Journal of Theoretical and Applied Information Technology
15th May 2021. Vol.99. No 9

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

AUTOMATIC OFFENSIVE LANGUAGE DETECTION IN ONLINE USER GENERATED CONTENTS

[1]AIGERIM TOKTAROVA, [2,3]GULBAKHRAM BEISSENOVA, [4,5]MARAT NURTAS, [2]PERNEKUL KOZHABEKOVA, [6]ZHANAR AZHIBEKOVA, [2]ZLIKHA MAKHANOVA, [2]BIBIGUL TULEGENOVA, [2]NAZIRA RAKHYMBEK, [5,7,8]ZHARASBEK BAISHEMIROV

[1]Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan
[2]M.Auezov South Kazakhstan University, Shymkent, Kazakhstan
[3]University of friendship of people's academician A. Kuatbekov, Shymkent,Kazakhstan
[4]International Information Technology University, Almaty, Kazakhstan
[5]Kazakh-Britain Technical University, Almaty, Kazakhstan
[6]Asfendiyarov Kazakh National Medical University
[7]Abai Kazakh National Pedagogical University, Almaty, Kazakhstan
[8]RSE Institute of Information and Computational Technology CS MES RK, Almaty, Kazakhstan
[9]Tenaga National University, Kuala Lumpur, Malaysia

E-mail: aigerim.toktarova@ayu.edu.kz, Toktar.aigerim@list.ru, gulbakhrambeissenova@gmail.com

ABSTRACT

Profanity on the pages of social networking networks is growing the number of problems social networking has. Easy and automated steps are necessary to control the amount of content that is generated on a daily basis. There is a significant research question concerning language instruction rather than the implementation methods.

We are creating a dataset of internet users in Kazakhstan where they use social networks and the media to share their opinions. According to this report, it is the first time anybody has ever done a study focused on complaints from multiple social networks. We have classified our index in a variety of respects, one of which is to use derogatory terms.

Furthermore, our results will not only explore the roots of offensive language, but will also present concepts that help in differentiating such types of offensive language, such as offensive language and cyberbullying. We use machine learning approaches to access the data sets we can use for the automated study of offensive language on social media. The results show that recognizing offensive language on social networks is a task that can be solved automatically and produces excellent results.

Keywords: *Offensive language, Hate Speech, Machine Learning, Detection, Classification, Natural Language Processing, Social Networks, Social Media.*

1. INTRODUCTION

The progress of science and technology is currently accompanied by the intensive introduction of new information technologies in many areas of human activity. The development of the Internet leads to an uncontrolled exponential growth in the amount of various information, most of which is presented in text form [1].

With the development of the capabilities of telecommunications systems, scientific and technological progress generates both positive consequences and negative results, which subsequently awakens the phenomenon of social deviation, in particular, in Web content [2]. The information space of the Internet contains a lot of resources that carry information of various types,including destructive ones [3].

Destructive data is data that in a destructive and harmful manner impacts a mainstream audience's consciousness and actions. In general, social networks and the Internet are a "convenient" environment for the organization of disruptive knowledge and psychological control, including for the exploitation of persons, social classes and culture as a whole [4].

In fact, the task of identifying destructive information can be reduced to the task of classification. This task allows you to assign text information to the class of texts containing

destructive information or to the class of texts that do not contain destructive information. It is possible to assign information to one of the two classes by searching the text for so-called destructive content indicators, which include profanity, calls for mass riots, terrorism, drug trafficking, and others [5-7].

The modern global network contains a huge amount of heterogeneous information, which in its content can be considered as malicious. The discovery of sources is an important task, since their dissemination and use can lead to serious negative consequences both at the local level, affecting the interests and rights of individuals, and at the global level, which is reflected in international disputes and conflicts [8].

As an example of the detection of malicious information objects (IO), we can cite parental control systems. The role of IO in such systems is the information provided by such Internet services as websites, social networks, online chats, gaming and media portals, and others. In this case, the denial of access to the IO is performed based on the analysis of data flows transmitted from the corresponding Internet resource to the end user. A sign for such a ban may be the presence of malicious information containing, for example, profanity, calls for illegal actions or instructions that promote an unhealthy lifestyle [9-11].

The very task of detecting malicious information can be considered as a task of categorizing IT, in which illegitimate categories are defined in advance. Systems designed to solve this problem can be based both on the manual construction of classification rules, and with the involvement of automatic means of their generation. It is the latter type of such systems that is of the greatest interest to researchers in connection with the constant growth, development and popularization of such a promising scientific direction as machine learning [12].

The article discusses the issue of improving the effectiveness of detecting malicious IO by the example of the problem of classifying web pages using various machine learning methods and combining them.

2. RELATED WORKS

The task of detecting malicious information on the Internet can be reduced to the classification of web pages, in which a number of categories are pre-determined by the system administrator as containing illegitimate content. In this area, there are many works devoted to the construction of both expert systems and fully automatic systems [13].

The CONSTRUE system presented in [14] is based on production rules created manually by an expert operator. This system is designed to classify economic and financial news and correlate the analyzed text to one of 674 categories. The classification accuracy for the CONSTRUE system is more than 90 %. The disadvantage of such a system is that its maintenance in a consistent state requires the regular involvement of specialists who perform the addition and correction of production rules.

The approach to content categorization with automatic generation of classification rules is considered by researchers in [15]. Their proposed rule format is disjunctive normal form (DNF). The algorithm for generating rules is based on the sequential replacement of one of the conjuncts and the further addition of a new conjunct until one hundred percent coverage of the training sample is built (i.e., such a set of rules that will provide an error-free classification of training elements). This algorithm performs a heuristic search for such rules: the algorithm does not provide finding the minimum number of DNF conjuncts. In addition, unlike a decision tree, conjuncts united by a single rule using this algorithm are not mutually exclusive.

