RBIPA: AN ALGORITHM FOR ITERATIVE STEMMING OF TAMIL LANGUAGE TEXTS

V. Indumathi, S. SanthanaMegala

School of Computer Studies, Rathnavel Subramaniam College of Arts and Science, Coimbatore, India

ABSTRACT

Cyberbullying is currently one of the most important research fields. The majority of researchers have contributed to research on bully text identification in English texts or comments, due to the scarcity of data; analyzing Tamil textstemming is frequently a tedious job. Tamil is a morphologically diverse and agglutinative language. The creation of a Tamil stemmer is not an easy undertaking. After examining the major difficulties encountered, proposed the rule-based iterative preprocessing algorithm (RBIPA). In this attempt, Tamil morphemes and lemmas were extracted using the suffix stripping technique and a supervised machine learning algorithm for classify the word based for pronouns and proper nouns. The novelty of proposed system is developing a preprocessing algorithm for iterative stemming; lemmatize process to discovering exact words from the Tamil Language comments. RBIPA shows 84.96% of accuracy in the given Test Dataset which hasa total of 13000 words.

KEYWORDS

Rule-Based Preprocessing, Cyberbullying, NLP, Tamil Stemmer, Lemmatization, Machine Learning

1. INTRODUCTION

Online harassment known as "cyberbullying" takes place on social networking sites. These networks are used by criminals to gather data and information that will enable them to commit crimes, such as locating a susceptible victim. Researchers have therefore been focusing on establishing strategies and tools to identify and stop cyberbullying. In order to effectively identify cyberbullying situations, recent research has concentrated on cyberbullying monitoring systems. The main impression behind such systems is to extract some features from social media texts and then build classifier algorithms based on those extracted features to detect cyberbullying [1].

Bully is commonly defined as any communication that makes an individual or a group feels offended, resentful, annoyed, or insulted because of characteristics such as color, ethnicity, sexual orientation, nationality, race, or religion. When compared to physical abuse, these types of conversations last forever on social media platforms such as Facebook, Twitter, and YouTube, and affect the individual's mental state, causing depression, insomnia, and even suicide. As a result, identifying and restraining hate speech is critical. Finding such content is a crucial problem, though, because of the enormous amount of user-generated multilingual data on the internet, particularly social media platforms [2]. Many people from different demographics and linguistic backgrounds use social media sites to exchange information and interact with others. Furthermore, during the conversation, these speakers frequently combine their mother tongue with a second language. To deal with abusive content posted by users, many social media networks currently use a manual content screening method [3].

DOI: 10.5121/ijnlc.2023.12303

Social Mediasare ideal for detecting bullies. Bully Text classification can be performed on social media data in the form of positive, negative, or neutral. Handling posts in languages other than English, such as Tamil, has not progressed significantly. Extending detection into Tamil comments is vital for improving performance and enhancement while retaining all user comments. The best pre-processing techniques increased the efficiency of this analysis. Tamil morphology, like that of the other Dravidian languages, is very rich and agglutinative. In a language with such a rich morphological structure, deeper research at the word level is necessary to extract the meaning of a word from its morphemes and categories [4].Cyberbullying detection in Tamil is still a hot topic in the research community. In this context, analyzing Tamil comments is not a straightforward process. The first step in this sequence is to translate non-English comments into English, and then the complementary steps are applied for classification. Data sparsity is defined as spelling errors, emphasized words, and contraction words. NLP can be used to reduce data sparsity during pre-processing. Through Tamil Bully text classification, the accuracy of the classification can be greatly improved.

Effective preprocessing techniques are needed for the best analysis of Tamil language comments. Preprocessing techniques need to be framed with a unique set of procedures on stop word removal, compound forms, stemming, and lemmatization. Creating a stemmer for any language necessitates research into the derived forms that a word in that language can take. Tamil is a language with many inflections. In Tamil, words are made up of a root and one or more affixes. Affixes can be either prefixes or suffixes. The majority of Tamil affixes are suffixes. A word may receive an unlimited number of suffixes [5].

