Spring 2022

Deep learning models and Natural Langauge Processing on Twitter tweets to detect Online Harassment, Hate speech and Cyberbullying

Scientific Writing - ENGL 151
Grant Proposal
Dr Lobna Mohamed

Abdullah Kamal Muhammad
201801271
s-abdullahkamal@zewailcity.edu.eg

Abstract

Freedom is a double-edge weapon that can help humans to live a better life and also to destroy the principles of humanity and behavior. Humans invented social media to facilitate communication among humans all over the world, but they did not expect to use it in a wrong way. For example, Microsoft co-operated with tweeter to provide its users with an interesting and modern experience when they talk and teach an Artificial Intelligence (AI) account. This AI account should learn from humans' interactions, but unfortunately bad users gave her hate speech and harmful content which she started to talk with []. Twitter could not prevent her from that and deactivated the account. It this research, the method of how it is possible for social media platforms to detect harmful content and let AI account learn from only constructing data. To do that, it is very good to use new approaches like Knowledge Graph (KG), Natural Language Processing (NLP) and Sentiment Analysis to facilitate the detection of harmful content by linking between concepts, harassment speech and crimes to prevent them []. It is expected to get an accuracy of 91% and recall measurement of 0.9 by using KG method []. There will be many useful implications in preventing harmful content form different social platforms.

Keywords: Natural Language Processing, Knowledge Graph, Sentiment Analysis, harassment speech

I Introduction

As people use more and more social interactions on online social platforms, research on safety and security in social media has grown tremendously in the previous decade. As a result, the number of nasty acts that take advantage of such infrastructure has increased. Because of the anonymity and mobility provided by social media, people

may hide behind a screen, making the breeding and spread of hate speech easy [].

To prevent malicious content, social platforms have used traditional methods such as traditional machine learning models that use the traditional data set. Unfortunately, they did not achieve the desired goal. In this proposal, new approaches will be introduced, such as KG which has higher accuracy, and then models will have higher accuracy for detecting malicious content.

The rapid technological development has helped to increase the spread of hate speech and the diversity of its forms, which increases the responsibility of the developers in the field of Natural Language Processing (NLP) technology to work to reduce this phenomenon [1]. Offensive content such as online harassment is defined as a variety of aggressive or sexually offensive images or words that are transmitted through social media platforms. Hate speech is related to people, countries, events, and crimes that cause psychological harm to specific users [2]. By analyzing the shared content Twitter, it was found that there are many forms of speech content that come from different communities. Therefore, these forms need new approaches to have the capability of classifying the shared content into negative and positive content [3].

Social media platforms are making do their best to detect the contents of hate speech and verbal harassment by using machine learning models that uses their traditional databases [4]. Twitter used its datasets to train Machine Learning (ML) models in order to increasing its ability of providing users with safely usage [5]. Those models used a quantitative and qualitative approaches as a mixed approach of to achieve acceptable accuracy according to [6]. Achieving the required accuracy needs a strong understanding of the meaning and the sentiment of the tweets.

Understanding texts represents the beginning of the right path to prevent the harmful content that is constantly increasing. Therefore, the Knowledge Graph (KG) represents an important and effective step in achieving full understanding of texts. Machine learning models use traditional structures of datasets which lacks relations among words in different sequences [6]. KG avoids that disadvantage by representing people, events, countries and crimes as entities linked by relations. KG uses these relations to conclude new information that differ between harmful and neural content [7].

It will be helpful to introduce a new approach that combine Deep Learning (DL) models that have a high computational power with KG datasets that have a high detection accuracy to perform the data analysis process and to prevent harmful content [8]. DL models mimic the same performance as human neurons, so they learn from the available data and with the mistakes they make, they take responses to improve their performance resulting in high accuracy. According to [8], DL models based on KG datasets achieve an accuracy of 91% which is a significant achievement compared to ML models.

Based on the above review, it is clear that the current research of content detection models does not provide the results required for user protection because it uses traditional databases that treat data as separate blocks without creating relations among them. The DL and KG models indicated that the expected accuracy would be similar to the KG accuracy which is 91%.

These arouse two important questions: Firstly, to what extent do social platforms do their best to prevent hate speech and offensive content shared by destructive users on social platforms by using the available datasets of tweeter and other platforms? Secondly, how can social platforms prevent hate speech and offensive content with massive exchangeable data and messages among users? The proposed model and

dataset is going to have good results and effect on preventing such a problem in social media. Therefore, the aim of the study is to discuss knowledge graphs, deep learning models and new approaches as useful tools that enable social platforms to detect the harmful content shared among users daily. Providing new ways to prevent hate speech, sexual content, and cursing language. And the hypothesis is that these new approaches will protect users and platforms from harmful content

II Methodology

A Research Paradigm

The research aims to improve the accuracy of detecting different kinds of hate speech shared on social media platforms by applying and using new approaches like deep learning models that have been trained on the data of Knowledge Graphs (KG). The researchers will use a mixed methods approach to measure the performance of the models. The quantitative approach will be applied to calculate the accuracy of the models results and the quantitative approach will be applied at the same time to calculate the recall and the precision measurements to investigate classification performance.

B Samples

The study will build KG as a modern technique to have the ability of linking among entities, events and concepts in a better way than the traditional datasets. Social media datasets are scalable dynamic structures with incremental data which need different

approaches to perform data processing tasks on them. KG is the best approach because it will get new insights from the relations among entities to have the complete overview to link among different words that can form hate speech sentences together [6]. The researchers will have two different types of KG; the first one is the built KG which will be obtained through linguists and the second type which will be built from scratch.To build KG, the research will use different datasets of twitter containing hate speech, harassment, and offensive tweets. The number of tweets needed will range from 100 to 200 million randomly selected tweets representing different sorts of hate speech. It will be necessary to use balanced data to obtain balanced results. Deep learning models will be implemented to train on the KGs to achieve high accuracy using supercomputers with a high computational power to perform the analysis

C Tools

The main tools that will be used in the research are supercomputers with high computational power to perform the needed analysis, training and testing of the models. These supercomputers will provide the researchers with the suitable speed and accuracy they are aiming to achieve. The supercomputers are the latest technology provided by Intel which is Core i9-10900 with speed of 2.80 GHz and with an extreme processor of TURBO BOOStT upto 5.20 GHz. They have a powerful RAM of 64GB and graphics RAM upto GB. The storage is very large to contain the massive amount of data, so the supercomputers have 10 TB of pure SSD hard drives providing the suitable speed. Windows 10 Professional will be the operating system. The supercomputers are supported with 2x USB with speed of 2 and 4X USD with speed of 3.2. Output and input devices like a DisplayPort, a HDMI 1.4 Port, a DisplayPort 1.2, a keyboard and a mouse will be needed [9].

D Procedure and Timeline

Firstly, the researchers will apply to have the data needed from twitter and some data analysis tasks will be performed to exclude irrelevant tweets. The time plan estimated 2

weeks to complete that task. The second step will be to buy useful built knowledge graphs that contain hate speech words, events, people's names and concepts. At the same time, dataset specialists will be hired to build new knowledge graphs by using the data acquired from Twitter [10]. The third step will be to preprocess the data and split them into training, validation and testing data to trace the performance of the model. The fourth step will be to calculate the quantitative and qualitative measurements of the model on the validation and testing data. The fifth and last step will be to compare the results of the deep learning of KG with the traditional results of machine learning models.

Training and testing steps will last for 1 month of training and testing the model based on the speed of the supercomputer that will be bought. The timeline of the whole research will last for a month and 2 weeks.

E Data Analysis

The training of the models, the data analysis of measurements, and the testing steps will be performed locally on Jupyter notebook to use the high computational power of the supercomputers. Measuring the accuracy of the models that will be used in the research depends mainly on the accuracy of KG to formulate relations among entities. The accuracy of KG can be determined by constructing the relations among entities and comparing them with the manual estimations of them. The researchers will choose random samples of relations to analyze manually to grant a better performance of KG. Related research indicates achieving accuracy of 91% by using KG.

F Budget

Table 1 Budget

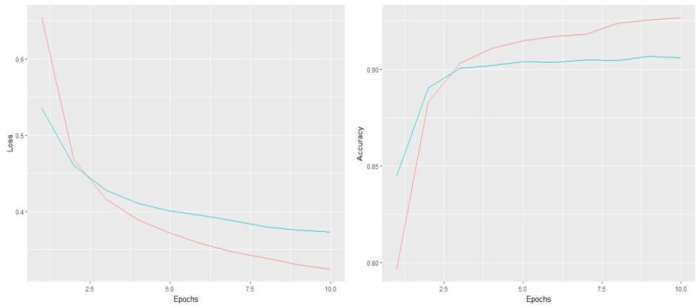
Details	Cost in $	Numbers of items and members

To afford buying needed datasets of twitter	10000$	-
The budget of supercomputers according to Amazon[3]	2850$	4
To afford hirning NLP assistant team	1000$	3
To afford hiring datasets specialists to build needed KG	500$	5

Total cost = 14350 $

III Expected Results

The expected value of recall will be 0.90, the expected value of precision will be 0.91, the expected value of F1-score will be 0.91, the loss function will be 36%, and the expected accuracy will be 91% [11]. The expected values mentioned above are the minimum expected values based on the measurements of KG and the results of the previous studies which will be compared with. As mentioned above, the modern approach will achieve a better performance to detect different hate speech content. These good results will qualify the research to be published in technical journals, such as IEEE and other engineering journals.



(a) Loss (b) Accuracy

Figure I: accuracy and loss functions graphs

IV References

[1] S. P. Kiriakidis and A. Kavoura, “Cyberbullying,” *Family & Community Health*, vol. 33, no. 2, pp. 82–93, Apr. 2010, doi: 10.1097/fch.0b013e3181d593e4.

[2] E. Reed, A. Wong, and A. Raj, “Cyber sexual harassment: A summary of current measures and implications for future research,” Violence Against Women, vol. 26, no. 12-13, pp. 1727–1740, 2019. doi.org/10.1177/1077801219880959

[3] F. Rodriguez-Sanchez, J. Carrillo- de-Albornoz, and L. Plaza, “Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data,” IEEE Access, vol. 8, pp. 219563–219576, 2020, doi: 10.1109/access.2020.3042604.

[4] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, “Machine learning techniques for hate speech classification of Twitter data: State-of-the-art, future challenges and Research Directions,” Computer Science Review, vol. 38, p. 100311, 2020. DOI:10.3390/fi13030080

[5] B. Sabir, F. Ullah, M. A. Babar, and R. Gaire, “Machine learning for detecting data exfiltration,” ACM Computing Surveys, vol. 54, no. 3, pp. 1–47, 2022. doi.org/10.1145/3442181

[6] A. Parihar, S. Thapa, and S. Mishra, “Hate Speech Detection Using Natural Language Processing: Applications and Challenges,” 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), 2021, doi: 10.1109/ICOEI51242.2021.9452882.

[7] F. A. Lovera, Y. C. Cardinale, and M. N. Homsi, “Sentiment Analysis in Twitter Based on Knowledge Graph and Deep Learning Classification,” Electronics, vol. 10, no. 22, p. 2739, Nov. 2021, doi: 10.3390/electronics10222739

[8] Y. Senarath and H. Purohit, “Evaluating Semantic Feature Representations to Efficiently Detect Hate Intent on Social Media,” 2020 IEEE 14th International Conference on Semantic Computing (ICSC), 2020, doi: 10.1109/ICSC.2020.00041.

[9] “Amazon.com: Dell OptiPlex 3080 Micro Desktop 10TB SSD 64GB RAM Extreme (Intel Core i9-10900 Processor with Turbo Boost to 5.20GHz, 64 GB RAM, 10 TB SSD, Win 10 Pro) PC Business Computer : Electronics,” *www.amazon.com*. https://www.amazon.com/Dell-OptiPlex-i9-10900-Processor-Business/dp/B09CCML9XQ/ref=psdc_565098_t4_B09987LVFH (accessed May 24, 2022).

[10] Z. Chen, Y. Wang, B. Zhao, J. Cheng, X. Zhao, and Z. Duan, “Knowledge Graph Completion: A Review,” *IEEE Access,* vol. 8, pp. 192435–192456, 2020, doi: 10.1109/ACCESS.2020.3030076.

[11] E. Barenholtz, N. D. Fitzgerald, and W. E. Hahn, “Machine-learning approaches to substance-abuse research: Emerging trends and their implications,” *Current Opinion in Psychiatry*, vol. 33, no. 4, pp. 334–342, 2020, doi: 10.1097/YCO.0000000000000611