2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-1451-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/Confluence51648.2021.9377077

Cyberbullying Detection and Prevention: Data Mining in Social Media

Daniyar Sultan
PhD student at al-Farabi Kazakh National University, Almaty, Kazakhstan

Azizah Suliman
College of Computing & Informatics, Tenaga National University, Kuala Lumpur, Malaysia,
azizah@uniten.edu.my

Aigerim Toktarova
PhD student at International Kazakh-Turkish University, Almaty, Kazakhstan

Batyrkhan Omarov
al-Farabi Kazakh National University, Akhmet Yassawi International Kazakh-Turkish University, Kazakhstan
batyahan@gmail.com

Satmyrza Mamikov
University of friendship of people's academician A. Kuatbekov, Shymkent, Kazakhstan

Gulbakhram Beissenova
M.Auezov South Kazakhstan University, University of friendship of people's academician A. Kuatbekov, Shymkent, Kazakhstan

Abstract — **This article is devoted to analysis of the articles on cyberbullying of children and adolescents and creating the methodology of working on the own work, which related especially to Kazakh social media. The article provides examples of similar works by foreign researchers, the United States, England and other European countries. Using the methods of theoretical, namely analysis, synthesis, and empirical: comparison and experiment, the analysis of works on this topic was carried out, and this work was aimed at analyzing the problem in the Kazakh space. In brief, we consider the issue of creating a parser, as well as collecting data for training machine learning and deep learning algorithms that could find and block texts containing a humiliating slope in real time. The article will be of interest to both novice specialists who are engaged in data analysis, and experienced ones to expand their horizons.**

The question of implementing deep learning algorithms requires further study, which is one of the parts of a large work being done, as well as a future topic for a subsequent article.

Keywords— Cyberbullying, Bullying, Classification, Machine Learning, Deep Learning, Social Media, Social Networks

I. INTRODUCTION

Cyberbullying nowadays getting more popular, especially during the pandemic period. Issues of mental and actual brutality that were already just in the social climate have moved to the virtual one. From the start, it appears to be that the type of such provocation is innocuous [1]. In any case, the contrasts among cyberbullying and conventional genuine harassing are because of the highlights of the Internet: secrecy, the presence of a wide crowd, the capacity to make assaults 24 hours per day, and the chance of adulteration [2]. This is stated in the study and Cyberbullying 2017 Russian Association of Electronic Communications - KFS. In this research work, a survey was conducted, as well as their own developments, which are not disclosed for public use, but the results of the study are available in open form. This study is based on a survey with precise questions. The survey involved 2,500 people in the 12-17 and 19-23 age groups. Presently, the issue of savagery on the Internet is turning out to be pertinent in light of the fact that the mental strength of young people is under danger. Mental wellbeing is perceived as "concordance of an individual both with himself and with the climate: others, nature, space"[3]. The victims of cyberbullying are hesitant to impart their issues to grown-ups, in light of the fact that they believe that they might be denied of Internet access [4]. By the by, in the period of data innovation, an advanced young person invests practically the entirety of his free energy in the Internet. In this way, as per the review of the Public assessment Foundation, directed in December 2015. (distributed on 01.03.2016), just 10 % of youngsters more than 6 years of age don't utilize informal communities [10]. For a more point by point investigation of the effect of cyberbullying on the mental soundness of young people, we will present a meaning of this wonder, just as feature its causes, types and jobs [5].

Till that moment we've discussed only about some surveys and their analysis. But we didn't cover cyberbullying detection or recognition strategies yet. Also, we've made a research on exact works aimed to detect cyberbullying texts in media space[6].

In [7][8][21], specialists presented their work, where they used DNN to recognize cyberbullying. In the light of their results it is clean that the classical ML algorithms were beaten. Also they have created neural network to detect. In their approach they used Formspring, Twitter, Wikipedia datasets to train their deep learning algorithm. During the research they faced the problem that the data set is lack of enough data. Generally all the work is well structured, but there were a few configurations, which were stated in wrong way or were not indicated.

Another approach used a data set of [22]. It was decided for the first, because of publicity of the data set: everyone can find it by the link (https://github.com/Mrezvan94/Harassment-Corpus). Also this data set consist of several types of bullying topics: racism, sexual, policy, intelligence.

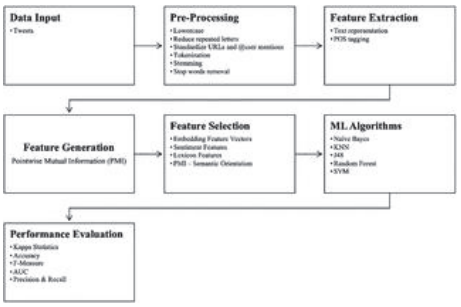

Fig. 1. Methodology used in research work [22]

On the collection step they had collected 24158 posts, where 21045 of them were not bullying one and 3113 give positive answer on the question of cyberbullying[9].

Among broadly agent tests of U.S. web clients matured 10-17 years, paces of detailed online provocation expanded by 83% throughout the former decade, from 6% in 2000 to 9% in 2005 to 11% in 2010 (Finkelhor, 2013; Jones, Mitchell, and Finkelhor, 2012). Another public investigation of long term olds found that a normal of 10% said they were tormented at any rate month to month on the web (Ybarra, et al., 2012a).[10] The rates are higher when analysts look all the more comprehensively at ugliness on the web (Levy, et al., 2012). For example, a public review directed in 2011 by MTV demonstrated that the greater part (56%) of youngsters matured 14-24 years have encountered boisterous attack through web-based media, and most of them (53%) said that the experience was profoundly disturbing (AP-MTV, 2011). In another 2011 overview, led by the Pew Research Center, 88% of teenagers matured 12-17 years who utilize web-based media said that they had seen others being mean or remorseless on a long range informal communication website (Lenhart, et al., 2011)[11].

This information was searched from "The Changing Landscape of Peer Aggression: A Literature Review on Cyberbullying and Interventions" review article from USA. This research gives us detailed information about cyberbullying growing up and how traditional bullying was declined and so on.[12][30] But this research was done physically. By physically we mean that there are no any information systems of software to automate. The one was done in survey format and simple analyzing, managing and grouping the collected information.

II. Materials and Methodology

The main idea of all these works in general terms was the same, but depending on the more deeply set task, different technologies were used. Somewhere there was more work with people, namely polling and then identifying some patterns, and somewhere machine and deep learning algorithms were used[13][33][34]. In studies where most of the work is done with algorithms, databases are used, and at this point the work again diverges in two directions: 1 - the data set is taken from open resources, 2-they build their own tools for data collection, then they do primary processing, cleaning, manual tagging, and then they experiment with algorithms[34].

[14] in his article it was investigated the plausibility of naturally perceiving signs of cyberbullying. Pivotal contrast with related examination is that we don't just model harasser 'assaults', yet additionally more verifiable types of cyberbullying and responses from casualties and spectators (for example all under one parallel mark signs of cyberbullying), since these could moreover demonstrate that cyberbullying is ongoing. The result of this work is not fully working information system to automatically detect cyberbullying, but it needs to be checked by people if some "sign" of cyberbullying will appear. The main problem in this work that data set is unbalanced. It means that only 4-7% of all data marked as cyberbullying positive and about all the recent data is neutral data[15][16].

Despite of such problems or issues they achieved pretty good results in the task automatically detect cyberbullying signals for English and Dutch language post[21]. The results fastened by AUROC(Area Under the Receiver Operator Curve) score. As known AUROC is more durable to unbalanced data.

Working with Twitter social network [22][23][33] did a binary classification for cyberbullying recognition. They have built a system to make semantic direction of each word from dataset and afterward utilized as information highlight in mix of other notable highlights to be specific, word implanting, conclusion highlights, and various expression level vocabularies that distinguish positive and negative logical extremity of assessment articulations[24]. A broad arrangement of tests were performed for recognizing cyberbullying conduct in twofold plan (either cyberbullying conduct exists in the tweet or not) and multi-grouping plan (none, high, medium, low) to identify seriousness in twitter posts. While in comparison with [25] and [26] they this author created a semantic core from tweets they got and did similarity to percentage of hate speech, while [25], [26] aimed to make only binary classification to check if the text conations bullying or not.

The highest point of general classifier execution was accomplished by Random Forest with SMOTE of having kappa measurement of 0.711, by and large classifier precision 91.153, and f-measure 0.898.

The highest point of general classifier execution in paired setting was accomplished by Random Forest to the AUC 0.971 and f-score 0.929. The noteworthiness of proposed highlights is featured by contrasting standard highlights and our proposed highlights in both multi-class grouping and in paired plan. The work requires even more detailed analysis and also requires a distributed dataset with average values of neutral posts and cyberbullying posts[27].

TABLE I. Comparison of related articles

Reference	Algorithms used	Metrics	Corpus size (words)
[1]	Binary classification, Linear Regression	F-score, Precision Recall, AUC, MSE	85,462
[23]	Naïve bayes, KNN, Decision trees, Random forest, SVM	F-score, Precision Recall, AUC	24,189
[24]	Logistic Regression, SGD Classifier, Random Forest, SVM, LGBM Classifier	Accuracy, Recall, Precision, F1 Score	37,373
[21]	LDA, SVM, Linear Regression, Multiple Classification, Deep Learning	Accuracy, Recall, Precision	15,874
[36]	Clustering, Classification, Dimensionality reduction	Accuracy, Recall, Precision	5000
[22]	LDA, Random Forest, KNN, Classification and multiple classification	Accuracy, Recall, Precision, MSA, MAE	17,546

In all these works methodology was similar and follow to similar logic:

1) *Collect relevant data, pre-processing*

Figure 2 shows how data can be collected in the huge media space. Generally, there are 2 types of parsing (based on API

and with the own system architecture). To work with API you need to connect it via token and for another sources of information required to connect via user agent. Therefore, depending on the task, you can choose: to work with API and make your work faster and easier or you will parse another web-resources scrapping the structure of the systems

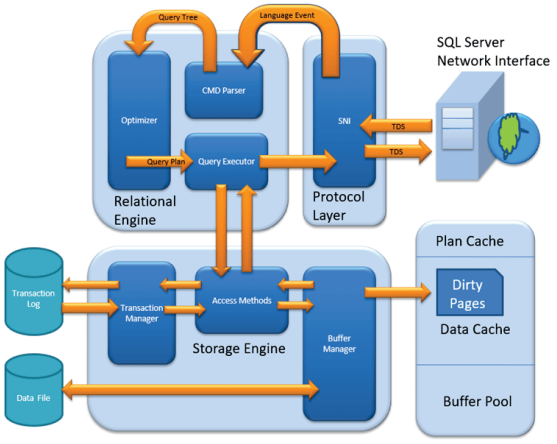

Fig. 2. Scheme of parsing data from web-resources [35]

or

2) *Find one in the open source, like git hub, kaggle and so on*

The easiest way start experience for newcomers in Data Science community is kaggle.com. It is the resource where customer can upload his own data set and give the task to the subscribers. For example, to build best predictive model based on the data. But almost every data is available for free. In this website data is collected to several groups. But in our case, we don't have or we didn't find relevant data set on our aim the research. [28]

3) *Test Machine Learning and Depp learning algorithms on this data set*

There are several Machine Learning algorithms which widely used in terms of predicting and classification[29].

- **Logistic regression** is a factual model that in its essential structure utilizes a strategic capacity to show a twofold reliant variable, albeit a lot more perplexing expansions exist. In relapse investigation, calculated relapse (or logit relapse) is assessing the boundaries of a strategic model (a type of paired relapse)
- **KNN (K-Nearest Neighbours)** is a non-parametric AI technique initially created by Evelyn Fix and Joseph Hodges in 1951. It is utilized for order and relapse. In every case input consists of the closest training set in feature example.
- **Linear Regression** is a direct way to deal with demonstrating the connection between a scalar reaction and at least one informative factors (otherwise called needy and free factors). The instance of one informative variable is called basic straight relapse; for mutiple, the cycle is called different direct relapse
- **SVM(Supper Vector Machine)** are managed learning models with related learning calculations that dissect information for order and relapse investigation. Notwithstanding performing direct grouping, SVMs can productively play out a non-straight characterization utilizing what is known as the portion stunt, verifiably planning their contributions to high-dimensional element spaces.
- **Random Forests** are a troupe learning technique for characterization, relapse and different assignments that work by developing a large number of choice trees at preparing time and yielding the class that is the method of the classes (order) or mean/normal forecast (relapse) of the individual trees. Irregular choice woods right for choice trees' propensity for overfitting to their preparation set. Irregular woodlands by and large outflank choice trees, however their exactness is lower than angle helped trees. In any case, information attributes can influence their presentation.

TABLE II. REPRESENTATION OF ACCURACY OF DIFFERENT WORKS

	[1]	[23]	[24]	[21]	[36]	[22]
Binary Classification	0.83	0.89	0.9	0.9	0.87	0.88
SVM	0.65	0.7	0.67	0.84	0.82	0.79
SGD	0.8	0.895	0.9	0.88	0.88	0.85
Random Forest	0.85	0.87	0.89	0.87	0.83	0.85
Logistic Regression	0.87	0.87	0.9	0.9	0.86	0.84
Neural Network	0.92	0.93	0.93	0.5	-	-
LDA	0.87	0.88	0.9	-	-	0.82

4) *Make comparative analysis between algorithms*

Here all of the authors did pretty good analysis with representation of their results on their topic. [25]. Most common marks to compare the algorithms were:

$$precision = \frac{TP}{TP + FP}$$ [17][18]

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

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

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$ [19]

Those 4 marks are the most commonly used in terms of testing if the model works in relevant way

5) *Documenting the work they did*

III. DISCUSSION

There are bunch of similar research works on this topic, one of them represented in the introduction section. This work focuses to a few points in the current writing on cyberbullying among the present youngsters. Specifically, we need a superior comprehension of which youth are most in danger of being menaces, casualties, and spectators on the

web, just as the conditions under which youth are destined to upstand against cyberbullying conduct. A significant part of this examination should include recording the qualities of existing enemy of cyberbullying endeavors presently utilized in U.S. schools and youth's gathering of these activities. This knowledge will help figure out which mediations merit fortifying and growing and which should be supplanted.[30][31][32]

Another research on school age individuals was done. This article is an audit of the writing of cyberbullying[34]. Primary discoveries are summed up with respect to issues of meaning of cyberbullying, contrasts, and similitudes with conventional harassing; its degree; the types of cyberbullying; the qualities of cyberbullies and cyber victims; the impacts of cyberbullying on the psychosocial improvement of youth; age and sexual orientation contrasts of cyberbullying; and saw reasons for cyberbullying. Also, the means that can be attempted by youth, guardians, educators, and schools to manage the issue and potential pathways for mediations, from a general wellbeing viewpoint, at the individual, class, hierarchical, and network levels are introduced from the writing. At long last, conceivable lawful arrangements getting from both crook and common law are introduced[35].

IV. OUR APPROACH

Using the previous experience of the works done, we changed the course of our research. It was decided to take the direction of automating the process of finding texts containing a humiliating slope. Further, after that, we began to study whether there are similar works on the territory of Kazakhstan or at least on the territory of the former USSR countries. As it turned out, there are several similar works in concept, but even they did not meet our requirements. Since we decided to create an autonomous information system that could automatically detect and take actions to prevent them[36]. Due to the lack of such tools, we came to the conclusion that we will have to do everything ourselves. Next, a plan was created, according to which it was necessary to take several subsequent steps, which in themselves are small projects, as well as topics of research articles:

A. Exploring the media space where textual information is most used
B. Use the methods of data collection in these portals and make an information system that makes it possible to analyze text information in real time.
C. After the data has been collected, do manual marking of the data
D. Experiment with conventional machine learning algorithms (Linear Regression, SVM, KNN...)
E. Create a collection of words from the data set to create the core of the Kazakh language
F. Collect a large amount of training data
G. Make a training of deep learning algorithms for natural language processing

CONCLUSION

The spread of interpersonal organizations on the planet is developing, which implies that an ever increasing number of individuals will encounter cyberbullying. Because of the pandemic, the prominence of online schooling and learning through the Internet is developing. To shield kids from cyberbullying, such messages should be naturally recognized and hindered.

In our examination, we made informational indexing to prepare to distinguish cyberbullying on the Internet. Specifically, we did research analysis on several same topics, did competitive analysis of them with our research direction and gathered a corpus(bag of expressions) of cyberbullying words in the Kazakh language, performed essential information handling and cleaning, and stamped them for the errand of parallel content order.

REFERENCES

[1] Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., ... & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. PloS one, 13(10), e0203794.

[2] X. Zhang et al., “Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network,” in 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016

[3] C. Van Hee et al., “Automatic Detection of Cyberbullying in Social Media Text,” arxiv.org, 2018.

[4] M. Dadvar, D. Trieschnigg, and F. de Jong, “Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8436 LNAI, Springer International Publishing, 2014, pp. 275–281.

[5] Kumari, K., Singh, J. P., Dwivedi, Y. K., & Rana, N. P. (2020). Towards Cyberbullying-free social media in smart cities: a unified multi-modal approach. Soft Computing, 24(15), 11059-11070.

[6] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Int. Conf. Learn. Represent., pp. 1–14, Sep. 2015.

[7] Chavan, V. S., & Shylaja, S. S. (2015, August). Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 2354-2358). IEEE.

[8] Jin-Liang Wanga, Linda A. Jackson , James Gaskin , Hai-Zhen Wang, “The effects of Social Networking Site (SNS) use on college student’s friendship and wellbeing”, Elsevier, Computers in Human Behavior 37 (2014).

[9] https://vk.com/

[10] https://www.instagram.com/

[11] https://vk.com/dev/methods

[12] https://vk.com/dev/

[13] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, “Detection of Harassment on Web 2.0,” in Proc. Content Analysis of Web 2.0 Workshop, Madrid, Spain, 2009

[14] Van Bruwaene, D., Huang, Q., & Inkpen, D. (2020). A multi-platform dataset for detecting cyberbullying in social media. Language Resources and Evaluation, 1-24.

[15] Lu, N., Wu, G., Zhang, Z., Zheng, Y., Ren, Y., & Choo, K. K. R. (2020). Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts. Concurrency and Computation: Practice and Experience, e5627.

[16] K. Reynolds, A. Kontostathis, and L. Edwards, "Using Machine Learning to Detect Cyberbullying," In Proceedings of the 2011 10thConference on Machine Learning and Applications Workshops

[17] https://vk.com/

[18] https://www.instagram.com/

[19] https://vk.com/dev/methods

[20] https://vk.com/dev/SDK

[21] Bechmann, A., & Vahlstrup, P. B. (2015). Studying Facebook and Instagram data: The digital footprints software. First Monday.

[22] https://www.instagram.com/developer/

[23] Stankov, U., Lazic, L., & Dragicevic, V. (2010). The extent of use of basic Facebook user-generated content by the national tourism organizations in Europe. European Journal of Tourism Research, 3(2), 105-113.

[24] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2605034/

[25] S. Agrawal and A. Awekar, "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms," Adv. Inf. Retr., pp. 141–153, 2018.

[26] Rezvan M, Shekarpour S, Balasuriya L, Thirunarayan K, Shalin VL, Sheth A. A Quality Type-aware Annotated Corpus and Lexicon for Harassment Research. Proceedings of the 10th ACM Conference on Web Science. New York, NY, USA: ACM; 2018. pp. 33–36.

[27] Cynthia Van Hee , Els Lefever , Ben Verhoeven , Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans and Veronique Hoste. Automatic Detection and Prevention of Cyberbullying

[28] Bandeh Ali Talpur, Declan O'Sullivan. Cyberbullying severity detection: A machine learning approach, 5.1 – 5.3

[29] Ptaszynski M, Eronen JKK, Masui F. Learning Deep on Cyberbullying is Always Better Than Brute Force. 2017; 8.

[30] Al-Garadi MA, Hussain MR, Khan N, Murtaza G, Nweke HF, Ali I, et al. Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges. IEEE Access. 2019;7: 70701–70718.

[31] Amgad Muneer, Suliman Mohamed Fati, A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter

[32] Van Royen K, Poels K, Daelemans W, Vandebosch H. Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. Telematics and Informatics. 2014

[33] Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. 1995

[34] Livingstone S, Haddon L, Vincent J, Giovanna M, Ólafsson K. Net Children Go Mobile: The Uk report; 2014. Available from: http://netchildrengomobile.eu/reports. [Accessed 30th March 2018].

[35] https://images.app.goo.gl/MHXE7kVjvCeuMSCm9

[36] Narynov, S., Mukhtarkhanuly, D., & Omarov, B. (2020). Dataset of depressive posts in Russian language collected from social media. Data in brief, 29, 105195.