1.1. Motivation

Tamil Comments processing still lags behind English Comments processing in many ways. The effectiveness of bully analysis was improved using the best pre-processing methods for Tamil Comments In such cases, a rule should be developed to improve the stemming procedures. In this manner, a Rule-Based Iterative Preprocessing (RBIP) algorithm has been proposed, and it handles the word with a language identifier before performing the stemming procedures and stem and lemmatize the words to identify the proper word.

1.2. Contribution

- The main contribution of the work is to identify bullying content in Tamil language.
- Proposed a rule based iterative pre-processing stemming algorithm for finding the root word of tamil comments.
- Accuracy and the rate of under- and over-stemming were evaluated in order to evaluate the performance of the proposed model.

The paper is organized as follows: In Section 2, the Related Works are presented. In Section 3, the Methodology used in the study is given. In Section 4, the Result and analysis are presented. In Section 5, the paper is concluded.

2. RELATED WORKS

According to [6], the majority online community members in non-native English-speaking nations use code-mixed text in their posts and comments. Due to the numerous difficulties in identifying offensive content and an absence of resources for Tamil, the process is made even more difficult.It showed the data recovery preparation techniques and their logical calculations.

In [7] claims that they have used a variety of feature extraction methods for NLP, including tokenization, lemmatization, and vectorization. They even analyzed that, in contrast to word2vec and bag of words, count vectorizer and TF-IDF are the two approaches providing very good accuracy in feature extraction. Therefore, we conducted a comparative analysis between these two feature extraction models to choose the best one. We found that count Vectorizer marginally outperforms TF-IDF in terms of accuracy.

According to [8], while there has been a lot of work done on stemming in languages such as English, Nepali stemming has very few works. The goal of this research is to develop a rule-based stemmer for Nepali text. It is an affix stripping method that separates the two different types of suffixes used in Nepali grammar. Only one negativity prefix is identified and removed. Research focuses on a variety of techniques for improving stemming performance, such as exception word identification, morphological normalization, and word transformation. On a simple TF-IDF based IR system and a simple news topic classifier utilising the Multinomial Naive Bayes Classifier, the stemmer is tested both intrinsically and extrinsically. The difference in performance between these systems with and without the stemmer is investigated.

According to [9], proposes a session-based framework for the automatic detection of cyberbullying from massive amounts of unlabeled streaming text. Given the large volume of streaming data from Social Networks that arrives at the server system, we include an ensemble of one-class classifiers in the session-based framework. The proposed framework addresses the real-world scenario in which only a limited number of positive examples are available for preliminary training. Their contribution here, is to detect cyberbullying automatically in real-world situations where labeled data is not readily available. Their primary findings indicate that the proposed method is level headedly effective for automatically detecting cyberbullying on Social Networks. The trials show that when knowledge from helpful and unlabeled data, the ensemble learner outperforms the single-window and fixed-window approaches.

According to [10], a rule-based stemmer for Urdu is proposed. Urdu is a synthesis of several languages, including Arabic, Hindi, English, Turkish, and Sanskrit. It has a complicated and varied morphology. This explains why linguistic processing of Urdu has not advanced very significantly. The process of changing a word into its root form is known as stemming. We separate the suffix and prefixes from the word by stemming. Word processing, spell checking, word parsing, word frequency and count research, search engines, and natural language processing are just a few of the applications it can be used in. A rule-based stemmer can distinguish between the affixes and identify the root word. They tested their system and got 86.5% accuracy.

3. RESEARCH METHODOLOGY

The paper proposed a model for cyberbullying identification on social media sites. Particularly the model (as shown in fig 1) provides an effective preprocessing algorithm to handle the Tamil Comments.