Predicates reflecting the signs were used as elementary conjuncts (atoms).:

1) the occurrence of a certain word (or phrase) from a local dictionary (a set of words containing concepts specific to one category) in the analyzed text;

2) exceeding the frequency of occurrence of a certain expression within the analyzed text by the specified threshold value.

The proposed approach allows you to keep the presentation of the rules in a format that is convenient for analysis by experts. At the same time, with the described method of generating rules, the generalizing ability of the system and the ability to process noisy data are lost.

The authors of the article [16] propose to accept the analyzed document as an array of real-valued coefficients, which represent the relative and absolute frequencies of occurrence of certain words in the classified text. Among these coefficients were highlighted [17-19]:

- the frequency of the word (TF, from the English Term Frequency);
- Inverse document frequency (IDF, from the English Inverse Document Frequency);

- the importance of the word (TD, from the English Term Discrimination), where TD = TF × IDF; and some others.

In [20], an approach is presented that allows us to assign each word its integral weight, including the probability of the appearance of this word, both within a certain category, and within the entire collection of documents, and taking into account the other categories.

A comparison of two machine learning methods, namely the Bayesian classifier and the decision tree, in the context of the text categorization problem is performed in [21]. The authors of this article emphasize that the decision tree shows the best performance on large training data sets, and the Bayesian classifier shows the best performance on smaller data sets. Moreover, the Bayesian classifier with the increasing number of treatable symptoms observed the situation of overfitting (on the test set performance of the classifier is reduced), and decision tree under these conditions, and sufficient training samples increases the efficacy of the classification.

The applicability of another popular machine learning method, namely, the support vector machine (MOV), to the problem of text classification is investigated [22], where the author highlights the ability of MOV to learn on both high-dimensional and sparse feature vectors. The solution to the problem of text classification in most cases has the form of linearly separable areas, for the separation of which the MOU can be used.

The article [23] describes two types of convolutional neural networks: direct signal propagation and with the transformation of the" bag-of-words " on the convolutional layer. As a result of experiments, the authors found that the first type of neural network demonstrates greater performance in terms of classification indicators compared to the second type of neural network.

In [24], a method is presented for extracting features within the text categorization task. The proposed modification of the genetic algorithm, as shown by experiments, allows us to achieve a more compact representation of the training vectors in terms of their dimension and improve the quality of classification of the analyzed text.

A common limitation for the above-mentioned works on the application of machine learning methods to the problem being solved is the use of single-component classifiers, which makes it impossible to train the model in parts and, in turn, makes it difficult to parallelize this process.

The analysis of works in this subject area shows the relevance of the topic under consideration. At the same time, despite their diversity, the task of developing a methodology designed to detect malicious IO on the Internet and combining machine learning methods and their combination remains a high priority in the research community.

3. RESEARCH METHODS

Traditional methods of content analysis used to study small volumes of text provide the researcher with a high level of control in solving analytical problems [25-27]. The methodological techniques presented in this paper are based on automated analysis of large text arrays of information on the Internet. The grouping and interpretation of the research results is based on the method of sound theory, which is consistent with the theory of relations presented above.

Among the methods for assessing the emotional background and extracting entities from tweets, the Dostoevsky neural network model and stemming with word root search were selected (as having the highest accuracy). In the presented models are used as methods of machine learning, algorithmic and classical methods of analysis. The developer of software tools is M. A. Kitov. The accuracy of the classifications was assessed using manual marking of a part of the data [28-30].

User messages were extracted from open sources on the Internet. The study collected and processed 1 377 879 tweets. Content analysis of messages was performed according to the following algorithm: data cleaning from irrelevant content and spam; text splitting into tokens and morphological analysis; text tonality evaluation [31].

As the results of the statistical analysis of words, the following values were used for each desired entity: the number of mentions; the number of positive mentions; the number of negative mentions; the overall average emotional intensity of the utterance. In this paper, only the last position is presented – the average emotional intensity of words and utterances [32].

4. DESCRIPTION OF THE CORPORA

Twitter was chosen as the platform for collecting the corpus because:

1. First, twitter users often express a subjective, emotionally charged opinion about something;
2. To express emotions, users use live, conversational language, which may contain slang

and profanity that enhance the tone of the messages;

3. When writing messages, users may make widespread mistakes that are corrected by the editors of news publications, but which must be taken into account when classifying texts from the Internet (for example, blogs or product review sites) by tone.

Using the twitter Streaming API [33], a text collection consisting of about 15 million short messages was collected, on the basis of which, using the method [34] and the filtering proposed by the author [35], a balanced corpus was formed, consisting of the following collections:

* collection of positive messages 114,991 entries;

* collection of negative messages 111,923 entries;

5. IDENTIFYING FEATURES FOR THE TASK OF TEXT CLASSIFICATION BY TONALITY

All documents from the training and test samples are dimensional feature vectors. Thus, the document is defined as a vector d = (w 1, w 2,..., w / V|), where V is the set of all unique unigrams from the training sample, and w i is the weight of the i-th unigram. For weighing unigrams in this paper, the following weight scheme TF-IDF is used, which is calculated by the formula:

$$tf.idf = tfx \log \frac{T}{T(t_i)} \quad (1)$$

hereafter, tf is the frequency of occurrence of the term in the collection (positive or negative tweets). T is the total number of messages in the positive and negative collections, and T(ti) is the number of messages in the positive and negative collections containing the term.

6. ANALYSIS OF APPROACHES TO THE SOLUTION

To solve this problem, the existing search methods used to find specific words in the text were analyzed using the example of searching and identifying profanity among destructive information [36-38].

Often, solutions aimed at organizing the search and identification of destructive information (for example, profanity) involve the analysis of search methods based on expert information processing. The result of the use of expert methods is the creation of so-called "black lists" of Internet sites containing undesirable information prohibited for distribution. The introduction and formation of such lists are aimed at the mandatory blocking of illegal Web pages, but due to the large amount of such content and the rather long time spent using only the human factor, the use of this method is not appropriate. Therefore, in the task of identifying destructive content, automated methods are used to improve the quality of the process execution [39]. These include the so-called thematic search (by dictionary) and intelligent data processing methods, the advantages of which are shown in Figure 1. In addition, due to the need for constant replenishment of the inventory, the use of thematic search in a "pure" form is impractical. In addition, when using simple methods of searching for destructive words, the system requires constant interaction with an expert in order to analyze web content most effectively. Such methods are not able to adapt the system to objective changes in the subject (problem) area of the system functioning, do not have the ability to independently acquire new knowledge and automatically replenish it, i.e. self-study [40].

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

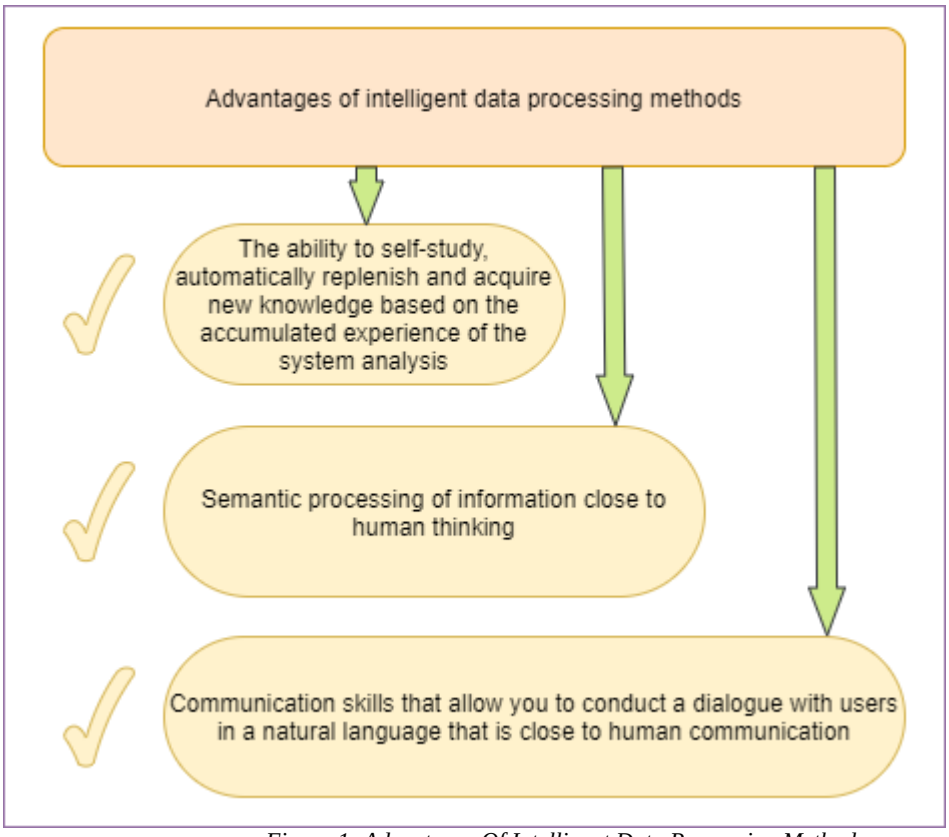

Figure 1: Advantages Of Intelligent Data Processing Methods

Thus, by endowing the developed system with properties and skills close to human factors, the search for dangerous content can show the most effective possibility of detecting destructive meaning in text information.

7. DATA COLLECTION METHODS

The article proposes a variant of the formal model of the implementation of this process within the framework of the engineering approach, according to which, in order to reproduce a linguistic object (phenomenon) using a computer, it is necessary to create a database of formalized data/knowledge describing this object (phenomenon), and build an algorithm for its functioning on its basis [41]. The database of the prototype of the above system was formed on the basis of a study of 215 English-language accounts of the social network Twitter with a total volume of 200,000 tweets, of which more than 4,000 tweets were analyzed in detail. According to the results of the analysis, 583 English-language tweets containing signs of a destructive "cyberbullying"strategy were identified [42-44]. Electronic verbal harassment was most often found in the posts of teenagers aged 11-17 years, as well as young people aged 18-35 years. Adolescent cyberbullying was carried out, as a rule, by groups of individuals, and electronic harassment in the youth environment was implemented according to the "one buller – one victim"scheme [45].

At the first stage of the database formation, we took into account the fact that participants in network communication often use forms of expression of thoughts that are far from traditional lexical norms. So, in the texts of tweets, there are abbreviations that reflect, in fact, the user's reaction or emotions (OMG-Oh my God, WTF-what the fuck), various abbreviations (u-you, Bout-about), intentional spelling mistakes, typos, etc. From a linguistic point of view, the lexical form of their representation is far from traditional, which makes it almost impossible to automatically analyze them until they are normalized. Therefore, a thorough analysis of the empirical material allowed us to identify all the words that are subject to lexical normalization in the array of tweets. For example, such tweet units as yoself, wtf, bae, bestest, dunno, sus, b4, etc.were selected as database elements and the correct spellings were recorded [46].

At the second stage of creating the database, eight classes of verbal markers were identified that clearly indicate the presence of electronic bullying in the messages of Twitter users and are therefore important for developing a procedure for automatically determining the means of its implementation [47-48]. Further, from the array of English-language tweets, specific lexical units that were part of previously defined classes were identified. The identification of cyberbullying markers was based on the knowledge of the English language, as well as taking into account a number of dictionary and reference sources, for example, The Oxford Advanced Learner's Dictionary, Merriam-Webster Dictionary, Collins Dictionary, etc. The verbal markers are:

1) taboo and obscene vocabulary (one-component, two-component, three-component and multi-component) - is used as a means of pronounced verbal aggression with the aim of social discrediting the victim, manipulating her by emphasizing her own superiority. This layer of vocabulary includes the words bastard, freak, moron, dirty cow, pussy eater, eat my shorts, dirty son-of-a-bitch, worthless excuse for a human being, etc., which clearly indicate the presence of electronic harassment in tweets;

2) words that name concepts related to the intimate life of a person - often used by bullers in order to hurt the feelings of the victim, humiliate and insult her. In many languages, the names of the genitals and types of sexual acts belong to taboo vocabulary and are often used to express a high degree of aggression of the addressee. This group includes the words ass, choad, penis, tities, vagina, etc.;

3) words referring to concepts related to sexual orientation and sexism - often, in order to achieve the above goals, victims are attributed and / or emphasize their non-traditional sexual orientation or sexist affiliation, using the words gay, lesbiana, pedophile/pedo, sexist, trans, etc., and the addressee is not necessarily a representative of a sexual minority;

4) words expressing a wish for evil and death – the use of such lexical units is due to buller's strong personal dislike and even hatred of the victim. The use of such lexical units as commit a suicide, die in inferno, slit your wrists, stop breathing, etc. shows that buller wants to emphasize his superiority over the victim, to prove his greatness;

5) words that express humiliation and insult of a person-most often used when electronic harassment occurs according to the scheme "several bullers – one victim". The lexical units cocky, dull, foul, ignorant, stupid, etc.are aimed at undermining the self-esteem and self-esteem of the victim through her constant criticism and understatement of her abilities. Thus Buller seeks to demonstrate his power over her;

6) words that name concepts related to nationality and racism, such as blackface, hack, nigger, racist, sociopath, etc. - allow the aggressor to inflict not only moral, but also social harm on the victim;

7) words that name animals - used as an insult, such lexical units pass into the category of invective words on a par with traditional profanity. The use of such animal comparisons as cow, donkey, pig, pooch, sheep, worm, etc. equates the characteristics of the victim with the characteristics of animals, which belittles its status in society and increases the buller's self-esteem by belittling it;

8) words that name people with physical and mental disabilities – the main purpose of using this vocabulary is to humiliate the honor and dignity of the victim, reduce her social attractiveness, destroy her as a person. Buller often uses language that indicates the victim's limited mental capacity when the victim is smarter than him, and Buller wants to prove otherwise. This group includes the words deaf, cripple, imbecile, moronic, weak-minded, etc.

8. DATA COLLECTION

8.1. Development of a corpus

It is important to collect text data for the study of what is spoken, feeling, or feels in texts. Unfortunately, it is hard to find suitable texts for this assignment when it comes to dealing with offensive language in a specific language. Many social networks and especially collections such as digital libraries, digital archives, and research databases, are all sharing catalogue and need to be filtered to specific categories of research topics. Our decision to create our own corpus of offensive language texts stemmed from the issues with our previous corpus, including its complexity and the fact that there is no separate subfield of Kazakhstan's language for offensive language. The corpus consists of two elements: text records containing offensive language, and open social network sites on which there are no offensive language posts. In order to collect data, we used parsing technologies. In Figure 2, you can see a basic data collection design.

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

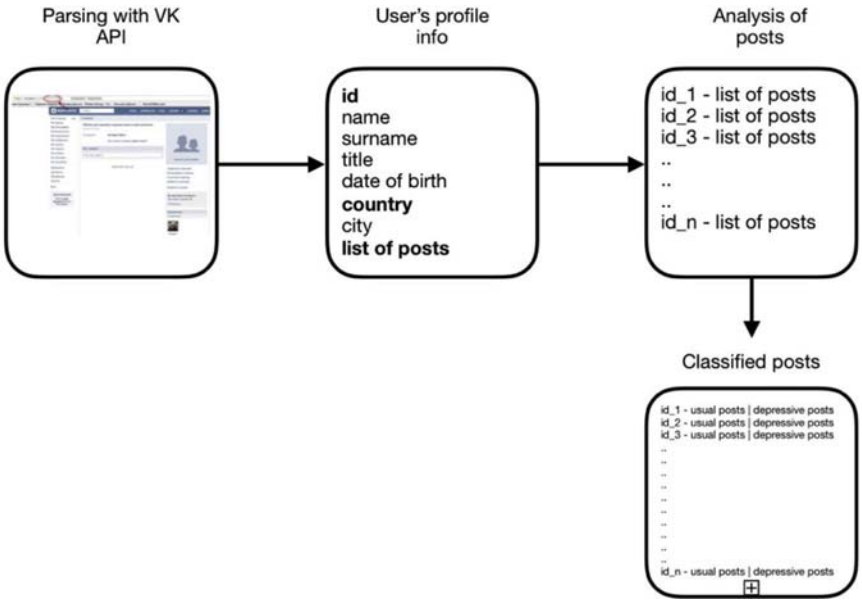

Figure 2: Data Collection

In the data analysis portion of the questionnaire, it was discovered that data from different data sources are needed in order to conduct the analysis (Vkontakte [49], Instagram [50], news portals). Now that the APIs have been made available to us, we learned on our own that they are perfect for collecting up-to-date knowledge. VK API methods used to collect the data from Vkontakte social network [51].

A parser for the research was created by using two separate parser forms. One simple and visual example of a practical and an equal financial system is this example of an operation of all forms of schemes as can be seen in Figure 2.

We had identified the code in the document by using this document's headers, and that helped us to download the HTML for this page. You could also get the responses one page at a time, which would cause the request and answer to be smoother, and less extraneous pieces of details would be available. Then, by changing the configuration of the site, i.e., by using various objects, we can obtain different data which will participate in the studies. When all the material is collected, it would be placed in a convenient format for subsequent retrieval by a vast number of users.

```
f = open("demofile2.txt", "a") # opening the file for editing
f.write("---") #record data
f.close() # close
```

All the measurements required for the data have been taken, and the workflow is now able to input the mathematical model. While this situation is quite identical to the previous one, the coding that is missed are meaningless and unimportant in the eyes of the developers. Because the computer can retrieve information from several sites at the same time, it will remain updated to guarantee that the information is as reliable as possible. Customization of the upgrade frequency is authorized without regard to the discrepancies between updates, at whichever intervals you desire. Our scenario specifies that the script adjusts the data every 15 minutes, meaning that all of the system's data will be available at any stage, and updates itself every 15 minutes.

8.2. Data Preprocessing

For the task at hand, we used a parser that collected two different datasets of texts—one for hateful texts and the other for optimistic texts—and applied this data to the outcome. The vectors in Figure 3 are schematically represented, as shown in the bottom of the figure, using the corpus. As can be seen on the left, blue-labeled words here signify neutral or unproblematic material, while green-labeled meanings imply content of hateful or insulting meaning.

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

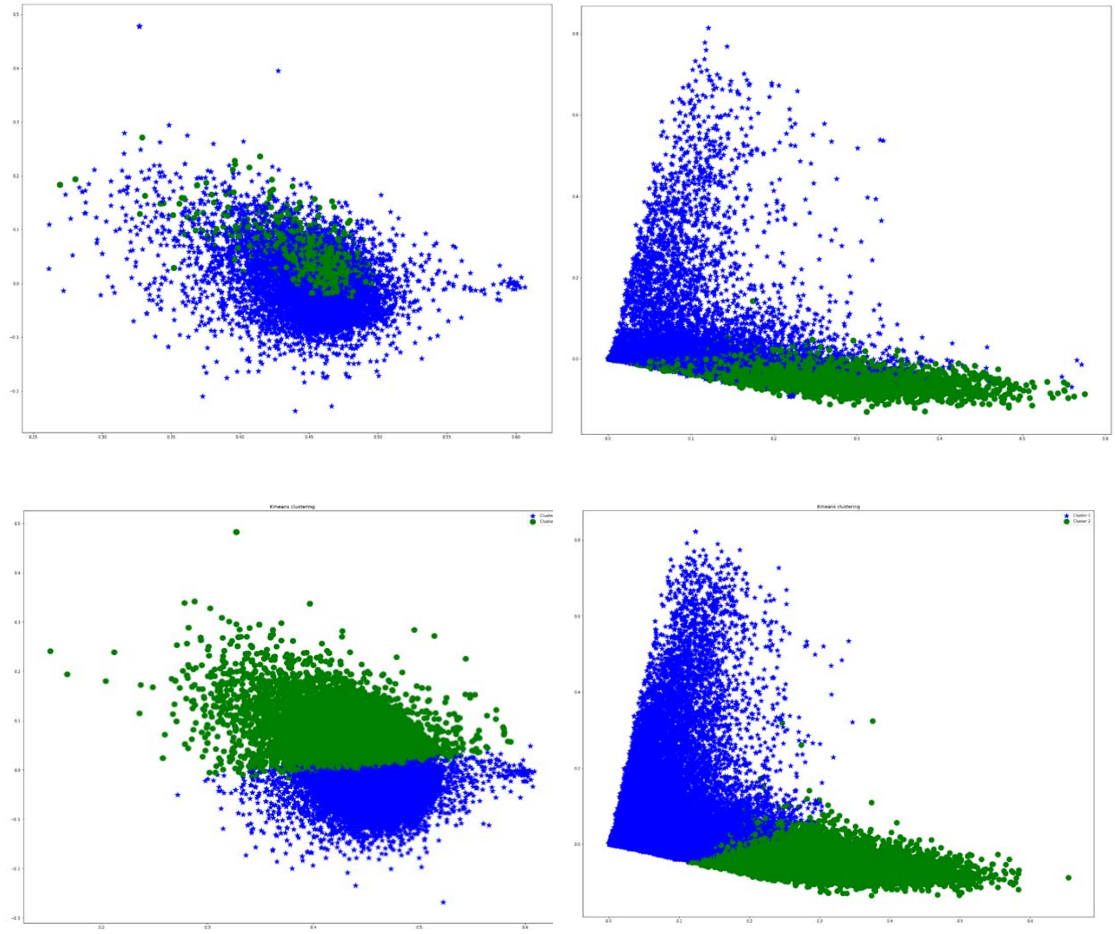

Figure 3: Graphical Representation Of Word2vec Vectors

These texts include hateful expression, which is not present in many other sites, and as such are categorized as blog posts that contain significant amounts of balanced content. When you see Figure 4, think of it as the markings on the shaped corpus. Black text on a white background appears to project objective texts, whereas red text on a red background is widely used to illustrate text that uses inflammatory language.

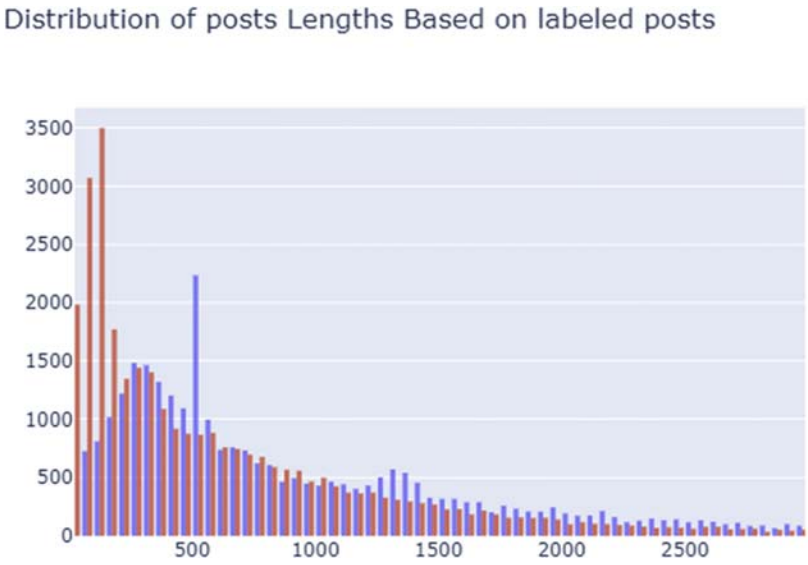

Figure 4: Distribution Of Texts By Labels (Red Color Is The Texts That Contain Offensive Language, Blue Color Is Neutral Texts)

9. EXPERIMENT RESULTS

In order to determine the success of the program, we depended on accuracy, recall, F1, and continuity. Formulas of each expression were given in expressions (2)-(5).

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2\,precision \cdot recall}{precision + recall} \tag{4}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{5}$$

The acronym TP denotes true positive values, and TN denotes true negative values. Meanwhile, FP denotes false positive values, and FN denotes false negative values.

A Receiver Operational Characteristic (ROC) is usually used in studies of binary classification to determine the amount of the classifier's classification precision. To find a binary classification gold standard, you must first provide a system of binarization to be able to submit the output data in. A simulation data collection is used to produce the outcome for each indicator, with an individual model for each of them. Yet each indicator is considered as binary predictions.

9.1 Classification Results

In order to properly classify internet offensive language, we used different deep learning algorithms. We saw good findings with Support Vector Machine, Random Forest, Naïve Bayes, K Nearest Neighbor, and Logistic Regression.

Before we could be absolutely sure that our algorithm was right, we created a receiver performance (ROC) curve by cross-validating the algorithm on our dataset. The ROC curve was applied in this experiment in order to help further grasp performance of a given activity in various threshold environments.

TABLE I. RESULTS OF OFFENSIVE LANGUAGE DETECTION

ML Method	*Accuracy*	*Precision*	*Recall*	*F1 score*
SVM	0.5062	0.6648	0.6647	0.8123
Decision Tree	0.8599	0.7017	0.7059	0.8068
Random Forest	0.7018	0.7268	0.6547	0.7697
KNN	0.8517	0.7305	0.7207	0.4892
Naïve Bayes	0.8024	0.7158	0.7728	0.7681

ML Method	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.8105	0.7325	0.4892	0.7104

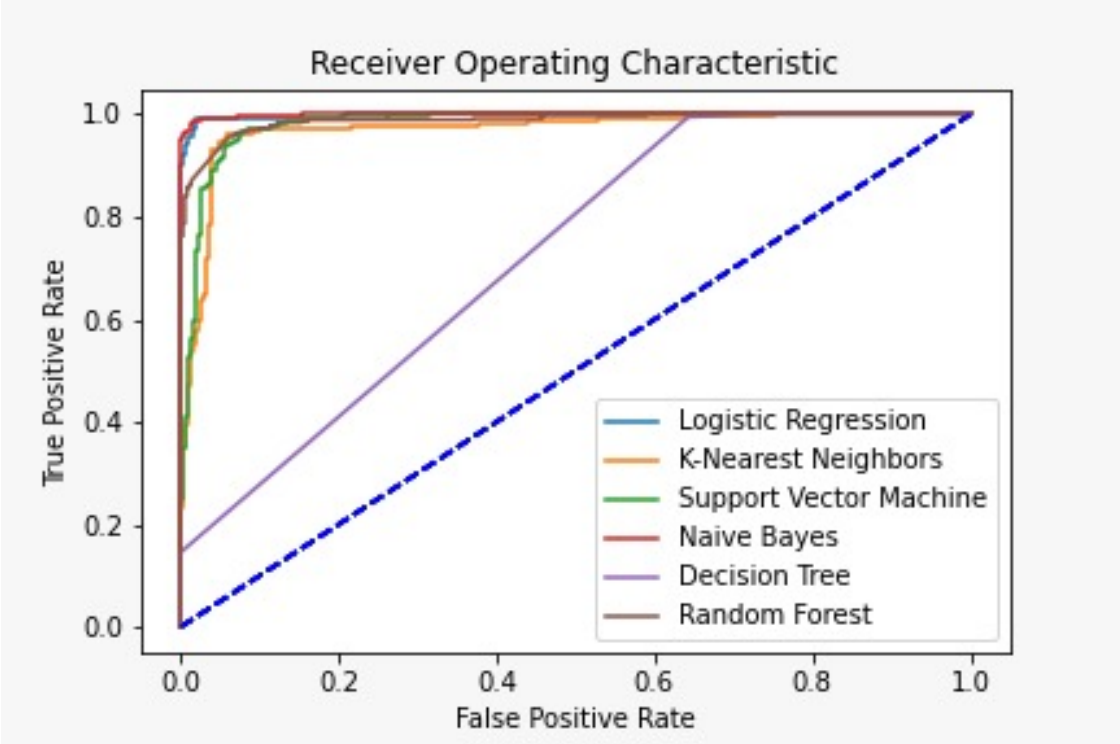

Figure 5: ROC Curves Result

The "steepness" of the ROC curve is significant, since it is a calculation of how far the X-axis curve needs to be broadened to increase the true positive outcome speed.

In Figure 5, toggling different extreme predictions on different train and test datasets built from re-calculations of K-fold validation is seen. After you stop taking each of these curves, when you do the average curve of and subset of the training range, you can see the variance of the curve and understand the variations of the curves. This is an indicator that how the classifier's performance changes depending on the training data and how the classifier's output differs depending on the amount of K-fold cross-validation iterations.

The graph appears to be smooth, and it appears as if the algorithm is competent at looking for distress, and has well-trained look for suicidal posts.

10 .CONCLUSION

Thus, in ensuring the information security of the Internet information space, a significant role is played by analyzing Web content to identify the presence or absence of a destructive orientation. Different approaches should be used to search for different types of destructive information. A possible solution to this problem is the joint use of several types of intelligent data processing methods, which will lead to further improvement of the efficiency of the system under development.

The use of the proposed combined approach is a promising and complementary method of identifying destructive information. The use of this method allows you to reduce the amount of calculations by combining them, increase the visibility and interconnectedness of the results of the analysis of particular problems, and thereby increase the efficiency of the analysis of dangerous content by increasing the validity and efficiency of management decisions.

At the same time, the feature of the proposed procedure is the ability of the system to automatically acquire new knowledge, which consists in self-replenishment of the dictionary. As a result of the research, the material was obtained, the analysis of which allowed us to conclude that the developed method can serve as a basis for the

implementation of a complex automated system for identifying destructive information.

REFERENCES:

[1] Cohen-Almagor, R. (2018). Taking North American white supremacist groups seriously: The scope and the challenge of offensive language on the Internet. International journal of crime, justice, and social democracy, 7(2), 38-57.

[2] Chetty, N., & Alathur, S. (2018). Offensive language review in the context of online social networks. Aggression and violent behavior, 40, 108-118.

[3] Udoh-Oshin, G. (2017). Offensive language on the Internet: Crime or Free Speech?.

[4] Mukeredzi, T. (2017). Uproar over internet shutdowns: Governments cite incitements to violence, exam cheating and offensive language. Journal of Pan African Studies, 10(10), 7.

[5] Alkiviadou, N. (2019). Offensive language on social media networks: towards a regulatory framework?. Information & Communications Technology Law, 28(1), 19-35.

[6] Omarov, B., Omarov, B., Issayev, A., Anarbayev, A., Akhmetov, B., Yessirkepov, Z., & Sabdenbekov, Y. (2020, November). Ensuring Comfort Microclimate for Sportsmen in Sport Halls: Comfort Temperature Case Study. In International Conference on Computational Collective Intelligence (pp. 626-637). Springer, Cham.

[7] Balica, R. (2017). The criminalization of online offensive language: It's complicated. Contemporary Readings in Law and Social Justice, 9(2), 184-190.

[8] Rieger, D., Schmitt, J. B., & Frischlich, L. (2018). Hate and counter-voices in the Internet: Introduction to the special issue. SCM Studies in Communication and Media, 7(4), 459-472.

[9] Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the reliability of offensive language annotations: The case of the european refugee crisis. arXiv preprint arXiv:1701.08118.

[10] Omarov, B., ANARBAYEV, A., TURYSKULOV, U., ORAZBAYEV, E., ERDENOV, M., IBRAYEV, A., & KENDZHAEVA, B. (2020). Fuzzy-PID based self-adjusted indoor temperature control for ensuring thermal comfort in sport complexes. J. Theor. Appl. Inf. Technol, 98(11).

[11] de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Offensive language dataset from a white supremacy forum. arXiv preprint arXiv:1809.04444.

[12] Pohjonen, M., & Udupa, S. (2017). Extreme speech online: An anthropological critique of offensive language debates. International Journal of Communication, 11, 19.

[13] Jakubowicz, A. (2018, January). Algorithms of hate: How the Internet facilitates the spread of racism and how public policy might help stem the impact. In Journal and Proceedings of the Royal Society of New South Wales.

Bortone, R., & Cerquozzi, F. (2017). L'offensive language al tempo di internet. Aggiornamenti sociali, 818, 827.

[14] Waltman, M. S., & Mattheis, A. A. (2017). Understanding offensive language. In Oxford Research Encyclopedia of Communication.

[15] Kim, K. H., Cho, Y. H., & Bae, J. A. (2020). Exploratory Study on Countering Internet Offensive language: Focusing on Case Study of Exposure to Internet Offensive language and Experts' in-depth Interview. The Journal of the Korea Contents Association, 20(2), 499-510.

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

[17] Barron, J. A. (2019). Internet Access, Offensive language and the First Amendment. First Amend. L. Rev., 18, 1.

[18] Omarov, B., Altayeva, A., & Im Cho, Y. (2017, June). Smart building climate control considering indoor and outdoor parameters. In IFIP International Conference on Computer Information Systems and Industrial Management (pp. 412-422). Springer, Cham.

[19] Claussen, V. (2018). Fighting offensive language and fake news. The Network Enforcement Act (NetzDG) in Germany in the context of European legislation. Rivista Di Diritto Dei Media, 3, 1-27.

[20] Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). Offensive language: A systematized review. Sage Open, 10(4), 2158244020973022.

[21] Celik, S. (2018). Tertiary-level internet users' opinions and perceptions of cyberhate. Information Technology & People.

[22] Guiora, A. N. (2018). Inciting terrorism on the internet: the limits of tolerating intolerance. In Incitement to Terrorism (pp. 135-149). Brill Nijhoff.

[23] Bilewicz, M., & Soral, W. (2020). Offensive language epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. Political Psychology, 41, 3-33.

[24] Niam, I. M. A., Irawan, B., Setianingsih, C., & Putra, B. P. (2018, December). Offensive language 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.

[25] Omarov, B., Baisholanova, K., Abdrakhmanov, R., Alibekova, Z., Dairabayev, M., Narykbay, R., & Omarov, B. (2017). Indoor microclimate comfort level control in residential buildings. Far East Journal of Electronics and Communications, 17(6), 1345-1352.

[26] Song, J., Lee, S., Kim, J., 2015. Crowd Target: Target-based Detection of Crowdturfing in Online Social Networks, in: Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security, CCS '15. ACM, NewYork, NY, USA. P. 793–804. DOI:10.1145/2810103.2813661.

[27] Adikari S., Dutta K. Identifying Fake Profiles in Linkedin, in: PACIS 2014 Proceedings. Presented at the Pacific Asia Conference on Information Systems

[28] Chu Z., Gianvecchio S., Wang H., Jajodia S. Who is Tweeting on Twitter: Human, Bot, or Cyborg? // Proceedings of the 26th Annual Computer Security Applications Conference, ACSAC '10. ACM, NewYork, NY, USA, 2010. P. 21–30. DOI:10.1145/ 1920261.1920265

[29] Omarov, B., Suliman, A., Kushibar, K. Face recognition using artificial neural networks in parallel architecture. Journal of Theoretical and Applied Information Technology 91 (2), pp. 238-248. (2016). Open Access.

[30] Cao Y., Li W., Zhang J. Real-time traffic information collecting and monitoring system based on the internet of things, in: 2011 6th International Conference on Pervasive Computing and Applications. Presented at the 2011 6th International Conference on Pervasive Computing and Applications, 2011. P. 45–49. DOI:10.1109/ ICPCA.2011.6106477

[31] Boshmaf Y., Logothetis D., Siganos G., Lería J., Lorenzo J., Ripeanu M., Beznosov K., Halawa H. Íntegro: Leveraging victim prediction for robust fake account detection in large scale OSNs. Comput. Secur., 2016. 61. P. 142– 168. DOI:10.1016/j.cose.2016.05.005

[32] Egele M., Stringhini G., Kruegel C., Vigna G. Towards Detecting Compromised Accounts on Social Networks. IEEE Trans. Dependable Secure Comput.PP, 1–1, 2015. DOI:10.1109/TDSC.2015.2479616

[33] Gupta, U., Chatterjee, A., Srikanth, R., & Agrawal, P. A Sentiment-and-Semantics-Based Approach for Emotion Detection in Textual Conversations // Neu-IR: Workshop on Neural Information Retrieval, SIGIR 2017, ACM. URL: arXiv:1707.06996

[34] Du, P., & Nie, J.-Y. Mutux at SemEval-2018 Task 1: Exploring Impacts of Context Information On Emotion Detection //Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018). New Orleans, Louisiana, June 5–6. 2018: 345–349

[35] Malmasi, S., & Zampieri, M. Detecting Offensive language in Social Media. 2017. URL: https://www.researchgate.net/publication/321902238

[36] Galán-García, P., Puerta, J.G.D.L., Gómez, C.L., Santos, I., & Bringas, P.G. Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying // Logic Journal of the IGPL, 2016, 24(1): 42-53

[37] Tahmasbi, N., & Rastegari, E. A Socio-Contextual Approach in Automated Detection of Public Cyberbullying on Twitter // ACM Transactions on Social Computing – Special Issue on HICSS 2018, 1, 4. doi>10.1145/3290838

[38] Adewole, K.S., Anuar, N.B., Kamsin, A., et al. Malicious accounts: dark of the social networks // Journal of Network and Computer Applications, 2017, 79: 41-67

[39] Ahmad, M., Agarwal, N., Jabin, S., & Hussain, S.Z. Analyzing Real and Fake users in Facebook Network based on Emotions // 11th International Conference on Communication Systems & Networks (COMSNETS). 2019. DOI: 10.1109/COMSNETS.2019.8711124

[40] Wani, M.A., Agarwal, N., Jabin, S., & Hussain, S.Z. User emotion analysis in

conflicting versus non-conflicting regions using online social networks. Telematics and Informatics. 2018. DOI: 10.1016/j.tele.2018.09.012

[41] Yang W., Boyd-Graber J., Resnik P. A dis-criminative topic model using document net-work structure //Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2016. – T. 1. – C. 686–696.

[42] Peng F., Schuurmans D. Combining na-ive Bayes and n-gram language models for text classification // European Conference on Infor-mation Retrieval. – Springer, Berlin, Heidelberg, 2003. – C. 335–350.

[43] Omarov, B., ANARBAYEV, A., TURYSKULOV, U., ORAZBAYEV, E., ERDENOV, M., IBRAYEV, A., & KENDZHAEVA, B. (2020). Fuzzy-PID based self-adjusted indoor temperature control for ensuring thermal comfort in sport complexes. J. Theor. Appl. Inf. Technol, 98(11).

[44] Jun S., Park S. S., Jang D. S. Document clustering method using dimension reduc-tion and support vector clustering to overcome sparseness // Expert Systems with Applications. – 2014. – T. 41. – No 7. – C. 3204–3212.

[45] Pliakos K., Geurts P., Vens C. Global mul-ti-output decision trees for interaction predic-tion // Machine Learning. – 2018. – C. 1–25.

[46] Abadi M. et al. TensorFlow: A System for Large-Scale Machine Learning //OSDI. – 2016. – T. 16. – C. 265–283.

[47] Dekhtyar A., Fong V. RE Data Challenge: Requirements Identification with Word2Vec and TensorFlow //Requirements Engineering Confer-ence (RE), 2017 IEEE 25th International. – IEEE, 2017. – C. 484–489.

[48] Jain A., Mandowara J. Text classification by combining text classifiers to improve the ef-ficiency of classification //International Journal of Computer Application (2250-1797). – 2016. – T. 6. – No 2.

[49] VK.com - Vkontakte Social Network

[50] Instagram.com - Instagram Social Network

[51] https://vk.com/dev/methods - VK API methods