iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** III **Month of publication:** March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67239

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

Automated Offensive Comment Detection Using Text Mining and Deep Learning

Mr. M.S. Sabari[1], Mrs. V. Indumathi[2], Mrs. G. Priyadharshini[3]
[1]*Assistant Professor, Department of CSE, Gnanamani College of Technology, Namakkal, Tamil Nadu, India*
[2]*Assistant Professor, Department of CSE, Gnanamani College of Technology, Namakkal, Tamil Nadu, India*
[3]*PG Scholar, Department of CSE, Gnanamani College of Technology, Namakkal, Tamil Nadu, India*

Abstract: ***Everyone has the right to freedom of expression. However, under the guise of free speech, this privilege is being abused to discriminate against and harm others, either physically or verbally. Hate speech is the term for this type of bigotry. Hate speech is described as language used to show hatred toward an individual or a group of individuals based on traits such as race, religion, ethnicity, gender, nationality, handicap, and sexual orientation. It can take the form of speech, writing, gestures, or displays that target someone due to their affiliation with a particular group. Hate speech has been more prevalent in recent years, both in person and online. Hateful content is bred and shared on social media and other internet platforms, which finally leads to hate crimes. The growing use of social media platforms and information exchange has resulted in significant benefits for humanity. However, this has resulted in several issues, including the spread and dissemination of hate speech messages. Recent studies used a range of machine learning and deep learning techniques with text mining methods to automatically detect hate speech messages on real-time datasets to handle this developing issue on social media platforms. Hence, this paper aims to survey the various algorithms to detect hateful comments and predict the best algorithms in social media datasets. And also implemented in real-time social environments to detect hate speech with mobile intimation.***
Index Terms: ***Social Media, Hate Speech, Machine learning, Deep learning, Text mining***

I. INTRODUCTION

Social media is a fashionable and, above all, simple means for people to publicly share their ideas and opinions while also interacting with others online. It has now become an integral component of human life. It is a stage in which people are easily harassed or abused by others, who express hate in various forms such as sexism, racism, politics, and so on. The use of these social media platforms for cyber tyranny, online nuisance, and blackmail is also on the rise. Social networking sites (SNS) have made it simple for us to connect with numerous societies or organizations that we are interested in. These sites have reached a significant number of individuals in society as a result of the development of numerous abilities such as high-speed internet and handheld devices. The majority of the handlers in these networks are under the age of thirty. Researchers have taken advantage of the vast amounts of data available on numerous social networking sites and undertaken extensive research in a variety of fields. Sentiment Analysis is a popular field of study that utilizes a lot of data from social media. The numerous sorts of social media are depicted in Figure 1.

Fig 1: Social media types

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

II. RELATED WORK

P. Fortuna and S. Nunes, et al.,...[1] examined the difficulties of perceiving hate speech, which is labeled in a variety of platforms and settings, and provides a unified description. This region has undeniable societal impact potential, particularly in online communities and virtual media systems. The advancement of automated hate speech identification requires the enhancement and systematization of shared assets, as well as recommendations, annotated datasets in many languages, and algorithms. Hate speech is a language that offends or degrades, or incites violence or hatred toward businesses, based on specific characteristics such as physical appearance, faith, descent, national or ethnic origin, sexual orientation, gender identification, or other characteristics, and it can occur in a variety of linguistic styles, even in a diffused bureaucracy or when humor is used.

A. Tolba, Z. Al-Makhadmeh, and others, [2] looked at 1500 samples to see if merging device learning approaches with NLP was beneficial. The automated approach was discovered to help improve the detection and prediction of hate speech on social networking websites. Furthermore, this method was found to be more accurate in detecting hate speech and to be more time-efficient than the traditional method. This is because the killer herbal language processing optimizing ensemble deep learning algorithm (KNLPEDNN) was used to analyses Twitter responses and forecast hate and non-hate posts with high accuracy. The proposed method used masses of Tweets as statistics during the self-learning system; it also categorized remarks from beyond facts evaluation, which efficiently decreased the misclassification charge.

