IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

ABUSIVE LANGUAGE DECTECTION USING NLP

[1]Kandarpa Venkata Abhiram, [2]Panigrahi Srikanth
[1]Student, [2]Assistant Professor
[1]Computer Science and Engineering,
[1]Gmr Institute of Technology, Rajam, India

Abstract: The growth of social media platforms has led to a significant increase in user-generated content. These platforms have given users the ability to produce, share, and exchange content in order to interact and communicate with one another. Threatening and abusive language spreads quickly on social media, but it can be stopped if we can find and take it down. These would have, given online bullying and haters new ways to express their hate to a bigger audience while frequently remaining anonymous. We require a strong and efficient automatic system to identify threat and abusive languages due to the increased popularity of social media platforms like Facebook, Twitter, Instagram, etc. and the excessive number of social media users. Many automated methods using machine learning, deep learning, and natural language processing (NLP) have been developed in the past due to the severe and frequent nature of this activity. This paper provides a thorough summary of the reducing techniques that the research in this field has suggested for detecting offensive content. A classification of several methodologies and features used by researchers in the detection process is offered based on algorithms and methods used in detection techniques. This study also covers the major problems that require detailed research in this field. Finally, some future research directions for creating reliable social media content detection systems are also mentioned.

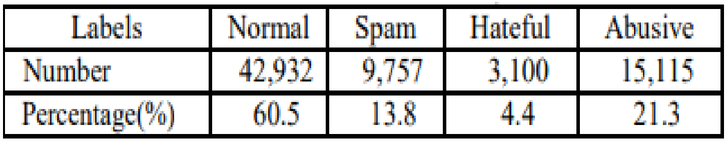
Index Terms* - *Social media; Abusive Content; Cyberbullying; Natural Language Processing; Machine Learning; Deep Learning.

I. Introduction

There is always a real chance that someone will be ridiculed or even harassed online, whether they are participating in comments, message boards, or social media. To combat abusive language, many internet companies have standards and guidelines that users are required to follow, as well as systems that use regular expressions and blacklists to catch bad language and remove a post, are used in conjunction with human editors. While automatically detecting abusive language on the internet is a crucial topic and task, the prior art is not very well-unified, stifling development. Several similar methods have been published in the last three years because previous research has spanned a variety of fields, including Artificial Intelligence (AI), Natural Language Processing (NLP), and Web Sciences. In addition, abusive language may serve as a generic term. In this paper, we aim to address the aforementioned shortcomings in the field while also developing a state-of-the-art method for detecting abusive language in user comments.

II. DATASETS

The lack of a standard dataset for hate speech detection makes it difficult to compare procedures and results using various data and comments. The datasets were developed for various goals, therefore they have distinct properties and show various forms of hate speech. Creating datasets for this activity takes time because the number of hateful instances in social networks is small, but a dataset must include a significant number of such cases. A number of datasets are also hidden to the general public. This could be due to concerns about privacy or the nature of the datasets, such as rudeness and inappropriate language. As previously noted, the goal of this suggested system is to create a classifier that uses BiRNN to classify text for a specific user comment. Table 1 displays the distribution of the 100k tweets in this dataset.



Labels	Normal	Spam	Hateful	Abusive
Number	42,932	9,757	3,100	15,115
Percentage(%)	60.5	13.8	4.4	21.3

FIG:EXAMPLE DATASET

III. LITERATURE SURVEY:

In paper [1] Nobata, Chikashi, et al. "Abusive language detection in online user content." Proceedings of the 25th international conference on world wide web. 2016.In addition to addressing the aforementioned gap in the area, algorithms presented in this research seek to create a cutting-edge way for identifying offensive language in user comments. The contributions made in this study are as follows:To outperform a deep learning system, this research uses supervised classification methodology with NLP features. utilises and modifies a number of the previous art features in an effort to compare their performance with the same data set. Add features from distributional semantics techniques to the feature list as well.Making a fresh data set of a few thousand user comments gathered from various domains public. This set comprises three judgments per remark and, for those considered abusive, a more detailed assessment of each comment's nature.

