Journal of Theoretical and Applied Information Technology
30th November 2022. Vol.100. No 22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AUTOMATED HATE SPEECH CLASSIFICATION USING EMOTION ANALYSIS IN SOCIAL MEDIA USER GENERATED TEXTS

AIGERIM TOKTAROVA[1], ZHALGASBEK IZTAEV[2], PERNEKUL KOZHABEKOVA[3], NABAT SUIEUOVA[4], ROBERT OWINO OPONDO[5], MUKHTAR KERIMBEKOV[6], ZHANNA ZHUNISBEKOVA[2]

[1]Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan
[2]M.Auezov South Kazakhstan University, Shymkent, Kazakhstan
[3]M.Auezov South Kazakhstan University, Shymkent
[4]Yessenov University, Aktau, Kazakhstan
[5]Nazarbayev Intellectual School- Physics and Mathematics, Shymkent, Kazakhstan
[6]University of friendship of people's academician A. Kuatbekov, Shymkent
E-mail: [1]aikerimtoktarova@gmail.com

ABSTRACT

As a result of the ease with which any viewpoint may be posted on social networking sites, online hate speech has become more widespread in recent years. This trend is mostly attributable to the fast growth of mobile computers and the Internet. Studies that were done in the past demonstrate that being exposed to hate speech online has substantial implications in real life for historically disadvantaged populations. As a result, there has been a lot of interest in research on the automatic identification of hate speech. Nevertheless, there has not been a lot of research done on how social networking sites might help identify communities that are prone to hate crimes. It is possible for hate speech to have an effect on any demographic group; however, certain groups are more susceptible to the effects of hate speech than others. For example, it is difficult for racial or ethnic groups whose languages have limited computing resources to automatically gather and evaluate online generated texts. This is to say nothing of the difficulty of automatically detecting hate speech on social networking sites. In this article, we present a method for the identification of hate speech posted on social networking sites, applying artificial intelligence methods in text processing and natural language processing. In order to detect, hate speech on social media, firstly, we collect data by using different keywords. Secondly, we apply machine learning algorithms to classify texts into several categories. Thirdly, we evaluate the proposed approaches and assess the result of hate speech detection problem in training and test sets.

Keywords: *Artificial Intelligence, Social Networks, Hate Speech, Classification, Detection.*

1. INTRODUCTION

Nationality, philosophy, religious beliefs, and economic standing are only few of the components that go into constructing an event or person. Even if this variety is something to be celebrated, it is important to note that the numerous personas that individuals carry might add up to result in several different types of discrimination. The above, in turn, may render a person more susceptible to hate groups, which are acts inspired by prejudice against the identity of a person [1]. Hatred is directed towards certain groups of people based on their membership in a particular category of human across the globe. The discourse of identification serves as the foundation for the argument that underpins the targeting of certain groups. The formation of certain narratives that are aimed to harmonize opposing personalities has been reported by racists as well as anti-racists [2]. Thus, the same idea could easily be implemented in multicultural nations where individuals are driven by communal hate merely targeting various ethnicities of people and rendering them more susceptible. This would be an example of a scenario in which the same theory might be applied. It has been claimed that incidences of purported hate crimes have occurred in a variety of nations, with the victim being a member of a disadvantaged population in each case. Simply being human makes a person susceptible to a variety of forms of damage, including physical injury, dependence on other people, and loss of authority.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The vulnerability might be attributed to one or more particular characteristics, situations, or kinds of groups [3].

Before classifying texts to extremist-related or neutral, we need to define danger criteria. One solution is to prepare a set of keywords. For the definition, a set of key phrases was prepared, applied to explore data in the Vkontakte social network [4]. Referring to the indicated keywords or phrases in the text, the software package infers that the text is applicable for further study. Figure 1 shows the entire data collection, analysis of posts, and classification of texts.

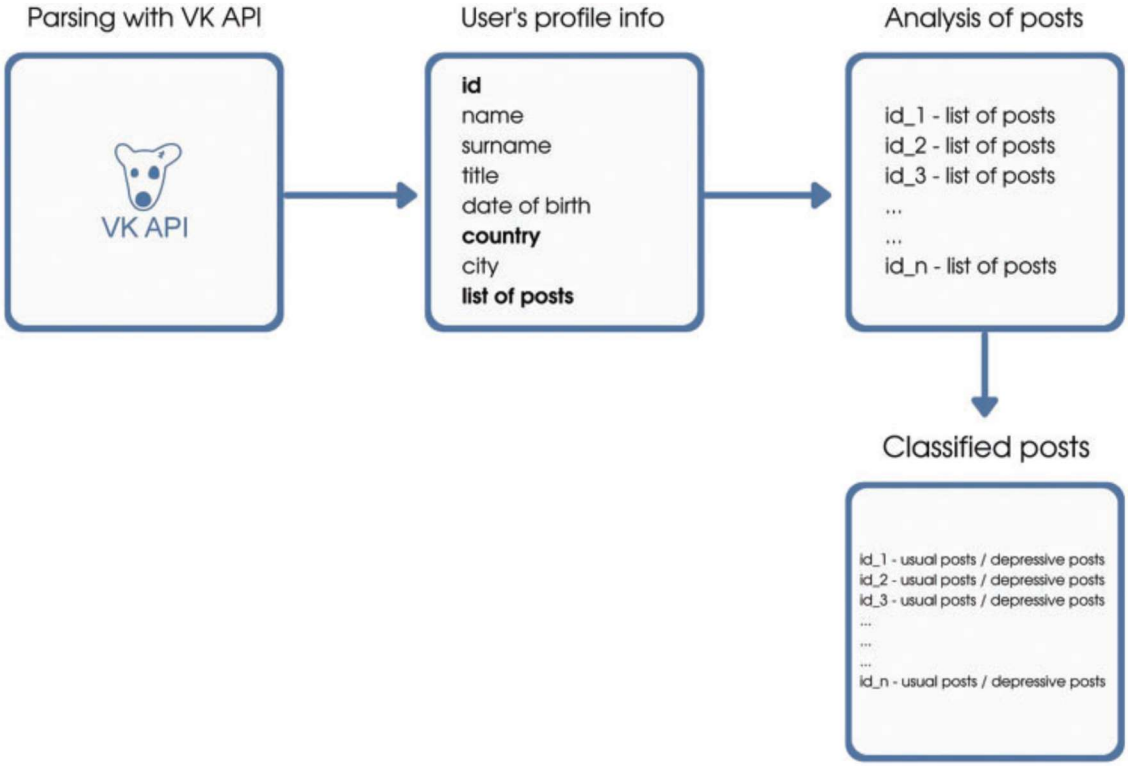

Figure 1: Scheme Of Data Acquisition, Analysis And Classification Of Posts

There has been a rise in the number of hostile acts that take use of the technology provided by online social networking sites alongside the meteoric rise in the number of social interactions that take place on social networking websites. Tweets or posts that include harsh language directed towards specific persons or groups of people are considered to be hateful messages on social networks [5]. The identification of such hostile speech is critical for conducting public sentiment research on one user group's attitude toward another user group and for preventing wrongdoing that is related with such speech. It is also helpful to filter tweets prior to providing topic recommendations or teaching artificial intelligence enabled chatterbots to learn from tweets.