R. Cao, R. K.-W. Lee, and T.-A. Hoang, et.al,…[3] developed DeepHate, a single deep-learning model for computerized hate speech detection in social media that uses multi-faceted textual representations. And do excellent experiments on three real-world, publicly available datasets. DeepHate consistently beats state-of-the-art approaches in the hate speech detection task, according to the test findings. After that, behavior empirically investigates the DeepHate version and provides perceptions into the notable capabilities that assisted in recognizing hate speech in social media. The salient feature evaluation additionally improves the explainability of our proposed model

Z. Waseem and D. Hovy, et.al,…[4] supply a data set of 16k tweets with hate speech annotations Also, consider which of the features we use provides the best identification results. We investigate the functions that increase hate speech identification in our corpus and find that, regardless of expected differences in geographic and phrase-duration distribution, they have little to no impact on overall performance and rarely improve over character-degree functions. Gender is an exception to this rule. And he provided a list of criteria based entirely on important race theory for identifying racist and sexist remarks. These can be utilized to obtain more records and address the problem of a small but highly prevalent group of people who are hateful. While the problem is far from being solved, we have discovered that a man or woman n-gram-based entirely method provides a solid foundation. Apart from gender, demographic data delivers minimal improvement, but this is due to a lack of coverage. To update future information and tests, we hope to improve area and gender type.

T. Davidson, D. Warmsley, et.al,…[5] classified tweets as either hate speech, harsh language, or neither. We train a model to distinguish between those categories and then examine the results to help it understand how we will distinguish between them. The findings suggest that fine-grained tags can aid in the detection of hate speech in a publication and highlight a number of important hurdles to effective classification. We conclude that future paintings must better account for context and heterogeneity in the use of hate speech. Also, they gathered tweets containing hate speech key terms using a crowd-sourced hate speech lexicon. We use crowdsourcing to categorize a pattern of these tweets into three groups: those that contain hate speech, those that merely contain offensive language, and those that have neither. To discriminate between these distinct classes, we train a multi-elegance classifier. An examination of the expectations and errors reveals when we can reliably distinguish hate speech from other objectionable words and when this distinction is more difficult. We discovered that racialist and homophobic tweets are more likely to be labeled as hate speech, whereas chauvinist tweets are more likely to be labeled as offensive. It's also more difficult to categorize tweets that don't contain blatant hate phrases.

P. Badjatiya, S. Gupta, et.al,…[6] Logistic Regression, Random Forest, SVMs, Gradient Boosted Decision Trees (GBDTs), and Deep Neural Networks were among the classifiers tested (DNNs). These classifiers' feature areas are specified in turn by project-specific embedding identified using three deep learning architectures: Feed Neural Networks, Convolutional Neural Networks (CNNs), and Long Short-Term Memory Networks (LSTMs). We explore typical spaces such as char n-grams, TF-IDF vectors, and Bag of Words vectors as baselines (BoWV). The complexity of the herbal language constructs makes this assignment very tough. We perform sizeable experiments with multiple deep mastering architectures to research semantic phrase embeddings to handle this complexity.

236

M. O. Ibrahim and I. Budi, et.al,…[7] built an Indonesian Twitter dataset for abusive language and hate speech recognition, including detecting the objective, category, and severity of hate speech. This research discusses multi-label written content grouping for abusive language and hate speech detection in Indonesian Twitter, including detecting the target, category, and level of hate speech using device learning processes with Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest Decision Tree (RFDT) classifiers and Binary Relevance (BR), Label Power-set (LP), and Classifier Chains (CC) as information transformation techniques. Term frequency, orthography, and lexicon functions were among the function extractions we employed. The results of our experiments reveal that the RFDT classifier uses LP during the fashionable period since the transformation approach provides high-quality accuracy with a short calculation time.

