on monochromatic data for high-resource languages.

CURRRENT SYSTEMS

Several experiments utilizing various deep learning and transfer learning approaches were conducted in order to identify problematic YouTube videos [13]. Digital networks have developed into an indispensable component of our daily activities. Thus, its posts have a bigger social impact. Depression, anxiety, and insomnia may result. The nature of provocative language varies based on the target audience. Identifying the nature of offensiveness necessitates multiple classifications. A collection of Tamil-English data was compiled as part of a competition to discover problematic Dravidian languages. The information was designated as: (1) Not offensive; (2) A personal insult to the perpetrator; (3) Specific harassment and insults; (4) Injure others by insulting them; and (5) Offensive, but not malicious This is a classification challenge requiring value predictions for six distinct classes. The objective is to categorize text data according to offensive methods, such as Individual-Targeted-Offense, Group-Targeted-Offense, Other-Targeted-Offense, Offensive-Untargeted, and Not-Tamil. Moreover, the classification includes a non-offensive category. A multilingual framework is used to classify the textual data. This framework has been trained on more than 100 languages, allowing it to classify text data in many languages with precision. The performance of the classification is measured using three metrics: precision, recall, and the class not-offensive harmonic-score. The model with the maximum weighted average is deemed superior. The classification task's findings are reported in the text. The average weighted F1-scores on the test datasets were 0.7346 and 0.7444, which are relatively high values. For the same dataset, additional CNN models scored lower F1-scores of 0.6112 and 0.6111.

A general summary of a categorization problem and the solution process: It illustrates the significance of specificity, retention, and F1-score metrics in evaluating classification performance and exhibits the

Table 1. Comparison of various optimizers with language models.

	Model with optimizers	Accuracy
Kannda	xlm-roberta-base (Adadelta)	73.75%
	xlm-roberta-base (Adagrad)	74.13%
	xlm-roberta-base (Adam)	73.49%
	xlm-roberta-base (AdamW)	72.97%
	xlm-roberta-base (Adamax)	73.87%
Malayalam	xlm-roberta-base (Adadelta)	95.45%
	xlm-roberta-base (Adagrad)	96.20%
	xlm-roberta-base (Adam)	95.75%
	xlm-roberta-base (AdamW)	96.25%
	xlm-roberta-base (Adamax)	96.25%
Tamil	xlm-roberta-base (Adadelta)	78.78%
	xlm-roberta-base (Adagrad)	78.92%
	xlm-roberta-base (Adam)	79.38%
	xlm-roberta-base (AdamW)	78.71%
	xlm-roberta-base (Adamax)	79.72%

value of a multilingual framework for categorizing text data in different languages. The categorization task results demonstrate the efficacy of the employed strategy and the potential for additional enhancements.

They have employed a pre-trained machine to identify undesirable tanglish phrases [11]. Nearly all age groups utilize social media for both professional and recreational objectives in the Internet age. Yet, some incorrect language or speech, which may be expressed in real or online formats, may disappoint a group of people. Due to the small dataset and the varying spelling of tanglish depending on the user, it is difficult to automatically recognize unpleasant remarks in the Tamil language. The dataset is extracted from YouTube and other social media platforms with mixed data coded. The different sample sizes generate a plausible scenario. Preprocess data by removing punctuation, emoji, and other characters. Training was conducted with the training and development package. In order to recognize offensive information, a pre-trained transformer model with a customized output was used. The last few lines of the learned data were utilized for prediction before being moved to a linear layer. The data were then resampled to improve accuracy. The score in the examination was 0.61, whereas the training score was 0.91. In Table 1, results are described in detail. As seen, pre-trained models with mean pooler provided the greatest outcomes. Due to model overfitting, there is a difference between test and train scores.

Researchers have attempted to address a new issue that has arisen on social networking sites as a result of inappropriate content posted by an individual or group [8]. Hence, the reader may face psychological troubles or mental anguish. In countless instances, it has led to riots. Therefore, it is vital to filter offensive content effectively. The data set for the first challenge was taken from YouTube. It was in Malayalam exclusively. Twitter provided the data set for the second task, which includes Malayalam and Tamil text. Around 83% of the data set is labelled NOT. This custom model was pre-trained in 100 languages, including Tamil and Malayalam. For training, the Adam optimizer was utilized. The final linear output layer was used for classification. With a F1 score of 0.94 on task 1 Malayalam, 0.84 on task 2 Tamil, and 0.76 on task 2 Malayalam, it was decided that the findings were superior. The performance of the system diminishes considerably over time. This will be examined in an upcoming study.

They have classified message level texts into offensive and non-offensive categories. Important social media services are now required [13]. Textual sentiment information may convey the emotions of the sender. In countries where multiple languages are spoken, English is widely used for written communication. In this informal style, the spellings of many words are uncommon and, in most cases,

distinct. Hence, it becomes difficult to recognize offensive content with precision. The organizer-supplied data set was utilized for all three competitions. The data collection contained all three varieties of code-mixed texts. Character and sub word embeddings were provided. Function 1: Malayalam to English, function 2: Tamil to English, and function 3: Malayalam to English. Future works are anticipated to feature a word-by-word focus and a hierarchical framework.