In paper [2]Davis, Dincy, Reena Murali, and Remesh Babu. "Abusive Language Detection and Characterization of Twitter Behavior." arXiv preprint arXiv:2009.14261 (2020).The main goal is to concentrate on numerous abusive behaviours on Twitter and determine whether or not a communication is abusive. The suggested BiRNN is a better deep learning model for automatically detecting abusive speech, according to results of comparisons between Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) approaches for various abusive behaviours in social media.The suggested method for detecting abusive language was assessed using two measures, namely accuracy and F1-measure. In the future, the effectiveness of the proposed system will be assessed using a variety of domain datasets, including those from Facebook, Wikipedia, Twitter, and other online communities, in order to generalise user behaviour.

In paper [3] Vidgen, Bertie, et al. "Challenges and frontiers in abusive content detection." Association for Computational Linguistics,2019.This paper discusses about the issues constrain, the performance, efficiency and generalizability of abusive content detection systems. Abusive content detection is a pressing social challenge for which computational methods can have a hugely positive impact. In this paper methods deals with critical insights into the challenges and frontiers facing the use of computational methods to detect abusive content. They differ from most previous research by taking an interdisciplinary approach, routed in both the computational and social sciences.

In paper [4] Rajamanickam, Santhosh, et al. "Joint modelling of emotion and abusive language detection." arXiv preprint arXiv:2005.14028 (2020).Aiming to tackle this problem, the natural language processing (NLP) community has experimented with a range of techniques for abuse detection. While achieving substantial success,these methods have so far only focused on modelling the linguistic properties The aim of our work is to investigate the relationship between emotion and abuse detection, which is likely to be independent of the biases that may exist in the annotations. They proposed a new approach to abuse detection, which takes advantage of the affective features to gain auxiliary knowledge through an MTL framework

In paper [5] Kanan, T., Aldaaja, A., & Hawashin, B. (2020). Cyber-bullying and cyber-harassment detection using supervised machine learning techniques in Arabic social media contents. Journal of Internet Technology, 21(5), 1409-1421.This paper mainly focused on detecting these phenomena on English text, few works studied this phenomenon on Arabic. To evaluate the performance of the classifiers, we use Recall, Precision, and F1-Measure. Future scope: it has provided a comprehensive comparison that would aid future research works in this direction.

In paper [6] Kaur, S., Singh, S., & Kaushal, S. (2021). Abusive Content Detection in Online User-Generated Data: A survey. Procedia Computer Science, 189, 274-281. The aim of this survey paper is to help newcomers and budding researchers to obtain an overall perspective of this research area by offering a thorough overview to gain insights of the area, including recent trends and proposed techniques. The researchers have successfully applied methods from the machine-learning field, with Bag-of-Words (BoW) and N-grams being the most frequently used features in classification.

In paper [7] Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019, August). Challenges and frontiers in abusive content detection. Association for Computational Linguistics.

These issues constrain the performance, efficiency and generalizability of abusive content detection systems. In this article we delineate and clarify the main challenges and frontiers in the field, critically evaluate their implications and discuss solutions. Detecting abusive content generically is an important aspiration for the field. However, it is very difficult because abusive content is so varied. Abusive content detection research is currently marked by too much of what Sartori labels 'high 'and 'low' level abstraction. In this paper we have summarized and critically discussed these issues and proposed and discussed possible solutions.

In paper [8] Urrutia Zubikarai, A. (2020). Applied NLP and ML for the detection of inappropriate text in a communications platform (Master's thesis, Universitat Politècnica de Catalunya). The main objective of this work, a model able to classify any sentence in Spanish in appropriate or non-appropriate have to be created. In addition, a study of different type of techniques and models should be done to extract conclusions and start a new line of work and investigation in the Next Ret SL company. The context and culture of each human, make this labelling more diffuse and difficult.There are many pre-processing tools available but, in this case, only 4 of the most used and suitable with the Tellfy environment are analysed. As Python is a very used language in Data Science community and almost all the scripts in the Tellfy project are in this language, only tools with support for Python have been analysed. (Spacy,NLTK Web).

In paper [9] Althobaiti, M. J. (2022). BERT-based Approach to Arabic Hate Speech and Offensive Language Detection in Twitter: Exploiting Emojis and Sentiment Analysis. International Journal of Advanced Computer Science and Applications, 13(5).

The experiments shows that BERT-based model gives the best results, surpassing the best benchmark systems in the literature, on all three tasks:

(a) offensive language detection with 84.3% F1-score,

(b) hate speech detection with 81.8% F1-score, and

(c) fine-grained hate speech recognition (e.g., race, religion, social class, etc.) with 45.1% F1- score.