I. Alfina, R. Mulia, et.al,…[8] produced a new dataset for hate speech identification in Indonesian, which encompasses hate speech in general, including religious, ethnic, racial, and gender hatred. We also did an initial research to see which combination of device learning rules and features produced the best results. The goal of the task is to find hate words in the Indonesian language. As far as I can tell, there hasn't been much research done on this subject. The most basic research we found has resulted in a dataset for religious hate speech, but the quality of this dataset is insufficient. The researchers wanted to construct a new dataset that included hate speech in general, such as hatred of religion, race, ethnicity, and gender. In addition, we conducted a preliminary investigation using the system learning approach.

To this point, machine learning has been the most widely utilized method of text classification.

M. O. Ibrahim and I. Budi, et.al,…[9] Developed a new Twitter dataset for detecting abusive language in Indonesian. Furthermore, tests in detecting abusive language in Indonesian social media were presented in order to defend the abusive phrases and writing patterns in Indonesian social media. In this paper, we create a new dataset and conduct research on abusive language in the Indonesian language. The test results show that NB outperforms SVM and RFDT in classifying abusive language in all instances using our dataset. When it comes to capability extractions, phrase unigram, and phrase n-gram combinations outperform alternative features such as NB, SVM, and RFDT. The test results also show that categorizing the tweet into three groups (non-abusive language, abusive but not offensive language, and offensive language) is more difficult than just determining whether the tweet is abusive or not. The classifier we used had trouble distinguishing whether the tweet was abusive but no longer offensive or offensive language in this case.

J. Salminen, M. Hopf, et.al,…[10] undertaken the development of an online hate classifier that runs on a mobile platform. This model works well for detecting hateful feedback across multiple social media systems, uses advanced linguistic functions, such as Bidirectional Encoder Representations from Transformers (BERT) (see "BERT" phase), and is made available to researchers and practitioners for similar use and improvement.

Then there was a lot of experimenting with different classification processes and feature representations (Logistic Regression, Nave Bayes, Support Vector Machines, XGBoost, and Neural Networks) (Bag-of-Words, TF-IDF, Word2Vec, BERT, and their aggregate). While all of the models appear to exceed the keyword-based baseline classifier, XGBoost with all of its features performs admirably (F1=0.92). According to the feature significance analysis, BERT skills have the most impact on the forecasts. Because the platform-specific effects from Twitter and Wikipedia are similar to their respective supply papers, the findings suggest the generalizability of the high-quality version. Also, make code freely available for use in real-world software systems as well as for further refinement by online hate researchers.

III. EXISTING METHODOLOGIES

Within the last decade, there has been a substantial increase in research on text classification in social media. Detecting and stopping the use of various sorts of abusive language in blogs, microblogs, and social networks is a particularly useful aspect of this work. In this study, we look at how to find hate speech on social media while separating it from popular vulgarity. We plan to employ supervised category algorithms Within the last decade, there has been a substantial increase in research on text classification in social media.

Detecting and stopping the use of various sorts of abusive language in blogs, microblogs, and social networks is a particularly useful aspect of this work. In this study, we look at how to find hate speech on social media while separating it from popular vulgarity. We plan to employ supervised category algorithms and a recently released dataset annotated for this purpose to construct lexical baselines for this paper. The basic steps for hate speech detection can show in fig 2.