Using machine learning algorithms, harmful content has been eliminated from social media. The social media have become the dominant form of communication [12]. Yet, some despicable individuals post harsh or damaging comments that incite enmity within a group of folks. Due to the vast volume of data uploaded every minute, we cannot manually eliminate offensive posts. Thus, we must create a machine learning model that automatically reads and classifies objectionable comments. The entire dataset is collected from YouTube, where code-mixing between Dravidian and Roman is prevalent. The dataset consisted of straightforward text lines. It was then broken into five categories. Each of the 35139 datasets was utilized. Initially, a number of machine learning models were analysed to lay the groundwork for the research. In addition, it was found that simple machine learning models are insufficiently effective. Standard ML models do not interpret a word in relation to its neighbours. Hence, networks that are neutral and capable of recognizing even nearby sentences are utilized. Combining varied models typically produces superior results. The team achieved different ranking for different languages. The dataset was sorted into five categories depending on offensiveness.

It is crucial, when filtering consumer products, to detect incorrect acceptable vocabulary platforms in local languages [14]. As a result, other languages like as Tamil, Malayalam, and Kannada struggle to recognize erroneous language. Due to the fact that market information for some well languages is often script and understudied, it is critical to provide facilities and understanding consumers research in order to further study in these languages. Scholars have proposed a coordinated effort to identify superfluous terms in Dravidian languages. The first joint challenge on difficult language identification in Tamil, Malayalam, and Kannada was released, and it was based on data that was accurately labelled and satisfactory. Owing to the employment structure, the model is evaluable in bilingual environments and contains code mixing. Several teams, in contrast, are proficient in all three languages. In this instance, the embedding type had a substantial impact on the findings.

PROPOSED SYSTEM

Data Set

Because of the millions of people that contribute to online media networks such as Twitter, YouTube, and Vimeo, the data saved on these sites is always changing. These platforms include Twitter, Facebook, and YouTube, among others. Twitter, Facebook, and YouTube are among websites that come into this category. It is feasible that this will have a substantial impact on both an individual's and an organization's reputation. This tendency has significantly raised the need for doing sentiment analysis and removing inaccurate terms from web [13].

It is probable that this is due to the site's extensive content, which contains not only songs but also instructions, product reviews, and movie trailers. YouTube users can broadcast their own creations to the platform's community and solicit feedback from other service users. It makes it possible for users to contribute more content in languages with limited resources by allowing them to do so. As a direct result of this, we concluded that we would utilize YouTube to collect user comments for our dataset. Given how ubiquitous movies are also some of Tamil, Malayalam, and Kannada speakers, we thought it would be wise to gather data on movie trailers. The availability of movie trailers in all three of these languages was a major reason in our decision to concentrate on this topic. Designed with a distinct number of trains, validation, and test cases for each language the system is possibly capable of comprehending. The problem that must be solved is known as a multi-class classification problem, and it asks the model to classify a possible piece of writing into one of six predefined categories [10]. The issue has been called a multi-class classification challenge.

Architecture

Because it represents the system's key components as blocks, the diagram of the system overview is able to give a detailed explanation of the system and its ability to detect vulgarity. At first, both the original and Romanized versions of the textual data are used. This is something that occurs throughout the entire process. After this is complete, a pre-training layer is built with the use of mask language modelling. A foundational education will be put to use in many ways, including enhancing students' ability to understand Dravidian literature. The classification model will be trained using a combination of the XLM-Robetra model and this pre-trained model. In this case, the term “sentence” refers to the method in question. Word tokenization is accomplished through “piece tokenization” in XLM-Robetra. For this purpose, we employ a Long Short-Term Memory (LSTM) that can store information in both directions to classify the text into useful chunks. Transliterating the decoded texts allows for an assessment of the results. This section describes the framework behind the Dravidian offensive language detection language models as shown in Figure 1. While developing a standard for cross-lingual classification, we tune XLMs in the same way that single-language models are adjusted for English classification.

RESULT

Several different sets of optimizers were examined and evaluated, but it was concluded that the Adagrad optimizer provided the best results when training the XLM-Roberta Model for the Kannada language. In spite of this, it was shown that the AdamW optimizer produced the best results when employed with this language. Utilizing the classification models, we were able to achieve the requisite 96.30% precision.

After conducting extensive training on the XLM-Roberta Model using a variety of different optimizers, it was determined that Adamax is the optimal collection of parameters for processing the Tamil language. After much deliberation, this conclusion was chosen. Using the language model as a basis, an accuracy score of 79.72% was not out of the question. Table 1 summarizes the findings of an examination into the similarities and differences between a large number of different optimizers.

CONCLUSION

After the data have been compiled, we are able to make the observation that there is a very large number of different combinations of optimizers that are available, and that some of these combinations have obtained results that are so forward-thinking that they are regarded as being state-of-the-art. We employed the XLM-Roberta model in conjunction with several other optimizers while we were training the model so that we might increase the model's overall performance. As a result of this, our efforts to train the model were fruitful and we achieved the desired results.

