AGGRESSION DETECTION USING MACHINE LEARNING MODEL

Sreesanth[1] , Aleena Ibrahim[2], Hasheem M.N[3], Shanavas K.A[4]

[1,2,3] B. Tech students, CSE, Ilahia College of Engineering and Technology, Muvattupuzha, Kerala.
[4] Assistant Professor, CSE, Ilahia College of Engineering and Technology, Muvattupuzha, Kerala.

---***---

Abstract -On social media, aggression has become a big area of tension. However, due to the rapid and increasing rate of content generation as well as the evolution of violent behavior over time, recently proposed machine learning (ML) algorithms to detect various types of violent behavior suffer from a lack. Based on the ML paradigm, this paper describes a real-time system for monitoring aggression on Twitter. Here, we are implementing the system with the Flair Model. This method updates its Machine Learning models gradually when fresh labeled samples are received, and it achieves similar accuracy, precision, and recall in ML models with 93% accuracy, precision, and recall.

Key Words: Twitter, API, aggression, tweepy, flair, machine learning

1. INTRODUCTION

Online Aggressive behavior has been rising recently, with instances of violent behavior being reported in variety of places. This practice has been rising on a variety of platforms like face book, Twitter, Instagram, YouTube etc. A lot of people who are using these social media are being bullied by others. Many popular platforms also has taken action to address these issues by adopting new features and methods because they are frequently getting a unfavorable attention in the media.

So, our paper deals with this issue. To address this problem, we have developed a Machine Learning Model. Our model detects aggressive behavior in Twitter in real time. We have given real-time tweets from the Twitter as input. It will then be detected as aggressive or not by the Flair Model. So, whenever a person tweets an aggressive comment we can detect them in real time using our model.

2. PROPOSED SYSTEM

2.1 Dataset

In Our system we have given tweets as input to the Model. These Tweets are extracted from the Twitter. For that, we have created a developer account on twitter. We have used tweepy module to extract the API. Using Access token, access token secret, consumer key and consumer key secret we have authenticated the API. We have extracted the tweets using this API. These are the input to our model.

2.2 Flair Model

We have used the Flair Model in our system. This model detect aggression over Twitter in Real-time. We have given real time tweets to train and predict the Model. This model classifies a tweet into aggressive and non-aggressive.

2.3 Backend

We have used python language to develop this flair model. latest version of python 3.5 is used. Various modules like tweepy, panda, numpy, re, time, pickele have been imported that is useful for implementing our work. Tweepy has been used for extracting the Twitter API which in turn used to extract the tweets from Twitter.

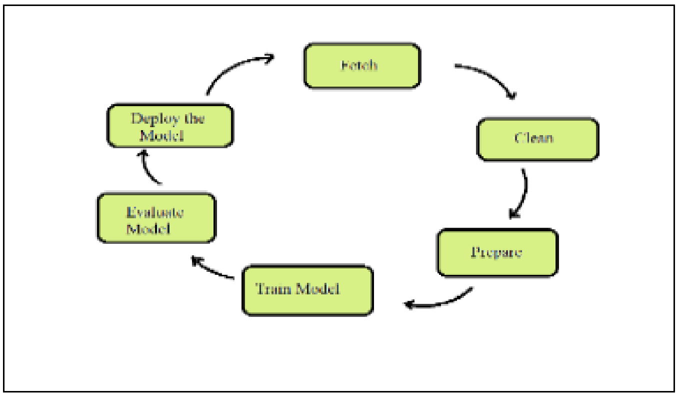

Fig -1: system model

3. METHODOLOGY

3.1 Preprocessing dataset

The extracted dataset may be ambiguous, erroneous and lot of unwanted samples may also appear. So, we need to preprocess the dataset to remove those. During the preprocessing stage unwanted characters, user handlers, http links, digits, special characters, retweet characters, additional spaces are also removed. Stemming and stop word removal is also done.