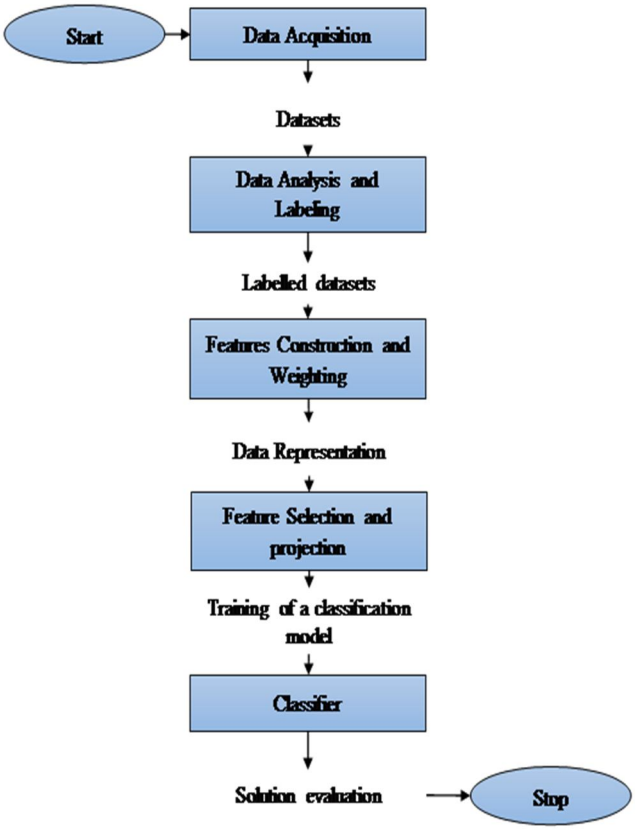


Fig 2: Steps for detecting hate comments

Most social media platforms have implemented individual policies to limit hate speech; however, enforcing these rules necessitates extensive manual labor to review each file. Some platforms, such as Facebook, have recently expanded the number of content material moderators. Automatic technology and methods may be used to speed up the reviewing process or devote human resources to positions that demand a thorough human examination. In this segment, we look at how automated hate speech identification from text works.

A. Keyword-Based Approaches

The employment of a key-word-based technique is a basic approach for determining hate speech. Text that contains potentially hostile keywords is detected using an ontology or dictionary. Hatebase, for example, maintains a database of pejorative words for a variety of companies in 95 languages. As terminology changes with time, such well-maintained artefacts are valuable. However, as we learned during our research into hate speech standards, using a vile slur isn't always enough to be considered hate speech. Keyword-based techniques are quick and simple to grasp. They do, however, face formidable obstacles. Detecting the most common racial insults could result in a highly specific device with low recall, where precision is the proportion of applicable from the set discovered and recall is the percent of relevant from the global population. In other words, a device that is based mostly on key phrases may be unable to detect hateful text that does not contain these phrases. In contrast, including terms that aren't often unpleasant (e.g., "garbage," "swine," and many others.) would result in far too many false alarms, increasing bearing in mind at the expense of precision.

B. Machine Learning Classifiers

A machine learning model uses samples of tagged textual material to create a classifier that can detect disliked speech using labels annotated by content reviewers. Several concepts were proposed and demonstrated to be successful in the hereafter. In this paper, we discuss an augmentation of the open-source structures used in the current study.

1) Content Preprocessing and Function Selection.

Textual content features suggesting hate should be retrieved to become aware of or classify user-generated content material. Individual phrases or sentences with obvious functions (n-grams, i.e., series of n consecutive phrases). Words can be stemmed to improve function matching by removing morphological distinctions from the root. Metaphor processing can extract functionalities as well. In textual content categorization, the bag-of-words assumption is often used. Under this approach, a submission is represented as a set of phrases or n-grams with no particular sequence. This assumption ignores a crucial aspect of languages, but it has proven useful in a variety of situations. There are several approaches for assigning weights to the phrases that are more important in this setting, including TF-IDF, for a current information retrieval overview. Aside from distributional features, phrase embedding, or assigning a vector to a phrase, is prevalent when using deep learning methods in natural language processing and textual content mining, and includes word2vec. The bag-of-words assumption is challenged by several deep learning designs, such as recurrent and transformer neural networks, which simulate the ordering of the words by processing over a succession of word embedding.

2) Hate Speech Detection Procedures and Baselines.

Text categorization models include Nave Bayes, Support Vector Machines, and Logistic Regression. With the assumption that the features do not interact with one another, Nave Bayes models classify changes without delay. SVMs and Logistic Regression are linear classifiers that anticipate lessons based on a mix of ranks for each attribute.