They reported that the Fast Text DL model outperformed an SMV classifier trained on character n-gram features. Mohaouchane, Mourhir, and Nikolov [48] explored the use of AraVec word embeddings and four DL models: Bidirectional Long Short-Term Memory (Bi-LSTM), Bi-LSTM with an attention mechanism, Convolutional Neural Network (CNN), and a combined model of CNN and LSTM. The experiments illustrated the outperforming results of the CNN over all other models.

In paper [10]Balayn, A., Yang, J., Szlavik, Z., & Bozzon, A. (2021). Automatic Identification of Harmful, Aggressive, Abusive, and Offensive Language on the Web: A Survey of Technical Biases Informed by Psychology Literature. ACM Transactions on Social Computing (TSC), 4(3), 1-56.This paper aim both at systematically characterizing the conceptual properties of online conflictual languages, and at investigating the extent to which they are reflected in state-of-the-art automatic detection systems. We discuss diverse research opportunities for the computer science community and reflect on broader technical and structural issues. we discuss the creation of datasets for online conflictual language detection.

In paper [11]Ballinger, N. (2022). Using a BERT-based Ensemble Network for Abusive Language Detection. In This model uses sub-models using the BERT architecture trained on datasets labelled for hate speech, offensive language and abusive language. The approach presented in this paper is able to outperform the standard BERT model, and Hate BERT, a re-trained variation of the BERT model used for detecting abusive language and other similar tasks. This allows the network to learn through a feedback loop instead of the traditional feed-forward neural networks that only use outputs as the input to the next successive layer. This approach was coded in Pytorch (using TensorFlow as the backend since BERT is a google-based model) and relies on many of the BERT functions from the transformers package. Pre-trained models are extended from Hate BERT and its training on the RAL-E dataset to be used when fine-tuning this new model.

In paper [12]Biere, S., Bhulai, S., & Analytics, M. B. (2018). Hate speech detection using natural language processing techniques. Master Business Analytics Department of Mathematics Faculty of Science.This paper also applies a current technique in this field on a dataset This classifier assigns each tweet to one of the categories of a Twitter dataset: hate, offensive language, and neither.The performance of this model has been tested using the accuracy, as well as looking at the precision, recall and F-score. The final model resulted in an accuracy of 91%, precision of 91%, recall of 90% and a F-measure of 90%.

In paper [13]Kshitiz, K., Singh, H., & Kukreja, P. (2017). Detecting hate speech and insults on social commentary using NLP and machine learning. Int J Eng Technol Sci Res, 4(12), 279-285.This paper involves determining ways to identify bullying in text by analysing and experimenting with different methods to find the most suitable way of classifying bullying comments. We proposed a efficient algorithm to identify the bullying test and aggressive comments and analyses these comments to check the validity. NLP and Machine learning is used for analysing the social comment and identified the aggressive effect of an individual or a group and other related works [14 – 28].

IV. METHODOLOGY

METHODOLOGY-01

The proposed system classifies comments from Twitter platform and the classes includes robust set of abuse related labels.
This work developed a BiRNN detection model by training the architecture with Twitter data. Figure 2 shows the basic architecture
of the proposed abusive language detection system.
The overall system architecture
can be considered as three main modules namely:

1. Text pre-processing module.
2. Embedding module.
3. Model generation.

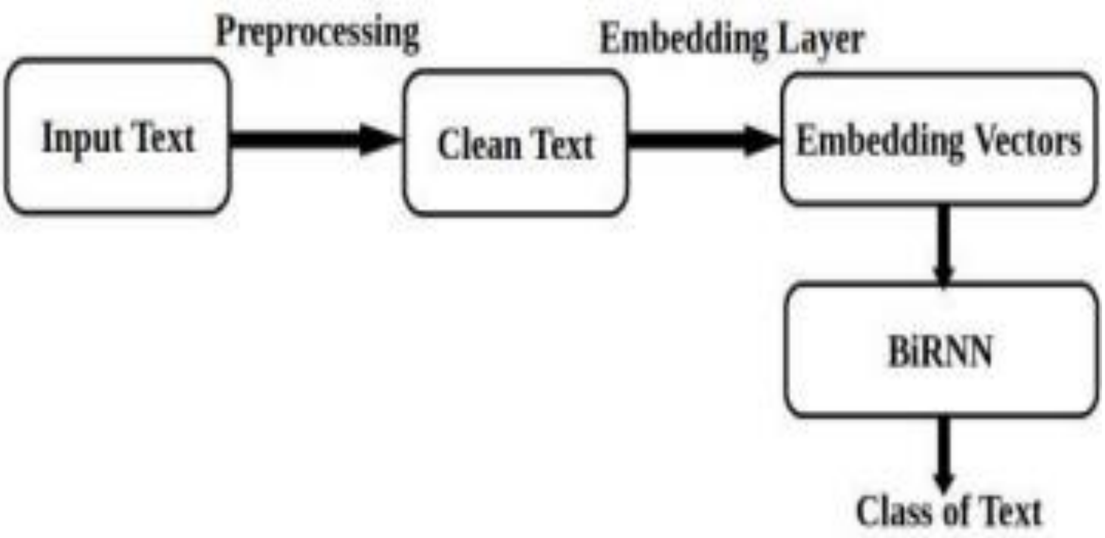


FIG:

METHODOLOGY-02

Ngram Features:
We use character n-grams to model the kinds of conscious or unconscious bastardizations of offensive words, in contrast to previous work in this area that either ignored unnormalized text or used simple edit distance metrics to normalize it.

Linguistic Features:
These features are designed to specifically look for words that inflame people, like using hate lists that already exist, as well as non-abusive language, like using politeness words or modal verbs. Some of these features are:

- Length of comment in tokens
- Average length of word
- Number of punctuations
- Number of periods, question marks, quotes, and repeated punctuation
- number of one letter tokens
- number of capitalized letters
- number of URLS
- number of tokens with non-alpha characters in the middle
- number of discourse connectives

Syntactic Features:
The features are essentially different types of tuples making use of the words, POS tags and dependency relations. These include:

- parent of node
- grandparent of node
- POS of parent
- POS of grandparent
- tuple consisting of the word, parent and grandparent

- children of node8

Distributional Semantics Features:

Numerous natural language processing applications have been successfully facilitated by the concepts of distributed and distributional word and text representations. There are three kinds of embedding-derived features that we use. The first two are based on averaging the word embeddings of all the words in the comment. This is essentially a shallow method used in sentiment analysis to approximate an embedding for a larger piece of text.A great advantage of learning distributed representation vectors for comments in this way is that the algorithm is not sensitive to comment length and it does not require specific tuning for word weights.As a disadvantage, this algorithm needs the constant retraining when new comments are added, which makes the model less efficient for the online applications.

METHODOLOGY-03: BiRNN

BiRNN connect two hidden layers of different directions to the same output. With this form of conceptual deep learning, the output layer can simultaneously obtain information from past (backward) and future (forward) states. This was developed to increase the amount of input available information to the network. For example, MLP have limitations on the flexibility of input data, as they demand their input data to be fixed. Standard RNN also have limitations, as future input information cannot be reached from the current state. BiRNN are particularly useful when an input context is needed. For example, in the case of handwriting recognition, awareness of the letters before and after the current letter can be improved.

The general BiRNN model which involves a text input of embedding followed by a hidden layer. The SoftMax function produce a vector that represents the probability of a list of classes and uses these to generate most probable class in output layer. In short, the network architecture of our model has the following structure:

Embedding Layer: This layer converts each word into an embedded vector.

- Hidden Layer: The hidden layer is a BiRNN. The output of this layer is a fixed size interpretation of its input.
- Output Layer: In the output layer, the interpretation learned from the output layer, RNN passes through a fully connected neural network with a SoftMax output node which classifies the text as abusive, hateful, spam or normal.

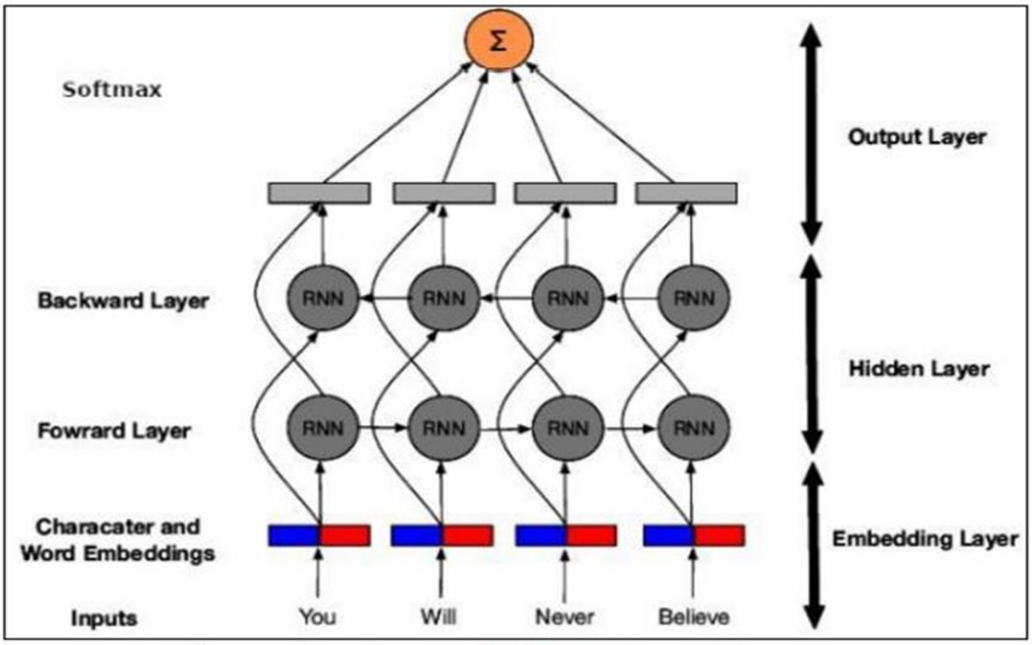


FIG: General BiRNN architecture

V. RESULTS AND DISCUSSION

Results-01-Tools and Libraries:

The tools and libraries used in the proposed system. The sub tasks in the text preprocessing phase has been carried out by Python

Pandas' library, Regular Expressions, Natural Language Tool Kit (NLTK), and some string handling functions in Python.

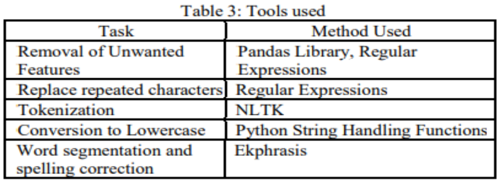


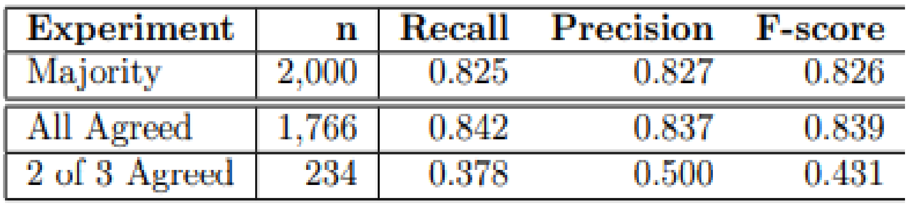
Table 3: Tools used

Task	Method Used
Removal of Unwanted Features	Pandas Library, Regular Expressions
Replace repeated characters	Regular Expressions
Tokenization	NLTK
Conversion to Lowercase	Python String Handling Functions
Word segmentation and spelling correction	Ekphrasis

Table: Tools used

Results-02

This paper evaluated the system when all three raters agreed (unanimous agreement) and where exactly two agreed (and we use their judgment). Although the number of "All Agreed" comments is dominant in this data (1,766 of 2,000), and the difference between labels by the majority vote and those by "All Agreed" are small, the results on the cases where all graders agreed have higher results compared to those with exactly 2 of 3 raters agreed (0.839 to 0.826). The evaluation results of our models on this data are shown:

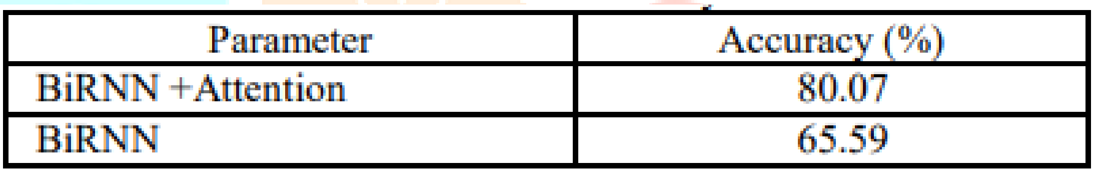


Experiment	n	Recall	Precision	F-score
Majority	2,000	0.825	0.827	0.826
All Agreed	1,766	0.842	0.837	0.839
2 of 3 Agreed	234	0.378	0.500	0.431

Table: Comparison table

Results-03

Many researches proved that the attention mechanism is an effective mechanism to obtain good results in NLP. Here the influence of attention layer [14] in deep learning model on validation accuracy considered. The model is trained with and without attention layer and results were analyzed. The accuracy with attention layer is 80.07% whereas without attention it is not increasing from 65.59% as show in the Table 4. That is attention layer between LSTM layer and output layer helps to focus on important words that have effect on classification. It will assign an attention score to each word in the text. Thus, it indicates the amount of attention that the model allocates for each word.