As a result of technological advancements such as the use of social media platforms for the purposes of communicating, expressing ideas, interacting with people, as well as discovering and disseminating information, vulnerability has become a rescuer. In spite of the fact that social networks provide an essential platform for communication that may take place in a quick and effective manner, same platforms are also used as a way of disseminating hate speech worldwide. This is as a result of the fact that widely used social networking sites, such as Facebook and Twitter, allow for the unrestricted dissemination of material that includes hate speech, sensitive information, and debatable subjects. According to research conducted by state-of-the-art researches and surveys of people who use the internet, online hate speech has major repercussions offline, both for organizations and for people. There is a significant correlation between the dissemination of hate speech and the actual commission of hate crimes against the community that is the focus [6].

The topic of the regulation of hate speech is one that is continuously being discussed [7-10]. It is not quite obvious at this time whether the best way to respond to it is via legal action or through some other approach (such as counter-speech and education [11]). The obvious damage that may be caused by hate speech [12-14] makes it essential to identify it, regardless of the methods that can be used to combat it. Automatic identification of offensive and hostile material is necessary because of the sheer amount of content that is produced online, especially in social media, as well as the psychological strain that comes with human moderation [15].

There has been a rise in the number of hostile acts that take use of the technology provided by online social networking sites alongside the meteoric rise in the number of social interactions that take place on social networking websites. Tweets or posts that include harsh language directed towards specific persons or groups of people are considered to be hateful messages on social networks. The identification of such hostile speech is critical for conducting public sentiment research on one user group's attitude toward another user group and for preventing wrongdoing that is related with such speech. It is also helpful to filter tweets prior to providing topic recommendations or teaching artificial intelligence enabled chatterbots to learn from tweets.

2. RELATED WORKS

2.1 Social Network Analysis for Hate Speech Detection

Because of the Internet, access to information and the dissemination of that information may occur at a rate that has never been witnessed before. It is also a good medium for the spread of content that is damaging or hateful because of this capability, in addition to the possibility of keeping one's identity [16]. Because of this, a significant number of researchers have investigated the difficulty of automatically detecting this phenomenon [17, 18], and there have also been a significant number of competitions devoted to the solution of problems that are comparable to this one (for instance, aggression [19], or hateful/offensive content [20]). The use of core templates or key words [21] is one strategy that may be utilized in order to address the challenge. The concept of combining the procedures of extracting characteristics with the more classic approaches of artificial intelligence is becoming an increasingly frequent approach. Bag-of-Words (BoW) is a method that was applied by Kwok and Wang [22], and they combined it with a Naive Bayes classifier in order to get their desired results. Grevy et al. [23] used BoW in conjunction with Support Vector Machines; however, we are going to employ it on its own without the SVMs (SVMs). Others, on the other hand, employed more complicated ways to offer features for conventional machine learning algorithms (such as SVM, Naive Bayes, and Logistic Regression) [24-27]. This was done because BoW had a high percentage of false positives [28], which led to this finding. Consequently, this was found out. Experiments were carried out by Salminen et al. [29] using a variety of other classifications in addition to the hate and non-hate categories (such as accusation, humiliation, and so on). They suggested use a support vector machine (SVM) with a linear kernel.

In today's world, customers are able to interact through social media platforms for a cost that is almost indistinguishable from zero marginal dollars. Any user who has access to a connection to the internet that is fairly priced has the capability of broadcasting any kind of message on these platforms, and it has the potential to reach a big number of people in a relatively short period of time. Everyone now has the ability to publish anything they want, and anybody who is interested in the information may get it. This quality has developed into one that is more democratic. Democracy is to blame for the huge societal shifts that have taken place as a result of recent events. The transformative potential of social networking sites carries with it a variety of challenges, some of which may be harmful to certain groups of people [30]. This predicament manifests itself in a number of various ways, one of which is the prompt acknowledgment by the authorities of many countries that the propagation of hate speech online is a big worry. This is one of the many diverse expressions that this dilemma takes. The No Hate Speech Movement is an initiative that has the endorsement of the Council of Europe, which is why it is called that. The fundamental purpose of the initiative, which has been given the official title Countering Online Hate Speech, is to provide assistance to countries in dealing with the problem. It should not come as a surprise to learn that the goal of making it illegal to express hatred in any form is the impetus behind the vast majority, if not all, of the present efforts being made in this field. On the other hand, the bulk of current activities are centered on doing research on the comments that are made on online forums or websites that are affiliated with radical groups. In addition, there have been a number of studies that have been written and published on the subject of determining what constitutes hate

speech [31]. Since relevant corpora are readily accessible, the great majority of research on hate speech has been carried out using data from the English language. This is due to the fact that relevant corpora are freely available. However, in more recent years, there have been a few studies that looked at the identification of hate language in Kazakh, Dutch, and Indonesian respectively.

2.2 Keyword Analysis for Hate Speech Detection

The great majority of the resources provide user-generated public material, the vast majority of which is comprised of posts made on microblogs. These items are often retrieved using a method that is centered on keywords, and the vast majority of the time, they make use of terms that have a connotation that is unfavorable. During the process of building corpora, we made use of keyword-based data collection tactics in order to address the problem of biases; some authors have progressed beyond the plain procedures based on the vocabulary and have welcomed the use of other collection techniques or combined collection strategies. In some cases, the keyword-based strategy is combined with fetching the entire timeframe from users or pages that are regarded hateful, i.e., when it is likely to find hateful items, or from discussion forums about controversial topics that can easily trigger a certain language [32-33]. In other words, the likelihood of finding hateful content is increased. This is done while keeping in mind the constraints that come with gathering material from a diverse group of individuals. The authors of employed a combined strategy to gather abusive and misogynistic tweets. This method included monitoring possible victims of hate accounts, downloading the history of identified haters, and filtering Twitter streams with keywords. The tweets that were discovered in were collected with the help of this approach In a few other instances, a kind of a preconceived classification is delegated to the reading material on the recovery source, with the assumption being created that all of the goods managed to gather from a certain source may be deemed to be hateful. This kind of categorization is assigned to the texts based on the assumption that all of the items gathered from a certain source may be deemed to be offensive. [34] use a data sets that was retrieved from an internet site that gathers sudden reports by Internet users of any material that contains HS or child sexual abuse. The corpus is then validated by experts who decide that more than 40% of the subject matter is not actually disturbing, and that only 3% of the content could be considered illegal. This is an extremely novel approach that has not been taken before.