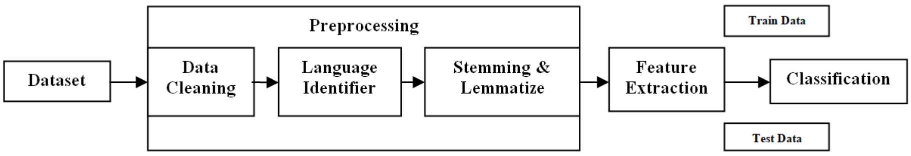
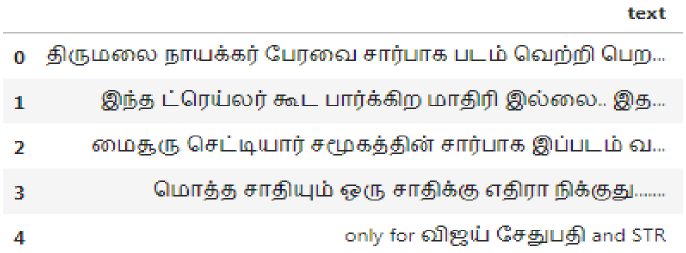


Figure 1: Proposed Preprocessing Model

3.1.Dataset

Dataset of Talk page edits of Wikipedia comments used for this model. In the dataset, 80% of the data was streamed for Training and 20% was used for Testing. Training data contains 52,000 observations and the test data contains 13,000 records respectively. Sample entries from the dataset given in fig. 2.



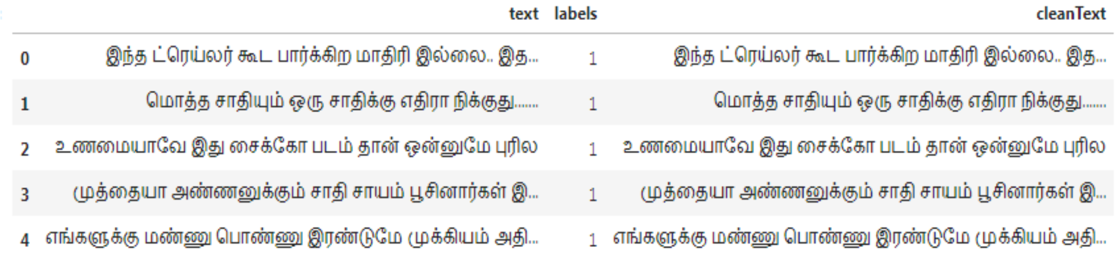
	text
0	திருமலை நாயக்கர் பேரவை சார்பாக படம் வெற்றி பெற...
1	இந்த ட்ரெய்லர் கூட பார்க்கிற மாதிரி இல்லை.. இத...
2	மைசூரு செட்டியார் சமூகத்தின் சார்பாக இப்படம் வ...
3	மொத்த சாதியும் ஒரு சாதிக்கு எதிரா நிக்குது.......
4	only for விஜய் சேதுபதி and STR

Figure 2: Dataset

3.2. Preprocessing

3.2.1. Data Cleaning

Data Cleaning is the process of removing hashtags, URLs, and user mentions, filtering special characters such as “&” and “$”, removing the double spaces, and removing emojis, punctuations, links, etc., which is not required for the corpus (as shown in fig. 3).



	text	labels	cleanText
0	இந்த ட்ரெய்லர் கூட பார்க்கிற மாதிரி இல்லை.. இத...	1	இந்த ட்ரெய்லர் கூட பார்க்கிற மாதிரி இல்லை.. இத...
1	மொத்த சாதியும் ஒரு சாதிக்கு எதிரா நிக்குது.......	1	மொத்த சாதியும் ஒரு சாதிக்கு எதிரா நிக்குது.......
2	உணமையாவே இது சைக்கோ படம் தான் ஒன்னுமே புரில	1	உணமையாவே இது சைக்கோ படம் தான் ஒன்னுமே புரில
3	முத்தையா அண்ணனுக்கும் சாதி சாயம் பூசினார்கள் இ...	1	முத்தையா அண்ணனுக்கும் சாதி சாயம் பூசினார்கள் இ...
4	எங்களுக்கு மண்ணு பொண்ணு இரண்டுமே முக்கியம் அதி...	1	எங்களுக்கு மண்ணு பொண்ணு இரண்டுமே முக்கியம் அதி...