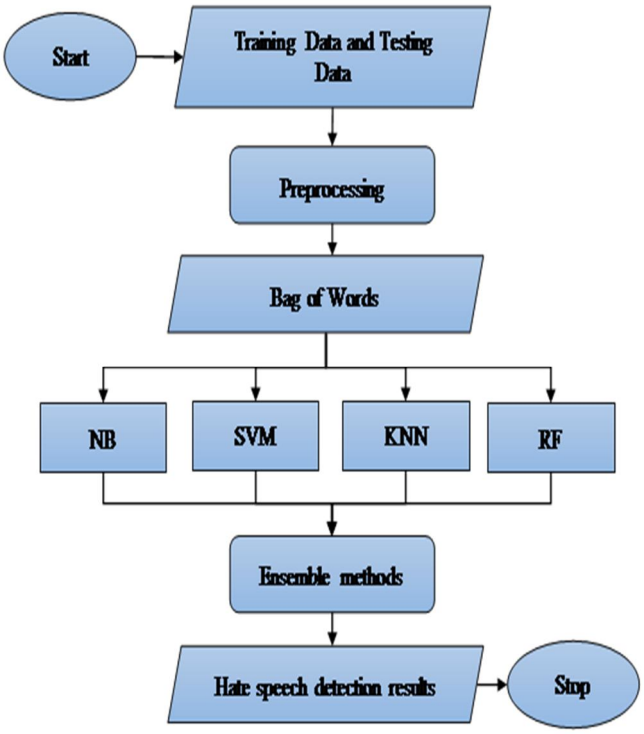


Fig 3: Existing methods for Hate speech detection

IV. PROPOSED METHODOLOGIES

The most effective way to meet new individuals is through social networking sites. People have discovered an illegal and immoral way to use social networking sites as their popularity has grown. The expression of hate and harassment are the most widespread and destructive misuse of online social media. Violence, hostility, bullying, coercion, harassment, racism, insults, provocation, and sexism are all examples of hate speech. These are a few of the most significant online risks to a social networking platform. To classify the data and determine if the remarks are hateful or normal, deep learning-based algorithms are applied.

The script is viewed as a collection of words by feed-forward networks.

RNN-based representations see the text as a collection of words and are useful for capturing word relationships and text structures.

For Term Count, CNN-based models are taught to recognize patterns in text, such as key phrases (TC).

Capsule networks have recently been applied to TC to address the information loss problem caused by CNN pooling operations.
The attention mechanism is active in categorizing related words in text, and it has evolved into a useful tool in DL model development.
Memory-augmented networks combine neural networks with an external memory that allows the models to read and write to datasets.
Graph neural networks are designed to capture interior graph structures of natural language, such as syntactic and semantic parse trees.
Finally, we can review various approaches such as machine learning and deep learning techniques in text classification in social media datasets. The following figure 4 shows the proposed framework and description.

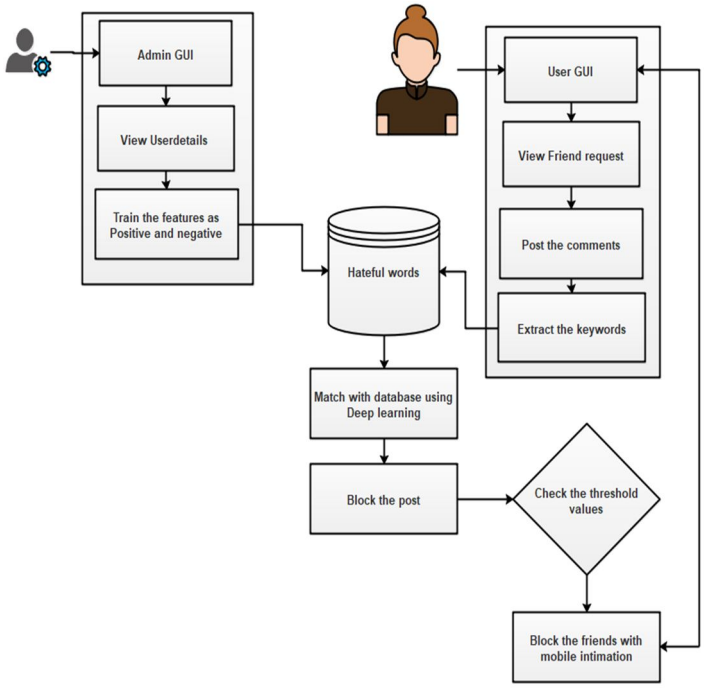


Fig 4: Proposed Work

The extraction and selection of a set of characterizing and discriminating features is the focus of most efforts in developing a robust deep-learning classifier. The steps of the text-mining algorithm are as follows:

- Tokenize text-based reviews as single terms
- Analyze unigrams, bigrams, and n-grams
- Remove stop words, analyze stemming words, and remove special characters
- Finally, extract key phrases
- Analyze extended words that can be substituted with right words

A database of categorized terms is created here, which is then used to check the words for any inappropriate words. If the communication contains any vulgar terms, the message will be submitted to the Blacklists, which will filter those words out. Finally, due to the content-based filtering technique, a message free of obscene terms will be posted on the user's wall. As follows is the suggested deep learning classifier:
Step 1: Initialize the neural network model
Step 2: Specify the layer type as convolution, max pooling, fully connected layers
Step 3: Activate the layers
Step 4: Specify the inputs and neurons
Step 5: Construct key terms as positive and negative
Step 6: Match with testing keywords
Step 7: Label as “positive” and “negative”

function INITCNNMODEL (θ, [n1–5])
layer type = [convolution, max-pooling, fully-connected, fully-connected];
layerActivation = [tanh(2), max(),softmax()]
model = new Model();
for$i$=1 to 4 do
layer = new Layer();
layer.type = layerType[i];
layer.input size = ni
layer.neurons = new Neuron [ni+1];
layer.params = θi;
model.addLayer(layer);
end for
return model;
end function

A system uses blacklists to automatically reject undesired messages based on both message content and message author relationships and characteristics. The extension of the collection of features evaluated in the classification process, a different semantics for filtering rules to better match the considered domain, to help the users Filtering Rules(FRs) specification, and a different semantics for filtering rules to better fit the considered domain.

V. EXPERIMENTAL RESULTS

In this chapter, we can construct the social network using ASP.NET as front end and SQL SERVER as Back end. The performance of the system can be analyzed in terms of F-measure parameter.

The performance of the system is evaluated using Precision, Recall and F-measure.

Precision =$\frac{TP}{TP+FP}$

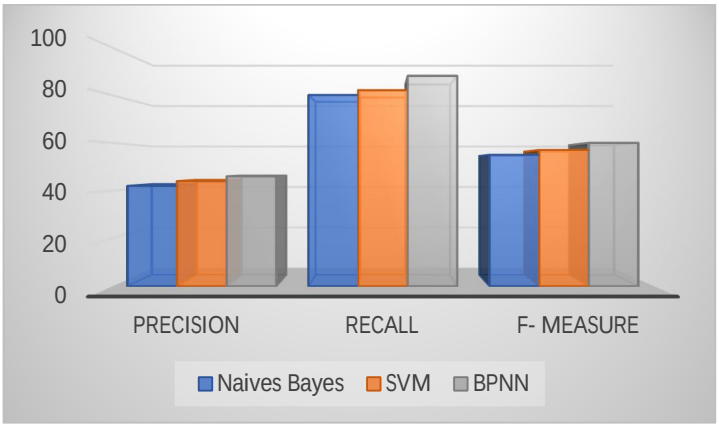
Recall =$\frac{TP}{TP+FN}$

F measure = 2* $\frac{Precision*Recall}{Precision+Recall}$

The performance evaluation result is shown in following table 1 and shows in fig 3.