The great majority of the corpora that were investigated in this work were collected by providing lists of keywords to the application programming interfaces (APIs) of social media sites. It is not always the case that these keywords are terms that are seen as being obviously harsh or insulting. In point of fact, they are chosen in such a manner as to be free of any possible harmful repercussions, and this is taken into consideration throughout the selection process [35]. This is done so that samples of both good and bad usage of hate speech as well as other types of abusive language may be collected . However, the technique for collecting data that is based on keywords does still introduce a bias into the data in terms of the topics that they cover, and as a consequence, it has an influence on the representativity of the corpora. Keywords are the most common way that researchers gather data.

The research carried out by Wiegand et al. (2019) investigates the likelihood of topic bias in a variety of abusive language corpora that were collected via the use of keyword searching. They acquire lists of phrases that have a substantial association with abusive microposts by computing their Point - wise Mutual Knowledge, and they utilize these lists to generate word filters. The experiment reveals that some datasets include a certain degree of topic bias, which has adverse effects for the use of these datasets in deep learning: an unsupervised system may learn that phrases associated to themes like as football are predictive of hate speech [36].

In the same way that hostile organizations have distinct speech patterns, communities that are comprised of individuals who are the targets of hateful speech also have language standards that are unique to their membership. These standards are based on the fact that these individuals are the targets of hateful speech. The term "support group" will be used to refer to them in a more comprehensive meaning. It is crucial to notice that support organizations and groups that spread hate speech against them frequently join in dialogue on similar problems, but with completely distinct purposes [37]. The conversation may be on the same topic, but the motivations for their participation are entirely different. Groups that promote fat shaming and communities that support people who are overweight or obese discuss the challenges that come with having a high body mass index (BMI),

and communities of women and communities of misogynists discuss the importance of gender equality. Because these subjects often cross over into one another, there is a considerable likelihood that they may utilize language that is comparable to one another, which may create confusion among those who are attempting to categorize them [38].

In addition, many strategies that are based on keywords employ provocative phrases that are already well-established and well-known to the general public in order to target certain groups. This is done in order to appeal to those specific populations. Even if the use of such keywords will undoubtedly capture some hateful speech, it is very common for people to communicate their hatred in ways that are not as overt, without resorting to typical slurs and other types of derogatory terminology. This is because it is more comfortable for them to do so [39].

Even while none of these terms is inherently antagonistic, when they are employed in this context, they aggressively degrade the group to whom they are ascribed. Hateful speakers, for example, often use the term "parasites" to refer to migrants and refugees, and they label African-Americans as "animals."

It is reasonable for us to assume that classifiers that were trained on phrases that were overtly antagonistic may miss posts that employed strategies that were more subtle or depending on the context. In addition, keywords may be masked in a number of different ways, such as via the use of homophones, misspellings, character substitutions (such as substituting letters with symbols), and so on. It is common practice to participate in these kinds of actions in order to avoid being discovered by keyword-based filters that are used by various websites. Taking action against hate speech in online communities "web of hatred" [40].

2.3 Features for Hate Speech Detection

The authors constructed bigram, unigram, and trigram features for each tweet after first changing each tweet to lowercase and then running it through the Porterstemmer. They gave higher weight to the features that had the greatest TF-IDF values. In order to gather information on the syntactic structure, we make use of NLTK to create Penn Part-of-Speech (POS) tagunigrams, bigrams, and trigrams. For the purpose of determining the overall quality of each tweet, we apply a reworked version of the Flesch-Kincaid Grade Level and the Flesch Reading Ease ratings. One phrase is all that is allowed to be included in each tweet at any one time. In addition to this, we assign scores to each tweet based on its sentiment by using a set of emotions that has been developed for use exclusively on social media. In addition to features that assess the total number of characters, words, and syllables in each tweet, we also have binary and count indications for hashtags, mentions, retweets, and URLs. The authors of the study, [41-43]. The issue of inflammatory language as well as the identification of hate speech by automation. As published in the proceedings of the annual AAAI conference on the web and social media.

When it comes to any text categorization task, the most obvious information to employ are surface-level features like bag of words. This applies to any activity that requires organizing text in some way. In point of fact, the majority of authors make use of feature sets that include unigrams as well as larger n-grams. This is because these types of phrases are easier to recognize. It is often asserted that these traits have a high capacity for predictive analysis. Despite this, n-gram traits are often combined with a broad range of other characteristics in a variety of works. For instance, [44] state in their most recent work that even though token and character n-gram features are the most predictive single features in their studies, combining them with all other features further increases performance. This is because combining them with all other features allows for more accurate predictions. The results of their research confirmed that this is the case.

Dealing with the problem of spelling variation is one of the most typical issues that arises while working with material that was provided by users as comments. The answer to this issue may be found in the use of character-level n-gram characteristics. The phrase "ki11 yrslef a$$hole," for instance, which is considered to be an example of hate speech, will most likely pose problems to token-based approaches due to the unusual spelling variations, which will result in very rare or even unknown tokens in the training data. This is an example of a phrase that is considered to be an example of hate speech. This particular phrase is an instance of something that is an example of something that is regarded to be hate speech. On the other hand, character-level strategies are the ones that have a greater probability of capturing the likeness to the canonical spelling of these tokens. This is because character-level techniques operate at a more granular level. A systematic comparison that

was carried out by [45] found that character n-gram features are more predictive of hate speech detection than token n-gram features are. This conclusion was reached as a result of the findings of the study.