Figure 3: Clean Text

3.2.2. Language Identifier

After cleaning the text, it is passed into the identifier to detect the Language. The dataset includes both texts as well. But for designing the model we included max of Tamil texts [14]. After detecting the language next level of stop-word removal is applied. We have manually created a set of stopwords for the Tamil Language (as shown in fig 4). 125 words included. Next Tokenization which is a process of splitting the words into tokens.

பல
ஆகும்
அல்லது
அவர்
நான்
உள்ள
அந்த
இவர்
சில
என்
போன்ற

ஒரு
என்று
மற்றும்
இந்த
இது
என்ற
கொண்டு
என்பது
என
முதல்
என்ன
இருந்து

வேண்டும்
வந்து
இதன்
அது
அவன்
தான்
பலரும்
என்னும்
பிறகு
அவர்கள்
வரை
அவள்
நீ

ஆகிய
இருந்தது
உள்ளன
வந்த
இருந்த
மிகவும்
இங்கு
மீது
ஓர்
இவை
இந்தக்
பற்றி

பின்
சேர்ந்த
ஆகியோர்
எனக்கு
இன்னும்
அந்தப்
அன்று
ஒரே
மிக
அங்கு
பல்வேறு
விட்டு

அதில்
நாம்
அதற்கு
எனவே
பிற
சிறு
மற்ற
விட
எந்த
எனவும்
எனப்படும்
எனினும்

மேலும்
பின்னர்
கொண்ட
இருக்கும்
தனது
உள்ளது
போது
என்றும்
அதன்
தன்

அடுத்த
இதனை
இதை
கொள்ள
இந்தத்
இதற்கு
அதனால்
தவிர
போல
வரையில்
சற்று
எனக்

பெரும்
அதை
பற்றிய
உன்
அதிக
அந்தக்
பேர்
இதனால்
அவை
அதே
ஏன்

வரும
வேறு
இரு
இதில்
போல்
இப்போது
அவரது
மட்டும்
இந்தப்
எனும்
மேல்

Figure 4: Tamil Stopwords

3.2.3. Stemming & Lemmatize

The terms stemming drop the conditions to their stem. It aims to identify the basic significance of a word. Tokens are stripped of their stemming step affixes and other lexical elements, leaving only the stem [11]. For example, “played” and “playing” both stem into “play”. In general stemming English words is easier compared to the Tamil Language. Here we have proposed a Rule-based Iterative Preprocessing Algorithm for Tamil Language stemming.

Suffixes are frequently used in Tamil to indicate tense, plurality, person, and other concepts. Therefore, a procedure is written for each category to handle the removal of the corresponding suffixes after the suffixes are classified into categories. There is a process to correct or recode the word's ending once the suffix is removed for each category in order to make it consumable for the following routine [13]. Every procedure check for the string's current length before removing the suffix. The suffixes are eliminated after the prefixes. Before moving forward and after removing a suffix, each method for removing suffixes verifies the length of the string and invokes the routine for repairing the endings.

Algorithm: Rule-based Iterative Preprocessing Algorithm

Input: List of Preprocessed Tamil Texts
Output: Stemmed Words (Root)

Step 1: Convert all the plural into singular
Step 2: Singular words matched with the suffix list. If a match is found then it will be written the root word. Else Again the rule iterated to find the exact match.
Step 3: Word length is calculated. In each Iteration word length is verified to maintain the unified.
Step 4: According to the identified suffix next possible suffix list is generated.
Step 5: Rules satisfied, Display the words.

3.3. Feature Extraction

In order to reduce the impact of tokens from the accessible corpus that occur frequently and are, as a result, experimentally less illuminating than features that occur very sometimes during the training of the model, we used TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF performs well compared to other vectorizers. TF-IDF works both CountVectorizer followed by TF-IDF Transformer. The transformer transforms a count matrix to a standardized Term-Frequency or Term Frequency-Inverse Document Frequency representation [12].

3.4. Classification

We have chosen a base model for our system which is Linear SVC. Linear SVC is based on SVM. Support vector machines (SVMs) are unit-powerful however versatile supervised machine learning, which is frequently applied for classification, regression, and outliers' detection [15]. SVMs area unit is common and memory economical as a result of the use of a set of coaching points within the call performance. Scikit-learn offer 3 categories particularly SVC, NuSVC, and LinearSVC which may perform multiclass classification.

