International Journal of Future Generation Communication and Networking
Vol. 13, No.2s, (2020), pp. 1633–1639

Diversified Troll Detection Mechanisms Using AI/ML Techniques

Mrunmayee Patil[1], Sarang Patil[2], Mayura Rane[3], Aishwarya Gaikwad[4,]

Dr. Shwetambari Chiwhane[5]

1, 2, 3, 4 B.E. Student, Dept. of Computer Engineering, NBN Sinhgad School of Engineering, Ambegoan, Pune-411041, Maharashtra, India
5 Professor, Dept. of Computer Engineering, NBN Sinhgad School of Engineering, Ambegaon, Pune – 411041, Maharashtra, India

Abstract

Online trolling is another form of harassment that made its way over the internet. Various troll detection measures should be implemented to handle this issue. It is necessary to prevent promoting the trend of trolling. This trend has been noticed to be rising on social media to make offensive and inflammatory remarks. A few manual measures like neglecting, reporting or blocking the trolls are being used, yet the rising number of trolls require an efficient automated approach. Some social sites block the trolls on the basis of set of troll words, nonetheless trolls counter these measures by purposefully misspelling their words or other clever methods. This paper emphasizes on various methodologies of anti-troll systems using AI/ML for adapting to the latest trolling techniques.

Keywords*: Machine Learning, Artificial Intelligence, Naïve Bayes, Sentiment Analysis, Twitter troll detection, Anti-Troll System*

I. INTRODUCTION

Now a days our millennial generation is using lot of social media platforms. That has become now bullying platforms for lot of people. One can easily get away by saying anything on such platforms. Today's need is having Anti Troll soft wares which detects trolling over the internet and identify the users. Lot of internet companies are trying to develop applications which can detect and help prevent online bullying but none of them are entirely successful. There are most of the soft wares which only identifies foul words but trolls are using very clever ways for bullying someone they can easily passes these obstacles. We need to create concrete system which identifies all aspects of bullies.

This paper explores the anti-troll systems that are currently in use and their working. It also focuses on various machine learning techniques have been used in these system. In particular this paper explains methodologies and techniques used in anti-troll detection systems. It also describes various results and there comparisons in form of charts. In the conclusion this paper suggest some future scope and opportunities for further development of these systems.

II. RELATED WORK

A. Literature Survey on Sentiment Analysis

Motivation of sentiment analysis is detecting the degree of negativity or positivity. For this process there are some important steps like:

1. Definition of task

2. Annotation rules

3. Collection of data

The terms used in defining the task are:

Repetitiveness: a lot of messages are sent by people who are potentially trolls.

Destructiveness: the messages sent by trolls are generally to express negative sentiments.

Deceptiveness: these messages may be misleading and want to achieve hatred and disagreements.

In annotation section we understand that a troll sends a message or messages which express negativity, irrational opinions or offensive comments. A person who posts such messages on a large scale is a troll.

ISSN: 2233-7857 IJFGCN

When we collect data, we are actually collecting samples which are used as notes or annotations. All the users who have posted in a minimum of 5 threads and minimum of 5 times in each of these threads are scored by Hater News which is an open source tool. After ranking these users by their Hater News scores, we select the highest ranked users. For example, if a person has a troll-score of 0.6, that means he is annotated as troll in 3 out of 5 threads. [3]

B. *Literature Survey on supervised machine learning for troll detection*

By implementing supervised learning, we can link trolling account or fake accounts to the real account to see if the fake account is being used and for what purposes using AI.
Based on the various comments and different posts posted by users, we need to check the authorship of that account. On this collected data, different cyberbullying techniques are implemented.
By analyzing group of different profiles who have some attributes in common, authorship identification is done.
This can be done by following steps:

1. Selection of various profiles
2. Collecting data and tweets from these profiles
3. Different features

Authorship identification techniques let us select various profiles which are similar to each other. All other profiles which don't have a relationship or similarity among them are ignored. There are rule and limitations set by the Twitter API in regards to collecting data. The requests count shouldn't be more than 350 per hour. For choosing user names, profiles and tweets, different java-based methods are used till we have a minimum of 100 unique tweets.
The few features which are used are language, tweets, twitter client and the time of posting. The language helps with geo positioning and helps in filtering. The text posted by the user, that is the tweet, lets us identify a unique writing style that is different for different users. The probability that users have different devices with which they can tweet is high. But many times they use their own preferred client. One more filtering method is provided by this. [2]