However, in order for these features to operate effectively, it is important for predicted phrases to appear in both the training data and the test data. Bag-of-words features often produce a high classification performance when utilized to the detection of hate speech. The detection of hate speech, on the other hand, is often carried out on very little snippets of text, which means that one might potentially run into a problem with the lack of data (for example, paragraphs or even individual lines). As a result of this, a number of different works address this issue by using a wide variety of different sorts of word generalization. This can be accomplished by first clustering the words in a document, and then using the IDs generated by the clustering process to represent groups of words as additional (generalized) characteristics. This can be done in two steps. Step one clusters the words in the document. Step two uses the IDs generated by the clustering process. Brown clustering is one example of a frequent approach that may be used for this purpose. Warner and Hirschberg have made use of it as a feature, and it's an example of how it might be used [46]. Latent Dirichlet Allocation (LDA) [47] creates for each word a topic distribution that reflects the degree to which a word corresponds to each subject. In contrast, Brown clustering generates hard clusters, which is another name for allocating each particular word to a certain cluster. In contrast to this, the Brown clustering method, which is responsible for producing the hard clusters, A method similar to this one has been used to the aforementioned data in order to identify instances of hate speech [48].

Considerations of language use also play an important part in the process of determining whether or not communication constitutes hate speech. Either a method that is more broad is employed to use linguistic characteristics, or the characteristics themselves are specifically suited to the task at hand.

Researchers from [49] study what occurs when you mix ngram features with tokens that have had POS information added to them. On the other hand, the performance of the classifier is not much improved by the addition of information about POS locations.

The study conducted by Chen et al. (2012) makes use of typed dependency connections since these connections take into account more in-depth syntactic information as a feature. Such connections offer the potential benefit that non-consecutive words that convey a (potentially long-distance) link may be recorded in one feature. This is an advantage that can be gained from such linkages. It is possible that this will be helpful in circumstances in which the connection is not immediately clear. As an example, in (4) there will be a dependency tuple that indicates the connection between the insulting term "pigs" and the group of people who are the subject of hatred. The notation for this tuple is going to be nsubj (pigs, Jews) [50].

3. MATERIALS AND METHODS

This section demonstrates whole framework of hate speech detection using machine learning techniques in natural language processing. In the first place, we collect data from social networking sites using hate speech related keywords. Figure 2 demonstrates overall framework of automated hate speech detection. In the second place, we develop a corpora of texts that classified to two parts that contains hate speech and does not contain hate speech related texts. In the third place, we train machine learning algorithms to classify hate speech related texts. In the third place, we test and evaluate the applied machine learning algorithms.

Journal of Theoretical and Applied Information Technology
30th November 2022. Vol.100. No 22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

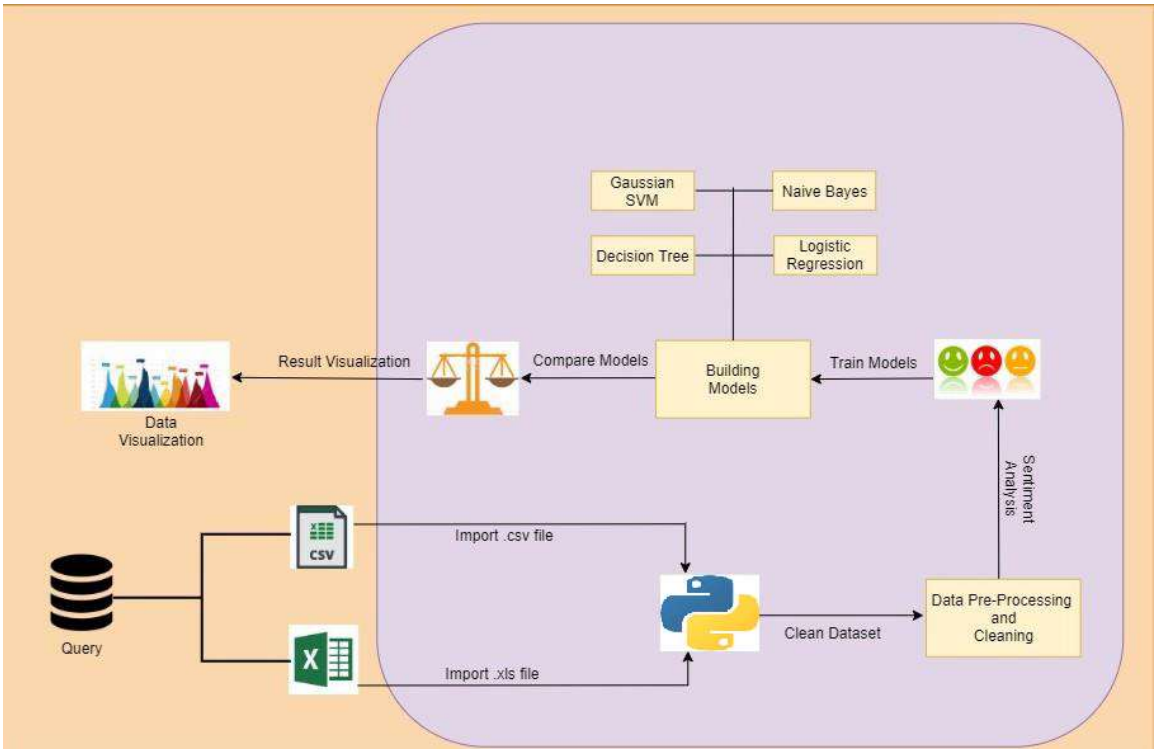

Figure 2: Framework Of The Proposed Study

Figure 3 demonstrates data collection and data preprocessing stage. In this stage, we get original text from social networks using hate speech related keywords. After that, we define different n-grams that in combination give hate speech related key phrase. Next stage is tokenization process. Tokenization involves splitting the raw text into tiny parts. The raw text is broken up into words and phrases that are referred to as tokens via tokenization. The context may be understood better or a model for natural language processing can be developed with the assistance of these tokens. By evaluating the order in which the words appear, tokenization provides assistance in deciphering the meaning of the source text. After tokenization we applied three methods as part of speech, stopwords removing, and named entity removing for data cleaning and preprocessing stage. Thus in the result, we will get preprocessed text that can be applied to train machine learning models.