Algorithm/ Performance measures	Precision	Recall	F- measure
Naives Bayes	42	80	55
SVM	44	82	57
BPNN	46	88	60

Table 1: Performance Table



(a)

Fig 5: Performance chart

From the above calculation, proposed neural network algorithm provide high level F-measure values than the existing Naives Bayes and SVM algorithm

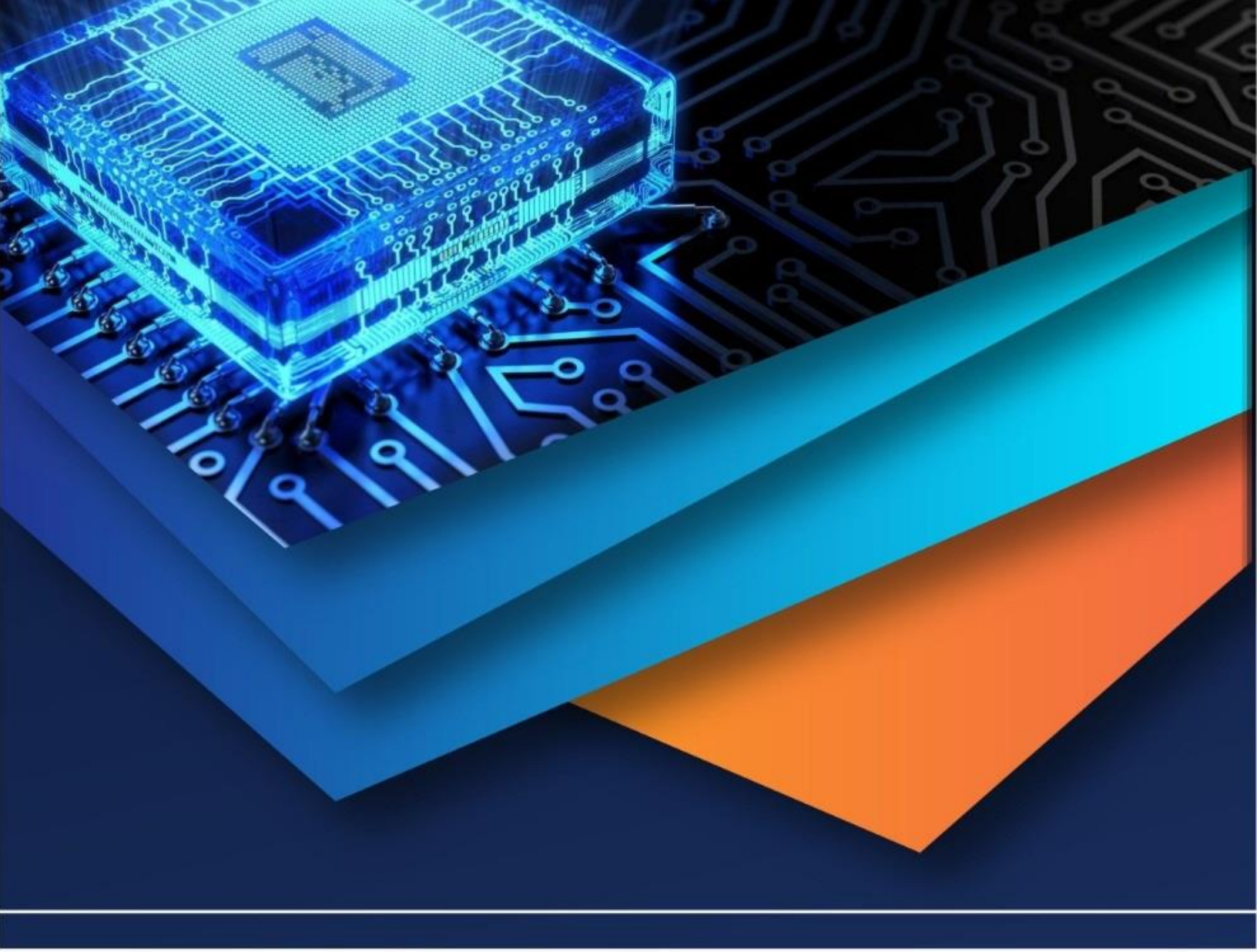
VI. CONCLUSION

We can survey the existing machine learning deep learning models in this research. We may conclude that deep learning models can be used to solve a variety of problems. The widely used machine learning and deep learning approaches for text classification were explored and compared in this work. We discovered that several forms of BPNN perform well in sequential learning tasks and solve the problems of disappearing and explosion of weights in standard text classification algorithms when learning long-term relationships in this work. Furthermore, the performance of BPNN models can be affected by hidden size and batch size.