C. *Literature survey on impact of different training Data Set on the Accuracy of Sentiment Classification of Naïve Bayes Technique*

The tweets collected need to be analyzed so as to assign labels. Using classifier labels are assigned to twitter data. Using Naive Bayes Technique tweets are classified either into a troll or not a troll. In Naïve Bayes, if a certain attribute is present then it labelled as "1" or else it is "0". By Naive-Bayes rule, probability of relevance for a document is calculated. It is assumed that attributes are not related to each other. For identification purpose, feature is also labelled as an attribute[11].
Classifying the tweets has various processes like collecting the tweets from twitter. Twitter data is obtained from Twython, a default library for getting the tweets from twitter. Preprocessing the tweets, dividing the tweets and classifying by trainer. In dividing the tweets, the training dataset is grouped into 5 different sets. While comparing, the validation part includes around 25 tweets. Grouping of selected tweets are done randomly. So these are some basic steps incorporated in this process[12 13]
The NLTK library from python is used to carry out sentimental analysis. Naïve-Bayes algorithm classifies sentiments for remaining tweets. Previous trained data is implemented as input for this purpose. [8]

Id	Tweet	Label
1	what wonderful day	1
2	The depression isn't over	0
3	Very awesome match	1
4	A clean but enjoyable game	1
5	It was a sad depression state	0
6	weather leads me to depression	0

ISSN: 2233-7857 IJFGCN

Fig. 1 Example of Tweets classified using naïve bayes algorithm

D. Literature survey on Sentiment Analysis using Naive Bayes Classifier Algorithms

Basically sentimental analysis is opinion mining which is used organizations, individuals as well as businesses. To carry out sentimental analysis, Naïve Bayes classifier algorithm is used. Firstly, a training set consisting of positive words and negative words is created. The positive words are labelled as class "1" whereas the negative words are labelled as class "0". This training set consists of 2005 positive words and 4783 negative words. New training sets can be made after scaling up this dataset. [7]
The accuracy of the predicted labels is analyzed through performance parameters. The performance is represented in a form of matrix which is called confusion matrix. Confusion matrix is plotted to sum up the performance of the learning model. A confusion matrix for classes "P" and "N" can be represented as-

		Predicted class: *P*	Predicted class: *N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Fig. 2 Confusion Matrix

TP- The actual class as well as the predicted class is positive.
FN- The actual class is positive but the predicted class is negative.
FP- The actual class is negative but the predicted class is positive.
TN- The actual class as well as predicted class is negative.

Performance parameters are as follows:
Accuracy-
It replies to the question of "How often is the classifier correct?"

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

Precision-
It replies to the question of "When it predicts the positive result, how often is it correct?"

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$= \frac{True\ Positive}{Total\ Predicted\ Positive}$$

Recall-
It replies to the question of "When it is actually the positive result, how often does it predict correctly"

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

F1 Score-
It is the weighted average of recall and precision

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

ISSN: 2233-7857 IJFGCN

E. Literature survey on Twitter Sentiment Classification Using Naïve Bayes Based on Trainer Perception

In this paper using Naïve bayes technique the tweets are classified into three categories i.e. positive, negative and neutral according to the trainer's perception. This perception may vary with different dataset and situations

A specific amount of tweets are taken into consideration for this process and some keywords are selected from tweets for perception training. For example, 50 tweets are selected, then 40 tweets are trained and remaining is the test data. The results were verified by the trainer which were obtained by classification using the Naïve Bayes technique.