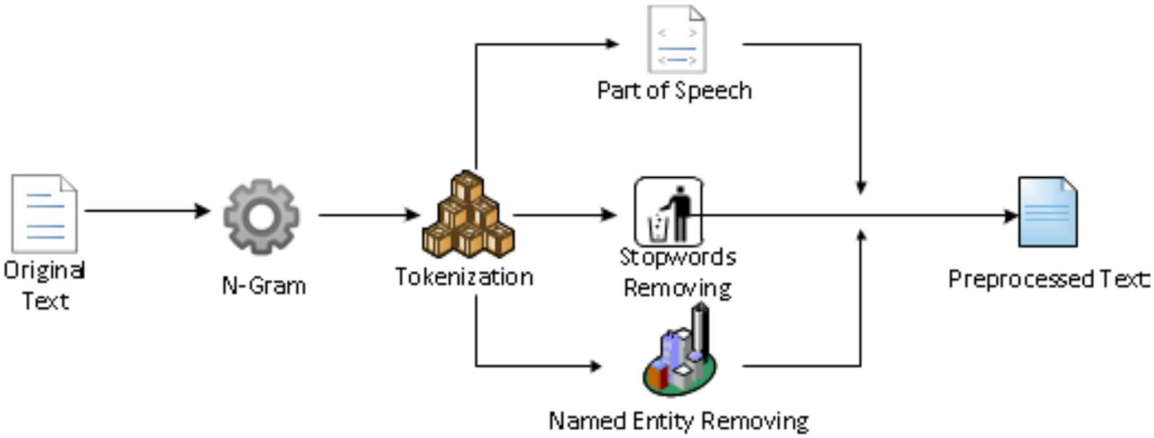

Figure 3: Description Is Placed Right Below The Figure

Thus, in this section we explained the proposed approach for hate speech related content detection in social networks that contain different stages like getting data, define n-grams, tokenization of texts, data cleaning using different techniques like part of speech, stopwords removing, named entity removing, and in the result getting the preprocessed texts.

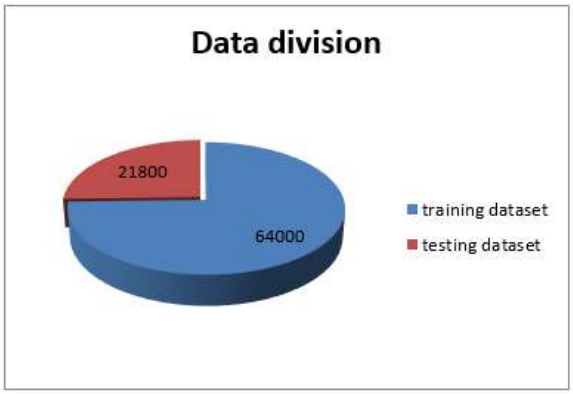

Figure 4: Model Testing

Figure 4 demonstrates a dataset that we collected from social networks. The dataset is divided into two parts as training set and test set. The proposed dataset consists of 85800 sentences. 64000 (74.6%) of them are applied to train the models, 21800 (25.4%) sentence are applied to test the model.

4. EVALUATION CRITERIA

The mean average precision (MaP) and the average recall (AR) at different degrees of intersection over union are the numerous metrics that are used in the process of evaluating the suggested model (IoU). In classification problems involving localization and object detection, the ratio of the areas of the bounding boxes is most commonly used as a metric to determine the reliability of the location of the bounding box. This is because the ratio of the areas of the bounding boxes is directly proportional to the accuracy of the location of the bounding box.

$$TPR = SEN = \frac{TP}{P} \quad (1)$$

Since the purpose of the assessment is to identify as many instances as possible from a population for a screening procedure, the number of false negatives should be maintained to a minimum while at the same time the number of false positives should be increased. As a consequence of this, it is necessary to calculate three major metrics: the true positive rate (TPR), the false positive rate (FPR), and the accuracy (ACC). In the world of medicine, the first factor is known to as sensitivity (SEN), and it is represented mathematically as equation (2) [51]:

$$FPR = \frac{FP}{N} \quad (2)$$

Where the number of true positive is TP, and the number of positive instances is P.

The estimation of the second term, true negative rate or specificity, expressed as equation (3) [52]:

$$TNR = SPEC = \frac{TN}{N} = 1 - FPR \quad (3)$$

The cumulative number of negative occurrences in the population is denoted by the letter N, the percentage of false positives is denoted by the letter FP, and the number of genuine negative samples is denoted by the letter N. On the other side, this statistic is best understood when seen as the ratio of genuine negatives to true negatives. This ratio is referred to as the specificity (SPEC) in the medical field, and it is represented by the equation (4) [53]:

$$ACC = \frac{TP + TN}{P + N} \quad (4)$$

Last but not least, precision is what establishes the equilibrium between genuine positives and genuine negatives. When the number of good and negative occurrences is not equal, this statistic has the potential to be a very helpful tool. The expression for this is the equation (5) [54]:

5. RESULTS

In this section, we divided the experimental results into two subsections. In the first subsection, we demonstrate results of hate speech training detection. In the further subsections, we present hate speech text detection results. In the second section, we demonstrate how the proposed model works in real time and show visual presentation. In addition, we indicate source images and hate speech analysis. In Subsection 3, we illustrate evaluation results of the proposed model by showing different evaluation parameters as precision, recall, f-score for each classified classes of road surface damages.

Journal of Theoretical and Applied Information Technology
30th November 2022. Vol.100. No 22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Evaluation of the proposed method by classes.

Algorithm	Accuracy	Precision	Recall	F1-score
Decision Tree				
Gradient Boosting decision tree	92%	88%	89%	90%
K nearest neighbours	90%	88%	90%	89%
Logistic regression	92%	91%	90%	91%
Neural Network	88%	89%	82%	83%
Naïve Bayes	89%	91%	90%	90%
Random Forest	91%	91%	92%	90%
Support Vector Machine	92%	93%	93%	91%

Figure 4 demonstrates model training and model testing results. The results show that the proposed model achieved to 95% accuracy in about 100 epochs. In the figure, blue line means accuracy of the model, orange line means validation accuracy in model training.

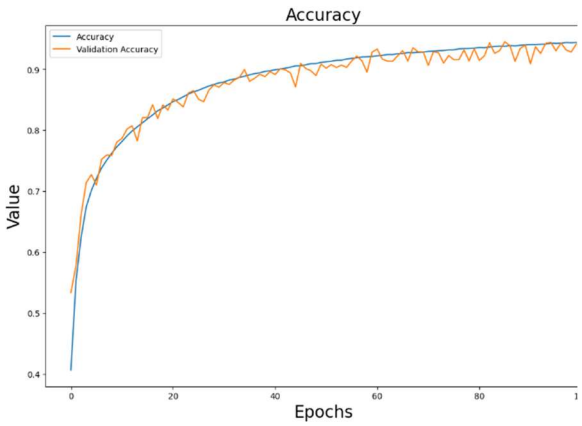

Figure 4: Model Training

Figure 5 demonstrates model testing results. As in the previous figure, in Figure 4, blue line stands for model accuracy and orange line stands for validation model accuracy in testing the proposed model. The figure show that the proposed framework can be successively detect hate speech related texts in social networks.