3.2 Feature extraction

For reflecting users' online presence and subsequently identifying the presence of abusive behavior, a wide range of criteria can be taken into account. Such features may be found in a user's profile, content they have uploaded, or social media platform. We will extract an array of user profiles, texts features, network features. Basic text features, structural and decorative elements, the sentiment conveyed in the posted content, and the use of swearing are all included in this category. Profile features include age of the account, number of posts in the account etc. Network features aims to measure the popularity of a user.

3.3 Training

In this system we are using the Flair model. The Flair model is trained with Realtime samples. Each samples are processed once and the model is updated. Therefore, this model remains always up- to- date.

3.4 Prediction

In this stage, the flair model is used to predict each samples label by calculating the likelihood that it belongs to each class label. Both labelled and unlabeled occurrences are subject to prediction for various reasons. The former is employed for identifying hostile tweets. The latter is helpful for assessing the model's classification performance by contrasting anticipated labels with actual ones

3.5 Testing

We test the model at this point. We provide the model with tweets as instances. It is then compared to the model, and the result indicates whether or not that particular tweet is aggressive.

4. CONCLUSION AND FUTURE SCOPE

People do inappropriate things out of sympathy and disregard for others' sentiments. Therefore, it is vital to censor or remove these contents in order to limit the dissemination of these kinds of communications. In this effort, we are creating a framework for social media aggressiveness detection. The Flair Model is used for training and testing, including feature extraction and preprocessing. Despite any temporary aggressive behaviors, the created ML model are incrementally updated and always current. We can detect the aggression in real time using flair model.

We have developed aggression detection model using Machine Learning model. This detects only in Twitter. In future we can implement similar models other social medias like Instagram, Facebook to detect aggressive users. We have implemented this single user similar idea can be used to monitor a kid by a parent.

REFERENCES

[1] Cyberbullies in Twitter: A focused review by Nicolas Tsapatsoulis, Vasiliki Anastasopoulou. IEEE 2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP).

[2] Multilingual Cyberbullying Detection System by Rohit Pawar, Rajeev R. Raje.

[3] Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network by Xiang Zhang, Jonathan Tong, *, Nishant Vishwamitra, Elizabeth Whittaker, Joseph P. Mazer, Robin Kowalski, Hongxin Hu, Feng Luo at the 2016 15th IEEE International Conference on Machine Learning and Applications.

[4] Unsupervised Cyber Bullying Detection in Social Networks by Michele Di Capua, Emanuel Di Nardo, Alfredo Petrosino on 23rd International Conference on Pattern Recognition (ICPR) Cancún Center, Cancún, México, December 4-8, 2016

[5] Cyberbullying Detection and Prevention: Data Mining and Psychological Perspective by Sourabh Parime, Vaibhav Suri on 2014 International Conference on Circuit, Power and Technologies

[6] A Study of Contact Network Generation for Cyber-bullying Detection by Mingmei Li, Atsushi Tagami in 2014 28th International Conference on Advanced Information Networking and Applications Workshops

[7] Facebook Watchdog: A Research Agenda for Detecting Online Grooming and Bullying Activities by Marlies Rybnicek, Rainer Poisel and Simon Tjoa in 2013 IEEE International Conference on Systems, Man, and Cybernetics

[8] Analysis of Cyber Aggression and Cyber-bullying in Social Networking by Tadashi Nakano, Tatsuya Suda, Yutaka Okaie, and Michael John Moore in 2016 IEEE Tenth International Conference on Semantic Computing.

[9] Mining Patterns of Cyberbullying on Twitter on Charalampos Chelmis, Daphney–Stavroula Zois, Mengfan Yao in 2017 IEEE International Conference on Data Mining Workshops.

[10] Real-Time Detection of Cyberbullying in Arabic twitter Streams by Djedjiga Mouheb, Masa Hilal Abushamleh, Maya Hilal Abushamleh, Zaher Al Aghbari, Ibrahim Kamel in 2019 10th IFIP International conference on New Technologies, Mobility and Security (NTMS)