Tweets collected are pre-processed and then given to naïve bayes classifier. By training and verifying the sentiment classification by the same person, we could achieve a high degree of accuracy using Naïve Bayes technique. This method is suitable to train and classify sentiment from twitter and other social network data.[5]

III. IMPLEMENTATION

A. System Architecture

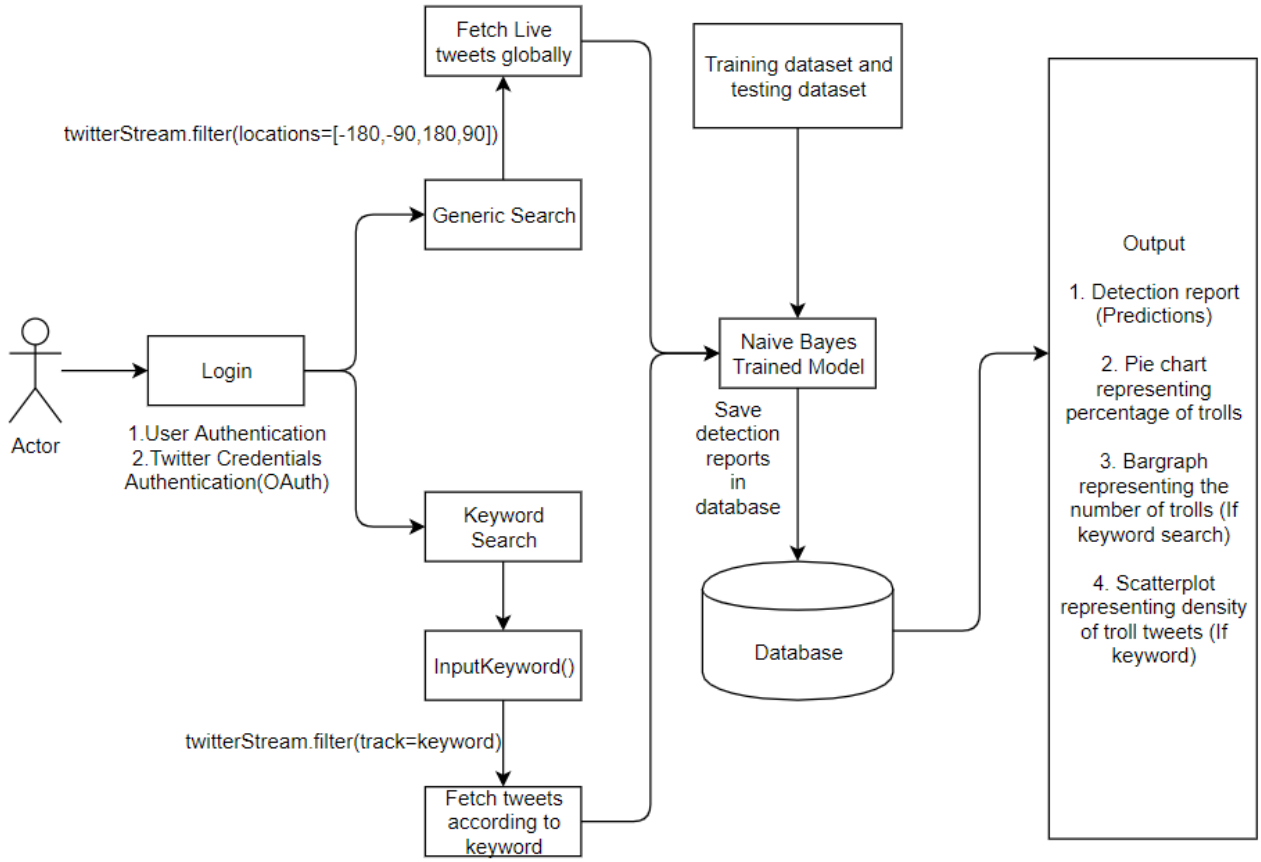


Fig. 3 System Architecture

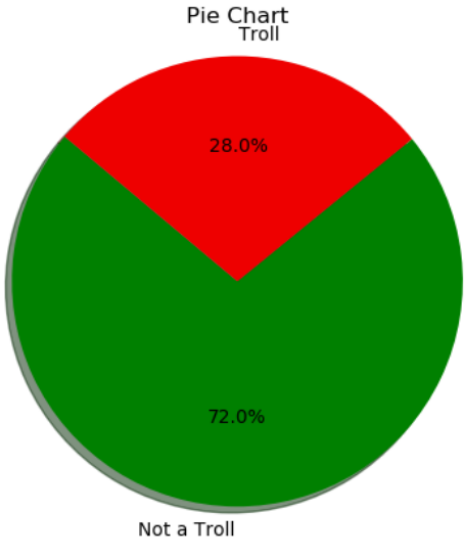
The above figure represents system architecture of the project. Two authentication processes take place as soon as the program starts running, first is user login and then the authentication of twitter credentials that helps us to access twitter data on live basis.