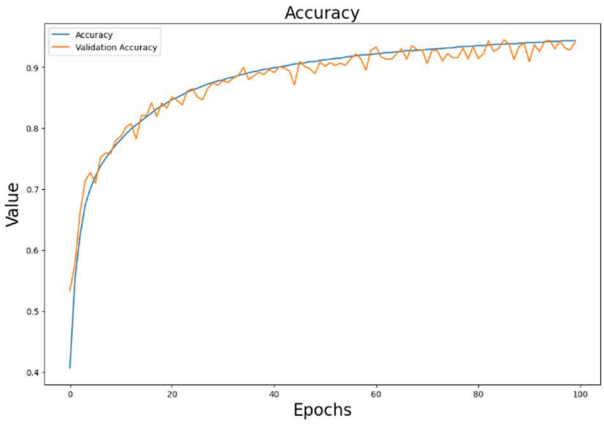

Figure 5: Model Testing

Figure 6 demonstrates confusion matrix for classification of hate speech related texts into three classes. Using confusion matrices, one can clear visualize exactly number of classification results in relation to other classes. There are three types of classes as hate speech related that is noted as 1, non-hate speech related that is noted as 0, and neutral class that is noted as 2 in our study. As we have three classes that should be categorized we have 3x3 confusion matrix. As the confusion matrix show, that the results are high and can be applied for hate speech detection on social networking sites.

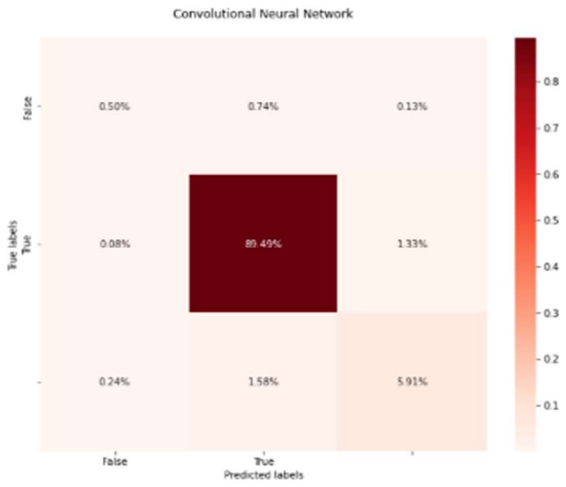

Figure 6: Confusion Matrix

Figure 7 demonstrates classification of hate speech related contents vs. other domain texts using different machine learning techniques. There, we applied machine learning models in different topics like jokes, news, toxic contents, spams, and advertising in order to understand permanency of the

proposed techniques to different datasets. The results show that, the proposed techniques give high accuracy in different types of datasets.

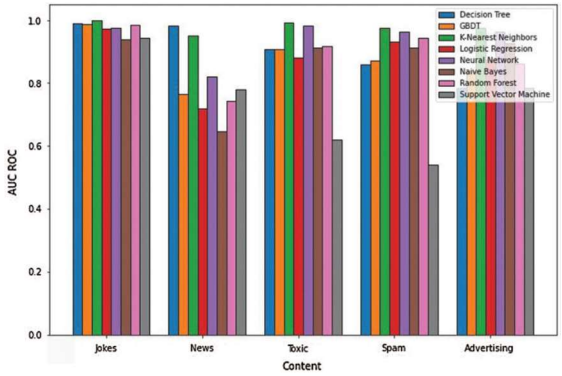

Figure 7: Classification for hate speech related content vs. other domain texts

Figure 8 demonstrates area under the curve receiver operating characteristics AUC-ROC curve for hate speech detection in social networks. It shows relations of false positive rate to true positive rate. The results give that hate speech detection accuracy fast achieved the high result. The results show, that the proposed framework give high detection accuracy.

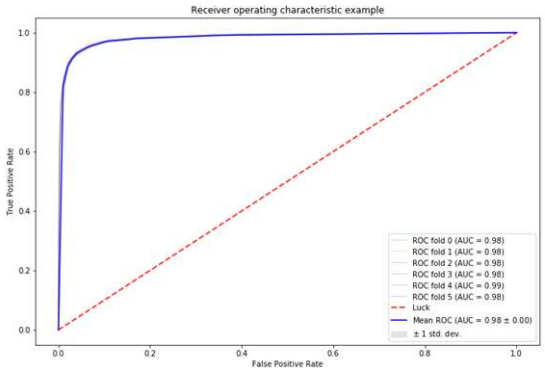

Figure 8: AUC-ROC curve for hate speech detection in social networks

6. CHALLENGES

The inclusion of several layers of complexity makes the already difficult task of automatically distinguishing offensive and/or hostile speech more difficult, and this problem is especially prevalent in social networking sites. There is a direct correlation between some of these issues and the restrictions that are imposed by approaches that are predicated on keywords. It is possible to obfuscate the meaning of words via a number of different means. This may be done on purpose in an attempt to evade the completely automatic moderation of material [60], or it may be an unintentional result of using social networks as a method of information transmission. Either way, it's important to be aware of the potential for either scenario to occur.

In addition, there are a great number of idioms that are not damaging in and of themselves; yet, if you employ them in the appropriate context, they are capable of being offensive [55-57]. However, even in the case of slurs, not only can various slurs have a varying degree of the offense that has been committed [58], but the event may also change based on different eras, as well as different uses of the same term, different users. One example of this is the differential in the use of slurs between members of the in-group and those of the out-group [59-60], which may be considered as an instance of this idea. If this component is neglected, it has the potential to contribute to the inherent bias in hate speech detection corpora, which in turn has the potential to play a role in the inherent prejudice in hate speech detection.

7. CONCLUSION

In conclusion, we were successful in developing a categorization model that distinguishes between hate speech and other types of communication. In addition to this, we developed text characteristics that were included into the training of our classification models. Our request question, which asked if machine learning might be used to identify instances of hate speech on social media sites, was answered by the final classification model that was constructed. It has been recognized, with regard to the body of study that exists, that certain terms are relevant for distinguishing hate speech versus communication that does not promote hatred. Tweets that were labeled with the hate hashtag included statements that were homophobic, racist, and sexist. Although this makes it simpler for us to identify hateful and offensive remarks, it raises the possibility that our algorithm may incorrectly categorize phrases if they do not include any of the standard hate terms. A further conclusion that may be drawn from these data is that hate speech may be aimed at a specific person, a specific group, or it may be employed without any specific target at all. The models that were employed all had scores of accuracy and precision that were somewhat above

average; hence, the performance of the final model was not very impressive. In addition, the hardware device that was employed for this job presented a great deal of difficulty owing to the physical characteristics it had.