4. RESULT AND ANALYSIS

Performance of the recommended algorithm was analyzed based on the stemmer performance and classification performance.

4.1. Stemmer Performance Analysis

Stemmer Performance analyzed using metrics such as Understemming, Overstemming, and Stemmer Effectiveness [11]. Above metrics are calculated using below equations (1), (2), and (3) respectively.

Understemming= [No. of words understemmed]/ Total Words *100%--- (1)
Overstemming= [No. of words overstemmed]/ Total Words *100%--- (2)
Stemmer Effectiveness= 100% - [Overstemming% - Understemming%] --- (3)

From the below fig.5 Proposed RBIPA algorithm have lesser count of unerstemming words compare to other stemmer algorithm. The proposed RBIPA gives a count of 520 words from the test dataset.

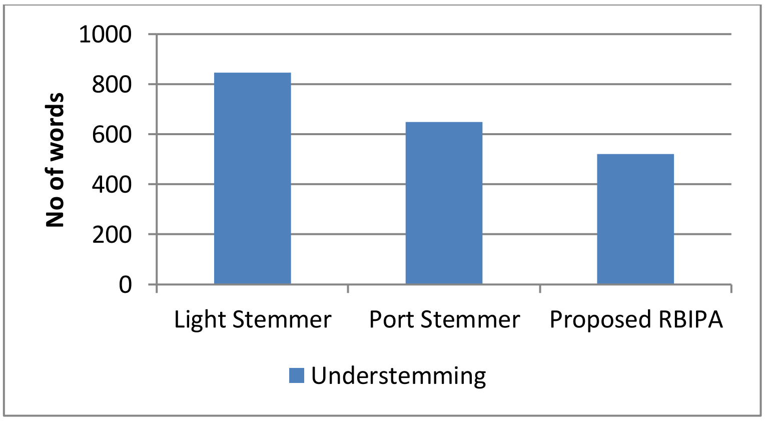


Figure 5: Understemming

From the below fig.6 Proposed RBIPA algorithm have lesser count of overstemming words compare to other stemmer algorithm. The proposed RBIPA gives a count of 1300 words from the test dataset.

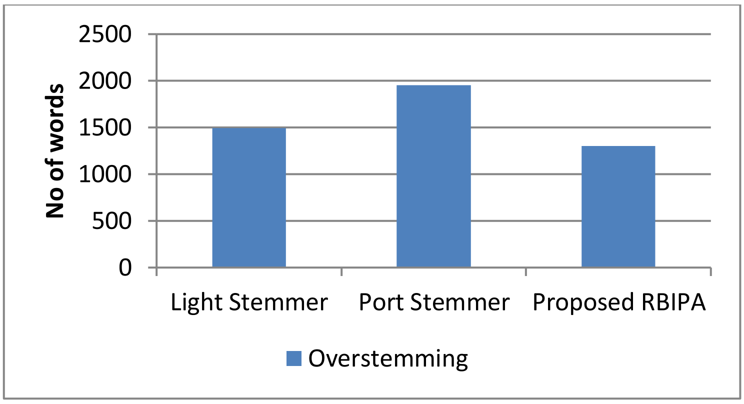


Figure 6: Overstemming

From the below fig.7 Proposed RBIPA algorithm finds max no. of root words compare to other stemmer algorithm. The proposed RBIPA gives a count of 11377 words from the test dataset.