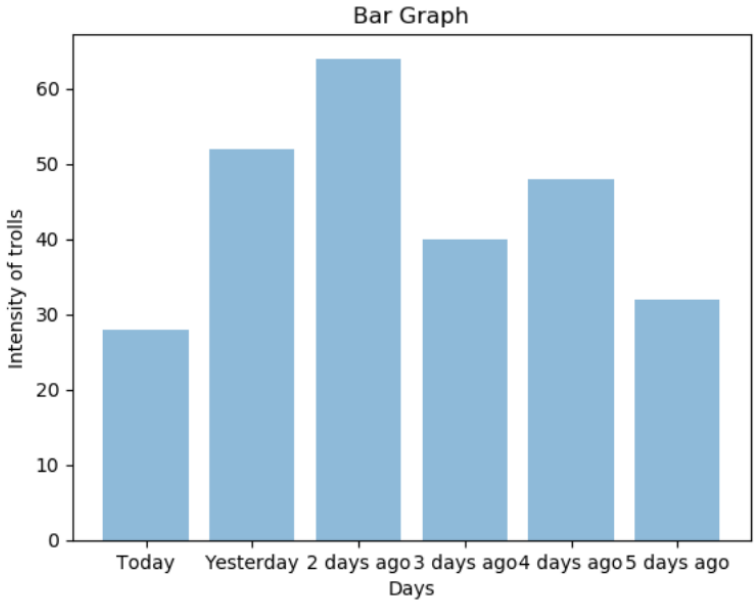
ISSN: 2233-7857 IJFGCN

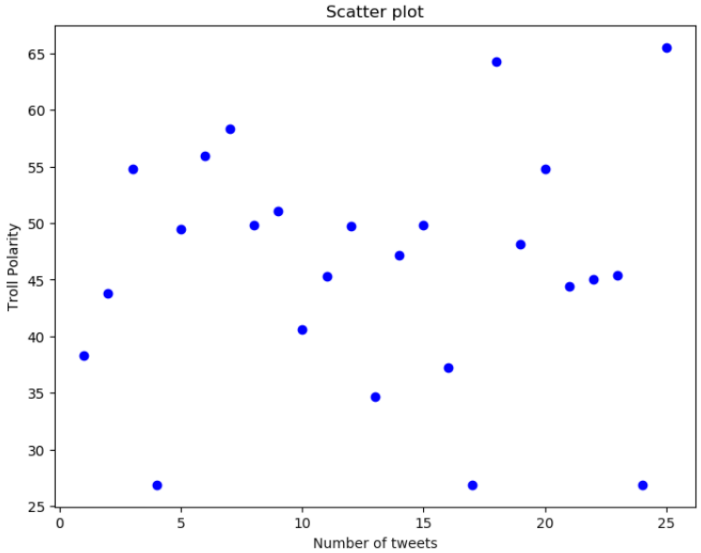
Our system works in two modules: - 1. Generic Search, 2. Keyword Search
In keyword search, we take an string as an input from the user and all live tweets containing the provided keyword gets stored in the database(Live tweets in gets stored in different database, old tweets get stores in different database.). In generic search, we set the location to collect tweets to global co-ordinates and collect all real time tweets and store them in the database. (Not possible to fetch old tweets when keyword is not provided).
We use a dataset labeled into two classes Troll Tweets and Not Troll Tweets to train our naive bayes classifier. This model then predicts the class of fetched tweets and stores the prediction in the database corresponding to the tweets.