Instead of relying just on datasets that are made up of commonly used terms associated with hatred, researchers doing future work should make use of databases that include hostile texts and settings. It is expected that the efforts would be able to construct models with improved accuracy and functionality. Additionally, they need to investigate the characteristics and reasons behind the individuals who participate in the dissemination of hatred on social media platforms.

REFERENCES:

[1] Abro, S., Shaikh, S., Khand, Z. H., Zafar, A., Khan, S., & Mujtaba, G. (2020). Automatic hate speech detection using machine learning: A comparative study. International Journal of Advanced Computer Science and Applications, 11(8).

[2] Mullah, N. S., & Zainon, W. M. N. W. (2021). Advances in machine learning algorithms for hate speech detection in social media: a review. IEEE Access.

[3] Alfina, I., Mulia, R., Fanany, M. I., & Ekanata, Y. (2017, October). Hate speech detection in the Indonesian language: A dataset and preliminary study. In 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS) (pp. 233-238). IEEE.

[4] Roy, P. K., Tripathy, A. K., Das, T. K., & Gao, X. Z. (2020). A framework for hate speech detection using deep convolutional neural network. IEEE Access, 8, 204951-204962.

[5] Al-Makhadmeh, Z., & Tolba, A. (2020). Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. Computing, 102(2), 501-522.

[6] Sutejo, T. L., & Lestari, D. P. (2018, November). Indonesia hate speech detection using deep learning. In 2018 International Conference on Asian Language Processing (IALP) (pp. 39-43). IEEE.

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

[8] Raufi, B., & Xhaferri, I. (2018, September). Application of machine learning techniques for hate speech detection in mobile applications. In 2018 International Conference on Information Technologies (InfoTech) (pp. 1-4). IEEE.

[9] MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. PloS one, 14(8), e0221152.

[10] Omar, A., Mahmoud, T. M., & Abd-El-Hafeez, T. (2020, April). Comparative performance of machine learning and deep learning algorithms for Arabic hate speech detection in osns. In The International Conference on Artificial Intelligence and Computer Vision (pp. 247-257). Springer, Cham.

[11] Sajjad, M., Zulifqar, F., Khan, M. U. G., & Azeem, M. (2019, August). Hate speech detection using fusion approach. In 2019 International Conference on Applied and Engineering Mathematics (ICAEM) (pp. 251-255). IEEE.

[12] Del Vigna12, F., Cimino23, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17) (pp. 86-95).

[13] Sohn, H., & Lee, H. (2019, November). Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In 2019 International Conference on Data Mining Workshops (ICDMW) (pp. 551-559). IEEE.

[14] Aulia, N., & Budi, I. (2019, April). Hate speech detection on indonesian long text documents using machine learning approach. In Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence (pp. 164-169).

[15] Mohapatra, S. K., Prasad, S., Bebarta, D. K., Das, T. K., Srinivasan, K., & Hu, Y. C. (2021). Automatic Hate Speech Detection in English-Odia Code Mixed Social Media Data Using Machine Learning Techniques. Applied Sciences, 11(18), 8575.

[16] Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018, January). All you need is" love" evading hate speech detection. In Proceedings of the 11th ACM workshop on artificial intelligence and security (pp. 2-12).

[17] Issayev, A., Ortayev, B., Issayev, G., Baurzhan, D., & Gulzhaina, A. (2022). Improving the supervisory competence of future teacher trainers with the help of innovative technologies.

[18] Rohmawati, U. A. N., Sihwi, S. W., & Cahyani, D. E. (2018, November). SEMAR: An interface for Indonesian hate speech detection using machine learning. In 2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) (pp. 646-651). IEEE.

[19] Anand, M., Sahay, K. B., Ahmed, M. A., Sultan, D., Chandan, R. R., & Singh, B. (2022). Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques. Theoretical Computer Science.

[20] Niam, I. M. A., Irawan, B., Setianingsih, C., & Putra, B. P. (2018, December). Hate speech detection using latent semantic analysis (lsa) method based on image. In 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC) (pp. 166-171). IEEE.

[21] Altayeva, A., Omarov, B., Suleimenov, Z., & Im Cho, Y. (2017, June). Application of multi-agent control systems in energy-efficient intelligent building. In 2017 Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS) (pp. 1-5). IEEE.

[22] Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE access, 6, 13825-13835.

[23] Zhou, Y., Yang, Y., Liu, H., Liu, X., & Savage, N. (2020). Deep learning based fusion approach for hate speech detection. IEEE Access, 8, 128923-128929.

[24] Nascimento, G., Carvalho, F., Cunha, A. M. D., Viana, C. R., & Guedes, G. P. (2019, October). Hate speech detection using brazilian imageboards. In Proceedings of the 25th Brazillian Symposium on Multimedia and the Web (pp. 325-328).

[25] Omarov, B., Orazbaev, E., Baimukhanbetov, B., Abusseitov, B., Khudiyarov, G., & Anarbayev, A. (2017). Test battery for comprehensive control in the training system of highly Skilled Wrestlers of Kazakhstan on National wrestling "Kazaksha Kuresi". Man In India, 97(11), 453-462.

[26] Defersha, N. B., & Tune, K. K. (2021). Detection of hate speech text in afan oromo social media using machine learning approach. Indian J Sci Technol, 14(31), 2567-78.

[27] Plaza-del-Arco, F. M., Molina-González, M. D., Urena-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. Expert Systems with Applications, 166, 114120.

[28] Doskarayev, B., & Kulbayev, A. (2017). Sport as an important factor of strengthening tolerance (The case of Kazakhstan). Revista ESPACIOS, 38(46).

[29] Omarov, B., Altayeva, A., Turganbayeva, A., Abdulkarimova, G., Gusmanova, F., Sarbasova, A., ... & Omarov, N. (2018, November). Agent based modeling of smart grids in smart cities. In International Conference on Electronic Governance and Open Society: Challenges in Eurasia (pp. 3-13). Springer, Cham.

[30] Sandaruwan, H. M. S. T., Lorensuhewa, S. A. S., & Kalyani, M. A. L. (2019, September). Sinhala hate speech detection in social media using text mining and machine learning. In 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer) (Vol. 250, pp. 1-8). IEEE.

[31] Koushik, G., Rajeswari, K., & Muthusamy, S. K. (2019, September). Automated hate speech detection on Twitter. In 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA) (pp. 1-4). IEEE.