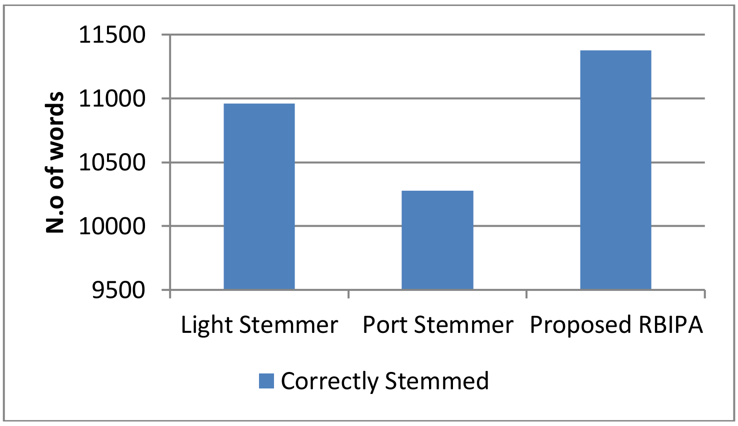


Figure 7: No of words stemmed to root word

4.2. Classification Performance Analysis

The accuracy of the stemmer algorithm is analyzed with previously available stemmers; the proposed RBIPA gives an accuracy of classification 87.52% on the test dataset which shown in fig 8.

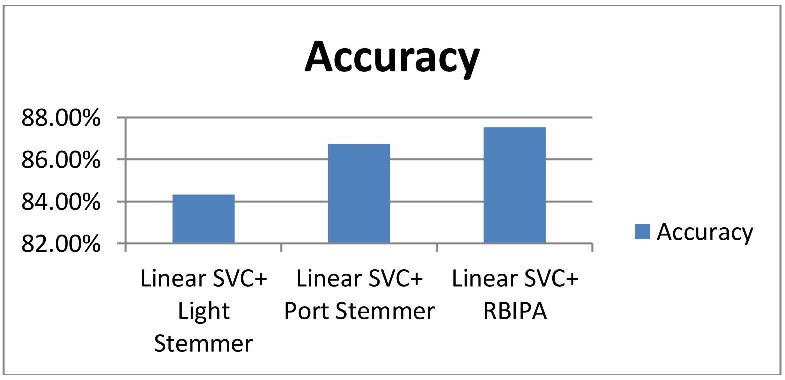


Figure 8: Accuracy

A comment classified and found that max of positive comments is more likely a good sign as in the fig 9.

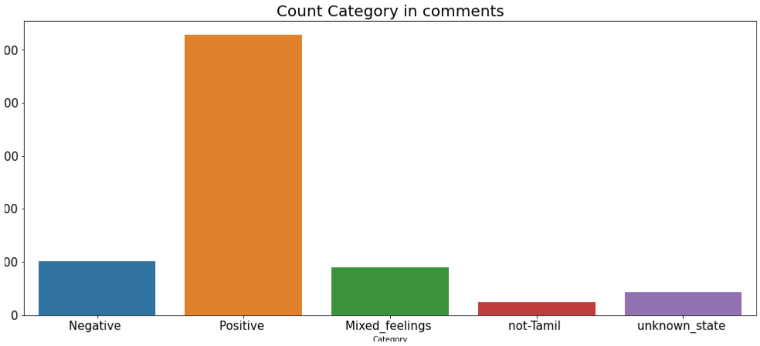


Figure 9: Comments Categorized

5. CONCLUSION

The evaluation of the accuracy of the Light Stemmer Algorithm provides 81.36% accuracy and the Rule-based Iterative Preprocessing Algorithm (RBIPA) shows 84.96% of accuracy in the given Test Dataset which has a total of 13000 words. On the Strength or Performance side, the proposed algorithm which is a Rule-based Iterative Preprocessing Algorithm (RBIPA) changes more words into their stems than the other two stemming algorithms with high accuracy in classification, Stemmer Effectiveness and No of root words identified. Based on these measures, found that proposed algorithm, Rule-based Iterative Preprocessing Algorithm (RBIPA) is the strongest stemmer, with Light Stemmer somewhat weaker but still quite strong.

REFERENCE

[1] Jalal Omer Atoum, "Cyberbullying Detection through Sentiment Analysis", International Conference on Computational Science and Computational Intelligence (CSCI), pp. 292- 297, 2020.

[2] Shankar Biradar and Sunil Saumya,"Transformer-based approach to classify abusive content in Dravidian Code-mixed text", Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pp. 100 – 104, 2022 Association for Computational Linguistics.