B. Analysis Results

Once the processing and prediction is done, output is displayed to the user. User will be shown the tweets classified in two classes. User can choose to report any tweet. A pie chart representing percentage of trolls, a bar graph representing the intensity of trolls since past 7 days to today, a scatter plot representing polarity of trolls.(For generic search, only pie chart and scatter plot can be displayed as we can't fetch old tweets when we search for tweets globally). User can also download the report in .pdf format if desired.







ISSN: 2233-7857 IJFGCN

Fig. 4 Graphical Representation of analysis on tweets

	precision	recall	f1-score	support
0	0.89	0.73	0.80	245
1	0.58	0.82	0.68	115
accuracy			0.76	360
macro avg	0.74	0.77	0.74	360
weighted avg	0.80	0.76	0.76	360

```
[[178  67]
 [ 21  94]]
0.7555555555555555
```
<----Confusion Matrix
<----Accuracy

Fig. 5 Metrics obtained after training model

IV. CONCLUSION

Thus we have implemented Anti-Troll System using Artificial Intelligence. We have used Sentiment analysis in this project. It is a platform for analyzing judgments available in text form from social media. This model used Naïve bayes algorithm to get better accuracy in order to classify troll tweets and not troll tweets. Our system provides various graphical analysis of the data, which helps user to identify ratio of troll tweets, intensity of trolls day wise as well as polarity of tweets. Our system also provides facility to user where hazardous tweets can be reported to the twitter server.

V. REFERENCES

1. Ushma B Bhatt, "Troll-Detection Systems Limitations of Troll Detection Systems and AI/ML Anti-Trolling Solution" , 3rd International Conference for Convergence in Technology (I2CT), April 2018.
2. Patxi Galán-García, "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, Volume: 24, Issue: 1 , Feb. 2016

3. Prakruthi V, "Real Time Sentiment Analysis of Twitter Posts", 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), Dec. 2018

4. Sahar A. El Rahman, "Sentiment Analysis of Twitter Data", 2019 International Conference on Computer and Information Sciences (ICCIS), April 2019

5. Mohd Fazil, "A Hybrid Approach for Detecting Automated Spammers in Twitter", IEEE Transactions on Information Forensics and Security, Volume: 13, Issue: 11, Nov. 2018

6. Zhao, Y. (2016). Twitter Data Analysis with R– Text Mining and Social Network Analysis. [Online] University of Canberra, p.40. Available at: https://paulvanderlaken.files.wordpress.com/2017/08/rdataminingslides-twitter-analysis.pdf.

7. Berna Seref, Sentiment Analysis using Naive Bayes and Complement Naive Bayes Classifier Algorithms on Hadoop Framework, 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Oct. 2018

8. Mohd Naim, Mohd Ibrahim, The Impact of Different Training Data Set on the Accuracy of Sentiment Classification of Naïve Bayes Technique

ISSN: 2233-7857 IJFGCN

9. Dr. C. Nalini, Shwetambari Kharabe, Sangeetha S," Efficient Notes Generation through Information Extraction", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-6S2, August 2019.

10. Udayan Birajdar, Sanket Gadhave, Shreyas Chikodikar, Shubham Dadhich, Shwetambari Chiwhane, "Detection and Classification of Diabetic Retinopathy Using AlexNet Architecture of Convolutional Neural Networks", Proceeding of International Conference on Computational Science and Application, online 05 January 2020, pp 245-253, Springer paper

11. Shwetambari Kharabe, C. Nalini," Robust ROI Localization Based Finger Vein Authentication Using Adaptive Thresholding Extraction with Deep Learning Technique", Journal of Advanced Research in Dynamical & Control Systems, Vol. 10, 07-Special Issue, 2018.

12. Shwetambari Kharabe, C. Nalini," Using Adaptive Thresholding Extraction - Robust ROI Localization Based Finger Vein Authentication", Journal of Advanced Research in Dynamical & Control Systems, Vol. 10, 13-Special Issue, 2018.

13. Shwetambari Kharabe, C. Nalini," Evaluation of Finger vein Identification Process", International Journal of Engineering and Advanced Technology (IJEAT), International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-6S, August 2019.

ISSN: 2233-7857 IJFGCN