REFERENCES

[1] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Comput. Surv., vol. 51, no. 4, pp. 1–30, Sep. 2018.
[2] Z. Al-Makhadmeh and A. Tolba, "Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach," Computing, vol. 102, no. 2, pp. 501–522, Feb. 2020.
[3] R. Cao, R. K.-W. Lee, and T.-A. Hoang, "DeepHate: Hate speech detection via multi-faceted text representations," in Proc. 12th ACM Conf. Web Sci., Southampton, U.K., Jul. 2020, pp. 11–20.
[4] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in Proc. NAACL Student Res. Workshop, San Diego, CA, USA, Jun. 2016, pp. 88–93.
[5] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in Proc. ICWSM, Montreal, QC, Canada, May 2017, pp. 15–18.
[6] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in Proc. 26th Int. Conf. World Wide Web Companion (WWW Companion), Perth, WA, Australia, Apr. 2017, pp. 759–760.
[7] M. O. Ibrohim and I. Budi, "Multi-label hate speech and abusive language detection in Indonesian Twitter," in Proc. 3rd Workshop Abusive Lang. Online, Florence, Italy, Aug. 2019, pp. 46–57.
[8] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," in Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS), Jakarta, Indonesia, Oct. 2017, pp. 233–238
[9] M. O. Ibrohim and I. Budi, "A dataset and preliminaries study for abusive language detection in Indonesian social media," Procedia Comput. Sci., vol. 135, pp. 222–229, Jan. 2018.
[10] J. Salminen, M. Hopf, S. A. Chowdhury, S.-G. Jung, H. Almerekhi, and B. J. Jansen, "Developing an online hate classifier for multiple social media platforms," Hum.-centric Comput. Inf. Sci., vol. 10, no. 1, pp. 1–34, Dec. 2020
[11] A. Jha and R. Mamidi, "When does a compliment become sexist? Analysis and classification of ambivalent sexism using Twitter data," in Proc. 2nd Workshop NLP Comput. Social Sci., Vancouver, BC, Canada, Aug. 2017, pp. 7–16.
[12] S. Yuan, X. Wu, and Y. Xiang, "A two phase deep learning model for identifying discrimination from tweets," in Proc. EDBT, Bordeaux, France, Mar. 2016, pp. 696–697.
[13] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on BERT model," PLoS ONE, vol. 15, no. 8, pp. 1–26, Aug. 2020.
[14] P. Burnap and M. L. Williams, "Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making," Policy Internet, vol. 7, no. 2, pp. 223–242, Jun. 2015.
[15] M. Wiegand, J. Ruppenhofer, and T. Kleinbauer, "Detection of abusive language: The problem of biased datasets," in Proc. HLT-NAACL, Minneapolis, MN, USA, Jun. 2019, pp. 602–608.
[16] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on BERT model," PLoS ONE, vol. 15, no. 8, pp. 1–26, Aug. 2020.
[17] P. Burnap and M. L. Williams, "Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making," Policy Internet, vol. 7, no. 2, pp. 223–242, Jun. 2015.
[18] M. Wiegand, J. Ruppenhofer, and T. Kleinbauer, "Detection of abusive language: The problem of biased datasets," in Proc. HLT-NAACL, Minneapolis, MN, USA, Jun. 2019, pp. 602–608.
[19] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. John, N. Constant, M. Guajardo-Céspedes, S. Yuan, C. Tar, Y. Sung, and R. Kurzweil, "Universal sentence encoder," in Proc. EMNLP, Brussels, Belgium, Mar. 2018, pp. 169–174.
[20] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter," in Proc. 13th Int. Workshop Semantic Eval., Minneapolis, MN, USA, Jun. 2019, pp. 54–63.

242



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY