International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 12 Issue 8, August 2023, Impact Factor: 7.751

Cyber Bullying Detection Using NLP and Machine Learning

Mrs. Priyanka[1] Kajal Kumari[2]

[1]Assistant Professor, Department of Computer Science & Engineering, Faridabad College of Engineering and Management, Haryana, India
[2]Research Scholar, Department of Computer Science & Engineering, Faridabad College of Engineering and Management, Haryana, India

ABSTRACT

Twitter is one of the social media that is often used by Indians. Not a few users use Twitter to carry out negative actions such as fraud, spreading fake news, writing things that tend to contain hate speech, to online bullying (cyber bullying). Twitter has developed features such as to reply restriction, rethinking methods, blocking, muting, and reporting to prevent cyber bullying. But to identify each and every tweets that is in context to cyber bulling or not is very cumbersome. Therefore, this study will classify tweets that contain cyber bullying from specified words using the Logistic Regression, Naive Bayes, Decision Tree, and Support Vector Machine (SVM) algorithm with the inculcation of NLP models like CBOW andN-Gram models mark them with the specified label. Cyber-bullying analysis on tweets with optimal accuracy of 94.9%. Consequently, cyber bullying detection system was implemented that demonstrates the validity of the proposed methodology.

INTRODUCTION

This section presents the preamble to cyber bullying or harassment in social networks; which is based on the messages made by users on the social network Twitter. The messages must be analyzed for each of their words and see if together they form a positive or negative proposition to classify it as a bullying or non-bullying message. The types of common users in this social network are user bullies(bully), user-troll(they bother everyone and don't contribute anything), and ordinary users (they don't bother anyone, they only use the services of the social network).From this, the thesis problem, the objectives and contributions to be achieved, and the organization of the thesis document are formulated. The great advances in information and communication technologies have given place for applications such as social networks to be easily introduced as everyday tools for work, studies, or entertainment. Thus, social networks have taken an important turn in the way of communicating and sharing information [1].Social media has allowed the development of techniques to analyze millions of data that are generated day by day. The processing of these data has become a fundamental piece in the definition of strategies like political [3], economic [4], or marketing [5]. The use of social networks to establish cyberbullying attacks was a practice little known in our environment, which worries the control authorities in all parts of the world. This is due to the lack of specialists in the forensic investigation in cyberspace, who can understand the problems of attacks and aggressions that occur in social networks, how the network works, how the system works, and what the vulnerabilities. According to the above, the importance of forming methods, techniques, algorithms, or models that make it possible to control the events of attacks on social networks is evident, which is why this work proposes to form a body that contains tweets, which will be classified according to their expressions as bullying or not bullying. This body will be used to train a neural network to predict new tweets and identify whether or not a text has signs of cyberbullying.

OBJECTIVE OF RESEARCH

Create a model for predicting cyberbullying attacks using machine learning techniques based on a training corpus for the classification and identification of texts with bullying characteristics.

1. Classify tweets as bullying and non-bullying, by categorizing the text based on keywords to form a corpus.

2. Train a machine learning model using the corpus data for the identification of bullying attacks using natural language text processing.
3. Verify the reliability of the model using the cross-validation technique to check the prediction accuracy of bullying text and not bullying.
4. Analyze the corpus of tweets using text mining techniques to identify the characteristics and patterns used in the tweets that have been classified as bullying and non-bullying.

REVIEW OF LITERATURE

CYBER-BULLYING

As defined by [11], it is also known as bullying. cybernetic, and is understood, as the constant and malicious damage done to a person considered as a victim, who is not able to defend himself by his means and is made or executed using electronic means such as the internet, mobile phones, or computers. Victims of cyberbullying are generally selected for meeting certain characteristics that identify them as weak both physically and emotionally. Also, I now feel different from other people, so they become an easy target for aggressors. The most frequent attacks are usually rumors, offenses, insults, threats, extortion, and intimidation, this through the use of electronic devices or means [12].

ARTIFICIAL INTELLIGENCE

According to [18],AI is in charge of the study of intelligence adapted to artificial elements, and oriented from the engineering approach, to create these elements so that they behave smartly. In other words, AI seeks to build systems that are used by machines, so that they perform and behave like a person, and thus would be said to be intelligent. Tasks, such as learning, the ability to adapt to changing environments, creativity, etc., are activities that are generally related to intelligence.

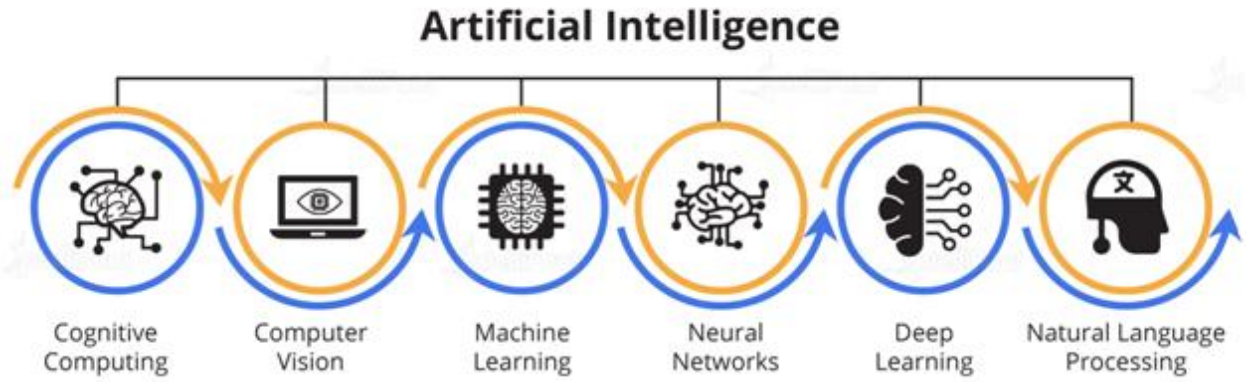
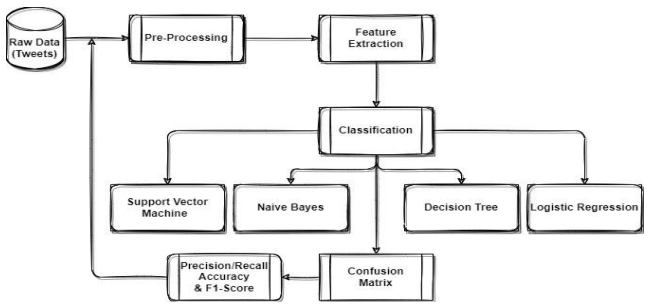


Figure 4: Various Domains of AI

RESEARCH STEPS

The overall steps of this research work consist of collecting data, preprocessing, classification, evaluation, and accuracy detection. Figure 3.1 shows the workflow of this research.



Data Collection
The first step in this research is data collection. Data was collected from Twitter social media in the form of English-language users' tweets. The keywords used to retrieve the data were 'spirit', 'nuts', 'idiot', 'dumb', 'ugly', 'village', 'tacky', 'sucks', 'arrogant', 'stupid', fat', 'hick', and so on. The data obtained is then stored in xlsx (Excel) format. Next, the labeling process is carried out, namely dividing the tweet data into two classes, the cyberbullying class with a Bulling label and the non-cyberbullying class with aN-B label.

METHODOLOGY

TWITTER
Currently, by the number of active users, Twitter is one of the largest exponents of microblogging networks. This network has becomeanimportant information tool with regard to official communication, mainly due to its ability to communicate what is happening in a short time and directly. Many prominent figures such as artists, athletes, businessmen, politicians and professionals, use their Twitter account as their official communication channel.Twitter allows you to post plain text messages with a maximum length of 280 characters. By default, messages are public, and any user can republish a message from another user, comment on it, and share it [76, 77].

Characteristics of Posts on Twitter
Among the main characteristics or elements that usually appear in a tweet, the following can be pointed out [76-78]:
Mentions: Messages may contain one or more references to other users on the network. This reference is achieved by including the user's name in the text, which appears with the @ symbol (at).Hashtags: They are used to represent a specific topic. The message may contain one or more of these tags, through which the tweets that refer to one of them in their text are grouped. They are identified by the initial symbol #.URLs: They allow you to include references to links on the Internet.

PYTHON
Nowadays, scientists of any branch use programming python as a one more tool to solve your problems. They need their tools are simple and efficient, so that they can concentrate on problems in their respective fields. Currently you can see a great trend to use Python in large research centers such as CERN (European Organization for
Nuclear Research) and by scientists in branches such as Bioinformatics, Neurophysiology, Physics, Mathematics, etc. This is due to the availability of visualization libraries, signal processing, statistics, algebra, etc.; easy to use and have very good documentation.

TOOLS SCIENTIFIC
The NumPy (Numeric Python) and SciPy (Scientific Python) packages are mainstays for scientific work today, these often emulate the functions available by Matlab (another scientific language existing ones) which makes it easier to transition from it. The Modular Data Processing (MDP) Toolkit, for example, offers functions for more specific and advanced topics such as Analysis of Principal Components and Independent Component Analysis, very useful for Machine Learning and NLP.

RESULT AND OBSERVATION

When classical pre-processing is done and using a NLP techniques like CBOW and N-Gram, results are obtained in accordance with the literature, sufficing only with taking unigrams despite the fact that they are few messages coupled with the vague writing style, In addition, the baseline has been exceeded. Initially it was thought that by including stop words to help bigram and trigrams would have a better performance in the classifiers, but it has been verified that the n-gram is more effective approach with cbow. Based on experimentation, it was found that by not including the two characteristicspre-processing (eliminate stop words and do stemming) the results may come out with a lower value.

The bulling and non-bulling messages or tweets captured by the tweet extractor correspond to a common language and colloquial, so these messages contain quite a few misspellings, terms coined on the fly and erroneous grammar. This makes it difficult to model language and have common characteristics in different messages that serve as help the classifier. There were also small annotated corpora compared to others given in tweets and this influenced the learning of the classifiers. Other messages required more contexts because the extractor might return a message from a group that belonged therefore this lack of information further complicated the fact of being able to classify a message. Subsequently, the different models of representation of the words, it was possible to appreciate thatthe best is that of grams, and the best classifier are that of Naïve Bayes, Decision Tree and Support Vector Machine. When the labeled tweets for bulling and non-bulling were analyzed and based on the results it is possible to say that if the fact of having these tweets in a messages will improve or worsen the results without the cbow and gram model. So these results indicate that if it is preferable to use grams and cbow

and add them as features to the training vectors of a classifier. Finally, when the combination of classifiers was analyzed, it was found that the Naive Bayes with the accuracy of 94.4%, Logistic Regression with the accuracy of 92.0%, Support Vector Machine with accuracy of 94.9% classifiers presented a excellent results when it was used with combining the cbow and n-gram collectively, whereas the decision tree is with 71.1% accuracy acting as the average classifier used. Therefore, the combination of cbow and grams with classifiers helps when choosing a method of weighting on the best classifiers for detecting bulling and non-bulling under the cyberbulling model.

At the end of the development of this thesis it was possible to see that it is possible to detect cyber-bulling automatically with few messages using robust techniques. Despite some limitations, based on the results obtained, we consider that the objectives of studying different processing techniques of the natural language, as well as apply machine learning techniques to meet the task of detecting cyber-bulling in the messages of the social network Twitter.

CONCLUSION

This chapter presents the conclusions that we have obtained at the end of this thesis. First, the limitations that arose during the study are mentioned development of this work. Then the achievements were addressed, based on the results obtained at the end of it. Afterwards, the contributions we consider that this work presents, from several approaches. Finally this scheme proposes several activities as future work to improve the one presented here.

REFERENCES

[1]. Kapoor, Kawal&Tamilmani, Kuttimani&Rana, Nripendra&Patil, Pushp&Dwivedi, Yogesh&Nerur, Sridhar. (2018). Advances in Social Media Research: Past, Present and Future. Information Systems Frontiers. 20. 10.1007/s10796-017-9810-y.
[2]. Myddleton, Johanna &Fullwood, Chris. (2016). Social Media Impact on Organisations. 10.1057/9781137517036_13.
[3]. Anstead, Nick. (2015). Social Media in Politics. 10.1002/9781118767771.wbiedcs050.
[4]. Dell'Anno, Roberto &Rayna, Thierry & Solomon, Offiong. (2015). Impact of social media on economic growth – evidence from social media. Applied Economics Letters. 23. 1-4. 10.1080/13504851.2015.1095992.
[5]. Zhao, Lanlin. (2022). Effect of Social Media on Marketing. 10.2991/assehr.k.220110.118.
[6]. https://www.statista.com/aboutus/our-research-commitment/2341/tanushree-basuroy
[7]. Nixon C. L. (2014). Current perspectives: the impact of cyberbullying on adolescent health. Adolescent health, medicine and therapeutics, 5, 143–158. https://doi.org/10.2147/AHMT.S36456
[8]. Nandhini, B. &Immanuvelrajkumar, Sheeba. (2015). Online Social Network Bullying Detection Using Intelligence Techniques. Procedia Computer Science. 45. 485-492. 10.1016/j.procs.2015.03.085.
[9]. Sarka, Dejan. (2021). Data Mining. 10.1007/978-1-4842-7173-5_7.
[10]. Thun, Lee Jia&Teh, Phoey& Cheng, Chi-Bin. (2021). CyberAid: Are your children safe from cyberbullying?. Journal of King Saud University - Computer and Information Sciences. 10.1016/j.jksuci.2021.03.001.
[11]. Hinduja, Sameer &Patchin, Justin. (2010). Bullying, Cyberbullying, and Suicide. Archives of suicide research : official journal of the International Academy for Suicide Research. 14. 206-21. 10.1080/13811118.2010.494133.
[12]. Marzano, Gilberto. (2022). Cyberbullying and Social Networking Sites. 10.4018/978-1-6684-5594-4.ch054.
[13]. SathyanarayanaRao, T S et al. "Cyberbullying: A virtual offense with real consequences." Indian journal of psychiatry vol. 60,1 (2018): 3-5. doi:10.4103/psychiatry.IndianJPsychiatry_147_18
[14]. Kim, S., Razi, A., Stringhini, G., Wisniewski, P. J., & De Choudhury, M. (2021). You Don't Know How I Feel: Insider-Outsider Perspective Gaps in Cyberbullying Risk Detection. Proceedings of the International AAAI Conference on Web and Social Media, 15(1), 290-302. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/18061
[15]. Dennehy, Rebecca &Meaney, Sarah & Walsh, Kieran &Sinnott, Carol & Cronin, Mary &Arensman, Ella. (2020). Young people's conceptualizations of the nature of cyberbullying: A systematic review and synthesis of quali tative research. Aggression and Violent Behavior. 51. 101379. 10.1016/j.avb.2020.101379.
[16]. An Educator's Guide to Cyberbullying Brown Senate.gov, archived from the original on April 10, 2011
[17]. Predoctoral, Andrea &Álvarez-García, David & Pérez-Fuentes, María. (2021). Anxiety and self-esteem in cyber-victimization profiles of adolescents. Comunicar. 29. 43-54. 10.3916/C67-2021-04.
[18]. Parwani, Ritwik. (2022). Artificial Intelligence.
[19]. Li, Luo. (2021). Artificial intelligence. 10.4324/9781003246503-9.
[20]. Haslwanter, Thomas. (2021). Machine Learning. 10.1007/978-3-030-57903-6_11.
[21]. Schuld, Maria &Petruccione, Francesco. (2021). Machine Learning. 10.1007/978-3-030-83098-4_2.