[3] Kadambini Swain, Ajit Kumar Nayak, "A Review on Rule-based and Hybrid Stemming Techniques", 2018 2nd International Conference on Data Science and Business Analytics, pp. 25-29, 2018.

[4] Vikram Singh and Balwinder Saini, "AN EFFECTIVE PRE-PROCESSING ALGORITHM FOR INFORMATION RETRIEVAL SYSTEMS", International Journal of Database Management, Vol.6, Is.6, pp. 13-24, 2014.

[5] Charangan Vasantharajan · Uthayasanker Thayasivam, "Towards Offensive Language Identification for TamilCode-Mixed YouTube Comments and Posts", SN Computer Science manuscript, 2021.

[6] Sunita, Vijay Rana, "An Effective Preprocessing Algorithm for Information Retrieval System", International Journal of Recent Technology and Engineering, Vol. 8 Is. 3, pp. 6371-6375, 2019.

[7] Karan Shah , Chaitanya Phadtare , Keval Rajpara "Cyber-Bullying Detection in Hinglish Languages Using Machine Learning", International Journal of Engineering Research & Technology, Vol. 11 Is. 05, pp. 439-447, 2022.

[8] Koirala, Pravesh and Aman Shakya. "A Nepali Rule-based Stemmer and its performance on different NLP applications." ArXiv abs/2002.09901, (2020).

[9] Nahar, Vinita, Xue Li, Chaoyi Pang, and Yang Zhang. "Cyberbullying Detection based on text-stream classification." AusDM 2013 (2013).

[10] Kansal, Rohit, Vishal Goyal, and Gurpreet Singh Lehal. "Rule-based Urdu Stemmer." COLING (2012).

[11] Vivek Anandan Ramachandran & Ilango Krishnamurthi, “An Iterative Suffix Stripping Tamil Stemmer”, Proceedings of the InConINDIA 2012, AISC 132, pp. 583–590, 2012.

[12] K. Arun, A. Srinagesh, “Multi-lingual Twitter sentiment analysis using machine learning”, International Journal of Electrical and Computer Engineering (IJECE) Vol. 10, Is. 6, pp. 5992-6000, 2020.

[13] Bharathi Raja Chakravarthi, Ruba Priyadharshini,Vigneshwaran Muralidaran, Navya Jose, ShardulSuryawanshi, Elizabeth Sherly, John P, McCrae, “DravidianCodeMix: sentiment analysis and offensive language identification dataset for Dravidian languages in the code-mixed text”, Springer, Vol. 56, pp. 765–806, 2022.

[14] Mrs. T. Pratheebha, Mrs. V. Indhumathi, Dr. S. SanthanaMegala, “An Empirical Study On Data Mining Techniques And Its Applications”, International Journal of Software and Hardware Research in Engineering, Vol. 9, Is. 4, pp. 23-31, 2021.

[15] V Indumathi and S Santhanamegala, “CLASSIFY BULLY TEXT WITH IMPROVED CLASSIFICATION MODEL USING GRID SEARCH WITH HYPERPARAMETER TUNING”. Advances and Applications in Mathematical Sciences, Vol. 21, Is. 9, pp.4973-4980, 2022.

AUTHORS

Mrs. V. Indumathi is a Research Scholar and working as Assistant Professor, School of Computer Studies, Rathnavel Subramaniam College of Arts and Science, Coimbatore. I am having 6 years of Industrial Experience. Business statistics, data mining, and full stack development are the areas in which I am most interested in conducting research and teaching. 5 papers were presented at national and international conferences, while 4 research papers were published in peer-reviewed international journals.

Dr.S.SanthanaMegala is an Assistant Professor in the Department of BCA. She is having 10 years of teaching experience. She completed her Ph.D degree from PRIST University, Thanjavur. Business statistics, data mining, data structures, and cloud computing are the areas in which I am most interested in conducting research and teaching. In the field of data mining, I have generated two M.Phil. candidates and am currently supervising four Ph.D. scholars. 13 papers were presented at national and international conferences, while 12 research papers were published in peer-reviewed international journals. "Legal Judgement Summarizer: A Companion for Structured Summary" is a topic covered in a book.