Parameter	Accuracy (%)
BiRNN +Attention	80.07
BiRNN	65.59

Table: Accuracy table

V. CONCLUSION

As the amount of online user generated content quickly grows, it is necessary to use accurate, automated methods to flag abusive language is of paramount of importance. Not addressing the problem can lead to users abandoning an online community due to harassment or companies pulling advertisements which are featured next to abusive comments. While there has been much work in this area in several different related fields, to date, there has not been a standard evaluation set with which researchers could compare their methods. Additionally, there have been several NLP methods used in prior work but these features have never been combined or evaluated against each other. In our work we take a major step forward in the field by first providing a curated public dataset and also performing several evaluations of a range of NLP features. We experimented with several new features for this task: different syntactic features as well as different types of embeddings features, and find them to be very powerful when combined with the standard NLP features. Character n grams alone fare very well in these noisy data sets. Our model also outperforms a deep learning-based model while avoiding the problem of having to retrain embeddings on every iteration. Next, we used our model to perform an analysis of hate speech over the course of one year, providing practical insight into how much data and what kind of data is necessary for this task. Most work has so far focused on abuse found in English, but it remains to be seen how our approach or any of the other prior approaches would fare in other languages. Given how powerful the two n-gram features were in English, these would probably fare well in other languages given enough training data. Another area of future work includes using the context of the comment as additional features. The context could include the article it references, any comments preceding or replied to, as well as information about the commenter's past behaviour or comments.

References

[1] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016, April). Abusive language detection in online user content. In Proceedings of the 25th international conference on world wide web (pp. 145-153).

[2] Davis, D., Murali, R., & Babu, R. (2020). Abusive Language Detection and Characterization of Twitter Behavior. arXiv preprint arXiv:2009.14261.

[3] Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019, August). Challenges and frontiers in abusive content detection. Association for Computational Linguistics.

[4] Rajamanickam, S., Mishra, P., Yannakoudakis, H., & Shutova, E. (2020). Joint modelling of emotion and abusive language detection. arXiv preprint arXiv:2005.14028

[5] Kanan, T., Aldaaja, A., & Hawashin, B. (2020). Cyber-bullying and cyber-harassment detection using supervised machine learning techniques in Arabic social media contents. Journal of Internet Technology, 21(5), 1409-1421.

[6] Kaur, S., Singh, S., & Kaushal, S. (2021). Abusive Content Detection in Online User-Generated Data: A survey. Procedia Computer Science, 189, 274-281.

[7] Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019, August). Challenges and frontiers in abusive content detection. Association for Computational Linguistics

[8] Urrutia Zubikarai, A. (2020). Applied NLP and ML for the detection of inappropriate text in a communications platform (Master's thesis, Universitat Politècnica de Catalunya).

[9] Althobaiti, M. J. (2022). BERT-based Approach to Arabic Hate Speech and Offensive Language Detection in Twitter: Exploiting Emojis and Sentiment Analysis. International Journal of Advanced Computer Science and Applications, 13(5).

[10] Balayn, A., Yang, J., Szlavik, Z., & Bozzon, A. (2021). Automatic Identification of Harmful, Aggressive, Abusive, and Offensive Language on the Web: A Survey of Technical Biases Informed by Psychology Literature. ACM Transactions on Social Computing (TSC), 4(3), 1-56.

[11] Ballinger, N. (2022). Using a BERT-based Ensemble Network for Abusive Language Detection.

[12] Biere, S., Bhulai, S., & Analytics, M. B. (2018). Hate speech detection using natural language processing techniques. Master Business Analytics Department of Mathematics Faculty of Science.

[13] Kshitiz, K., Singh, H., & Kukreja, P. (2017). Detecting hate speech and insults on social commentary using NLP and machine learning. Int J Eng Technol Sci Res, 4(12), 279-285.