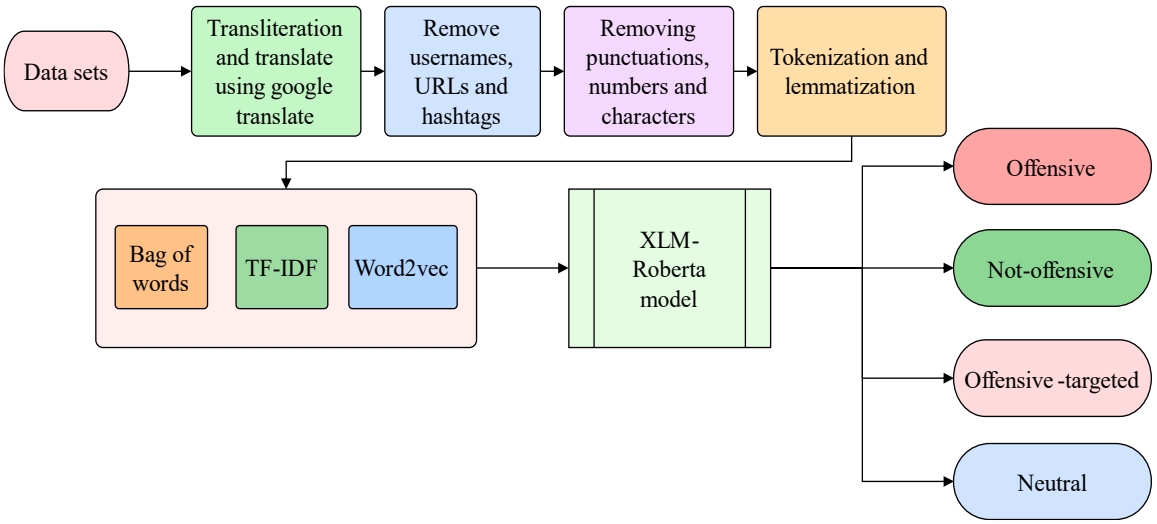

Figure 1. Architecture for offensive text detection.

REFERENCES

1. Alqarni M, Azim A. Low Level Source Code Vulnerability Detection Using Advanced BERT Language Model. In 35th Canadian AI Conf. 2022 May 27.
2. Risch J, Ruff R, Krestel R. Explaining offensive language detection. Journal for Language Technology and Computational Linguistics. 2020 Jul 1; 34(1): 29–47.
3. Husain F, Uzuner O. A survey of offensive language detection for the Arabic language. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP). 2021 Mar 9; 20(1): 1–44.
4. Djandji M, Baly F, Antoun W, Hajj H. Multi-task learning using AraBert for offensive language detection. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection. 2020 May; 97–101.
5. Andrew JJ. JudithJeyafreedaAndrew@ DravidianLangTech-EACL2021: offensive language detection for Dravidian code-mixed YouTube comments. In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages. 2021 Apr; 169–174.
6. Wiedemann G, Ruppert E, Jindal R, Biemann C. Transfer learning from lda to bilstm-cnn for offensive language detection in twitter. arXiv preprint arXiv:1811.02906. 2018 Nov 7.
7. Roy PK, Bhawal S, Subalalitha CN. Hate speech and offensive language detection in Dravidian languages using deep ensemble framework. Comput Speech Lang. 2022 Sep 1; 75: 101386.
8. Garain A, Mandal A, Naskar SK. JUNLP@ DravidianLangTech-EACL2021: Offensive language identification in Dravidian languages. In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages. 2021 Apr; 319–322.
9. Bharathi B. SSNCSE_NLP@ DravidianLangTech-EACL2021: Offensive language identification on multilingual code mixing text. In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages. 2021 Apr; 313–318.
10. Subramanian M, Ponnusamy R, Benhur S, Shanmugavadivel K, Ganesan A, Ravi D, Shanmugasundaram GK, Priyadharshini R, Chakravarthi BR. Offensive language detection in Tamil YouTube comments by adapters and cross-domain knowledge transfer. Comput Speech Lang. 2022 Nov 1; 76: 101404.
11. Kim Y, Dyer C, Rush AM. Compound probabilistic context-free grammars for grammar induction. arXiv preprint arXiv:1906.10225. 2019 Jun 24.
12. Clarke CL, Cormack GV, Lynam TR. Exploiting redundancy in question answering. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. 2001 Sep 1; 358–365.
13. Schwitter R, Mollá D, Fournier R, Hess M. Answer extraction towards better evaluations of NLP systems. In Proceedings of the 2000 ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding systems. 2000 May 4; 6: 20–27.
14. Irie K, Tüske Z, Alkhouli T, Schlüter R, Ney H. LSTM, GRU, highway and a bit of attention: An empirical overview for language modeling in speech recognition. In Interspeech. 2016 Sep 8; 3519–3523.