[32] Aljarah, I., Habib, M., Hijazi, N., Faris, H., Qaddoura, R., Hammo, B., ... & Alfawareh, M. (2021). Intelligent detection of hate speech in Arabic social network: A machine learning approach. Journal of Information Science, 47(4), 483-501.

[33] Onalbek, Z. K., Omarov, B. S., Berkimbayev, K. M., Mukhamedzhanov, B. K., Usenbek, R. R., Kendzhaeva, B. B., & Mukhamedzhanova, M. Z. (2013). Forming of professional competence of future tyeacher-trainers as a factor of increasing the quality. Middle East Journal of Scientific Research, 15(9), 1272-1276.

[34] Putri, T. T. A., Sriadhi, S., Sari, R. D., Rahmadani, R., & Hutahaean, H. D. (2020, April). A comparison of classification algorithms for hate speech detection. In Iop conference series: Materials science and engineering (Vol. 830, No. 3, p. 032006). IOP Publishing.

[35] Setyadi, N. A., Nasrun, M., & Setianingsih, C. (2018, December). Text analysis for hate speech detection using backpropagation neural network. In 2018 International Conference on Control,

Electronics, Renewable Energy and Communications (ICCEREC) (pp. 159-165). IEEE.

[36] Gaydhani, A., Doma, V., Kendre, S., & Bhagwat, L. (2018). Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. arXiv preprint arXiv:1809.08651.

[37] Aldjanabi, W., Dahou, A., Al-qaness, M. A., Elaziz, M. A., Helmi, A. M., & Damaševičius, R. (2021, October). Arabic Offensive and Hate Speech Detection Using a Cross-Corpora Multi-Task Learning Model. In Informatics (Vol. 8, No. 4, p. 69). MDPI.

[38] Rajput, G., Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2021, December). Hate speech detection using static BERT embeddings. In International Conference on Big Data Analytics (pp. 67-77). Springer, Cham.

[39] Florio, K., Basile, V., Polignano, M., Basile, P., & Patti, V. (2020). Time of your hate: The challenge of time in hate speech detection on social media. Applied Sciences, 10(12), 4180.

[40] Sakhipov, A., & Yermaganbetova, M. (2022). An educational portal with elements of blockchain technology in higher education institutions of Kazakhstan: opportunities and benefits. Global Journal of Engineering Education, 24(2).

[41] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In Proceedings of the international AAAI conference on web and social media (Vol. 11, No. 1, pp. 512-515).

[42] Ibrohim, M. O., & Budi, I. (2019, August). Multi-label hate speech and abusive language detection in Indonesian twitter. In Proceedings of the Third Workshop on Abusive Language Online (pp. 46-57).

[43] Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. PeerJ Computer Science, 7, e598.

[44] De la Pena Sarracén, G. L., Pons, R. G., Cuza, C. E. M., & Rosso, P. (2018). Hate speech detection using attention-based lstm. EVALITA evaluation of NLP and speech tools for Italian, 12, 235.

[45] Ali, R., Farooq, U., Arshad, U., Shahzad, W., & Beg, M. O. (2022). Hate speech detection on Twitter using transfer learning. Computer Speech & Language, 74, 101365.

[46] Mossie, Z., & Wang, J. H. (2018). Social network hate speech detection for Amharic language. Computer Science & Information Technology, 41-55.

[47] Sultanovich, O. B., Ergeshovich, S. E., Duisenbekovich, O. E., Balabekovna, K. B., Nagashbek, K. Z., & Nurlakovich, K. A. (2016). National Sports in the Sphere of Physical Culture as a Means of Forming Professional Competence of Future Coach Instructors. Indian Journal of Science and Technology, 9(5), 87605-87605.

[48] Fauzi, M. A., & Yuniarti, A. (2018). Ensemble method for indonesian twitter hate speech detection. Indonesian Journal of Electrical Engineering and Computer Science, 11(1), 294-299.

[49] Boishakhi, F. T., Shill, P. C., & Alam, M. G. R. (2021, December). Multi-modal Hate Speech Detection using Machine Learning. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 4496-4499). IEEE.

[50] Ali, M. Z., Rauf, S., Javed, K., & Hussain, S. (2021). Improving hate speech detection of Urdu tweets using sentiment analysis. IEEE Access, 9, 84296-84305.

[51] Jemima, P. P., Majumder, B. R., Ghosh, B. K., & Hoda, F. (2022, June). Hate Speech Detection using Machine Learning. In 2022 7th International Conference on Communication and Electronics Systems (ICCES) (pp. 1274-1277). IEEE.

[51] Arango, A., Pérez, J., & Poblete, B. (2020). Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). Information Systems, 101584.

[52] Rizoiu, M. A., Wang, T., Ferraro, G., & Suominen, H. (2019). Transfer learning for hate speech detection in social media. arXiv preprint arXiv:1906.03829.

[53] Bosco, C., Felice, D. O., Poletto, F., Sanguinetti, M., & Maurizio, T. (2018). Overview of the evalita 2018 hate speech detection task. In EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (Vol. 2263, pp. 1-9). CEUR.

[54] Nayel, H. A., & Shashirekha, H. L. (2019, December). DEEP at HASOC2019: A Machine Learning Framework for Hate Speech and Offensive Language Detection. In FIRE (Working Notes) (pp. 336-343).

[55] Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018, June). A dataset of Hindi-English code-mixed social media text for hate speech detection. In Proceedings of the second

workshop on computational modeling of people's opinions, personality, and emotions in social media (pp. 36-41).

[56] Romim, N., Ahmed, M., Talukder, H., & Islam, S. (2021). Hate speech detection in the bengali language: A dataset and its baseline evaluation. In Proceedings of International Joint Conference on Advances in Computational Intelligence (pp. 457-468). Springer, Singapore.

[57] Mossie, Z., & Wang, J. H. (2020). Vulnerable community identification using hate speech detection on social media. Information Processing & Management, 57(3), 102087.

[58] B.N. Singh, Bhim Singh, Ambrish Chandra, and Kamal Al-Haddad, "Digital Implementation of an Advanced Static VAR Compensator for Voltage Profile Improvement, Power Factor Correction and Balancing of Unbalanced Reactive Loads", *Electric Power Energy Research*, Vol. 54, No. 2, 2000, pp. 101-111.

[59] William, P., Gade, R., esh Chaudhari, R., Pawar, A. B., & Jawale, M. A. (2022, April). Machine Learning based Automatic Hate Speech Recognition System. In 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) (pp. 315-318). IEEE.

[60] Abderrouaf, C., & Oussalah, M. (2019, December). On online hate speech detection. effects of negated data construction. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 5595-5602). IEEE.