[14] Srikantrh, P., & Behera, C. K. (2022, July). A Machine Learning framework for Covid Detection Using Cough Sounds. In 2022 International Conference on Engineering & MIS (ICEMIS) (pp. 1-5). IEEE.

[15] Srikanth, P., & Behera, C. K. (2022, July). An Empirical study and assessment of minority oversampling with Dynamic Ensemble Selection on COVID-19 utilizing Blood Sample. In 2022 International Conference on Engineering & MIS (ICEMIS) (pp. 1-7). IEEE.

[16] Srikanth, P. (2021). An efficient approach for clustering and classification for fraud detection using bankruptcy data in IoT environment. International Journal of Information Technology, 13(6), 2497-2503.

[17] Panigrahi, S. (2020, April). Design and Analysis of Efficient Cluster Using Novel Dissimilarity Measure and Classification for High Dimensional Cancer Datasets. In Proceedings of the International Conference on Innovative Computing & Communications (ICICC).

[18] Panigrahi Srikanth , Kolla Saitejaswi and Dharmaiah Devarapalli, 'TEJU: Fraud Detection and Improving Classification Performance for Bankruptcy Datasets Using Machine Learning Techniques', international conference on sustainable computing in science, technology and management, ELSEVIER –SSRN ,2019.

[19] Panigrahi Srikanth, Dharmaiah deverapalli and Narsinga Rao M .R "Identification of AIDS Disease Severity Based on Computational Intelligence Techniques Using Clonal Selection Algorithm", International Journal of Convergence Computing –INDERSCIENCE Publications , Volume 2,Issue 3-4,page No:193-207.

[20] Panigrahi Srikanth "Clustering Algorithm of Novel Distribution Function for Dimensionality Reduction Using Big Data of OMICS", 2016 IEEE International Conferences on Computational Intelligence and Computing Research (ICCIC 2016), 2016. IEEE-2016, PP-1-6.

[21] Panigrahi Srikanth and Dr.N. Rajasekhar, "A Novel Cluster Evolution for Gene-miRNA Interactions Documents using Improved Similarity Measure", International Conferences on Engineering &MIC-2016 (ICEMIS -2016), IEEE-Morocco Section, IEEE-2016, PP-1-7.

[22] Panigrahi Srikanth and Dharmaiah Devrapalli, "A Critical Study of Classification Algorithms Using Diabetes Diagnosis", 2016 IEEE 6th International conferences on Advanced Computing (IACC 2016), IEEE 2016, (Google Scholar), 27-28 Feb 2016, PP-245-249.

[23] Panigrahi Srikanth and Dharmaiah deverapalli,"CFTDISM: Clustering Financial Text Documents Using Improved Similarity Measure", 2017 IEEE International Conferences on Computational Intelligence and Computing Research (ICCIC 2017), 2017.

[24] Dhamaiah Deverapalli and Panigrahi Srikanth,'A Novel Fuzzy Rules for Radial Basis Function Network Using BDNF with Type-2 Diabetes Mellitus', International Conference on Intelligent Computing and Communication Technologies - (ICICCT - 2019),springer conference.

[25] Dr.Nimmala Mangathayaru ,B Mathurabai and Panigrahi Srikanth "Clustering and Classification of Effective Diabetes Diagnosis: Computational Intelligence Techniques using PCA with KNN", Springer International Publishing AG 2018 Information and Communication Technology for Intelligent Systems (ICTIS 2017) me 1, Smart Innovation, Systems and Technologies (SIST-Vol-83), PP-426-440.

[26] Panigrahi Srikanth and Dharmaiah Deverapalli, "A Novel Cluster Algorithms of Analysis and predict for Brain Derived Neurotrophic Factor (BDNF) using Diabetes Patients", International Conference on computer and communication technologies - (IC3T 2016), Springer 2016, PP-109-125.

[27] Dharmaiah Deverapalli and Panigrahi Srikanth," Identification of Aids Disease using Genetic Algorithm", International conferences on Computational Intelligence and Soft Computing-(IBCB 2015), springer-2015,pp 99-111.

[28] Dharmaiah Deverapalli, Ch.anusha and Panigrahi Srikanth "Identification of Deleterious SNPs in TACR1 Gene Using Genetic Algorithm", International conferences on Computational Intelligence and Soft Computing-(IBCB 2015), springer – 2015,pp 87-97.