Detection of social media platform insults using Natural language processing and comparative study of machine learning algorithms

Sruthi Chiramel
Dept. of Computer Science
Frankfurt University of Applied Sciences
60318 Frankfurt a.M., Germany
chiramel@stud.fra-uas.de

Doina Logofătu
Dept. of Computer Science
Frankfurt University of Applied Sciences
60318 Frankfurt a.M., Germany
logofatu@fb2.fra-uas.de

Gheorghe Goldenthal
Dept. of Computer Science
Frankfurt University of Applied Sciences
60318 Frankfurt a.M., Germany
goldenthal@fb2.fra-uas.de

***Abstract*—The rise of the digital era has been the most ingenious aspect of the millennium. It brought the world closer ,facilitating growth in communication, ideas and thereby trade. The flare of various social media platforms such as Facebook, YouTube, Twitter etcetera precipitated a virtual world which with its many goodness also gave rise to perpetrators who behind their computer screen gained gratification by posting abusive comments in the social media platform.Curbing such activities is the need of the hour in order to protect the vulnerable victims from mental anguish. In this paper, we have proposed approaches to identify and classify social media insults. A comparative analysis between Support vector machines and Random forest has been elaborated in this paper.**

***Index Terms*—Insults, Social Media, Machine Learning, Natural language Processing, Support vector Machine, Random Forest.**

I. Introduction

This millennium has been the age of digital media.The world and people in different parts of the world are no longer as distant they used to be. They are all bound by the internet.A plethora of social medium platform available today has enabled people across the world to communicate efficiently and stay connected. However, along with all the advantages, there has also been a rise of the so called social media abusers, who post derogatory comments in the social media platform. This results in a need for filtering out such comments in order to safeguard the interests of regular users and also to curb the spread of unwanted negativity in the community.

II. Problem Description

Social media has entitled everyone to be an author, which has led to a rise of a new genre of social media abusers who are well-versed at defaming. In the world of web, these derogatory comments stay for ages and could be a reason for anguish amongst people. Cyber-harassment is the new-generation method of abuse and harassment by mode of web usage . It has become increasingly common, especially among teenagers, as the digital environment has flourished and technology has advanced [1]. It is the need of the hour, to address this issue. Although, there are laws enforced in each country related to social media abuse and social media etiquette's; it is a long drawn process to identify the culprit and engaging them in lawsuit. Instead, it would be interesting to counter technology with technology. This is when machine learning could be instrumental in eliminating such abuses at the first level of appearance.

III. Related Research

A. Detection of Abuse in Social Media

Cyber harassment can be essentially categorized under the traditional definition of bullying. One of the most path breaking analyis being formulated by Olweus D.[8]. Olweus described bullying in three categories

- i) **Intention** : The bully deliberately wants to abuse the victim ,
- ii) **Repetition** :The bully repeats the offense to cause the victim metal anguish and
- iii) **Imbalance of power** The bully targets a victim whom he perceives to be inferior to himself.

In order to address the issue of cyber harassment, the authors[9] in their paper 'Automatic detection of cyber-bullying in social media text', have proposed a possible solution with SVM classifier. Their experiment was based on Dutch and English datasets performing ten-fold cross validation and analysing the output on the basis of ROC (Receiver operator characteristic) as a performance metric.
Joni Salminen, in his research[10] has addressed solution for online hate comments using multi-platform data. 197,566 comments were collected from four platforms-YouTube, Reddit, Wikipedia, and Twitter wherein 80% of the comments were labeled as non abusive and the remaining 20% labeled as abusive. The data was run through several classification algorithms.

978-1-7281-9809-5/20/$31.00 ©2020 IEEE

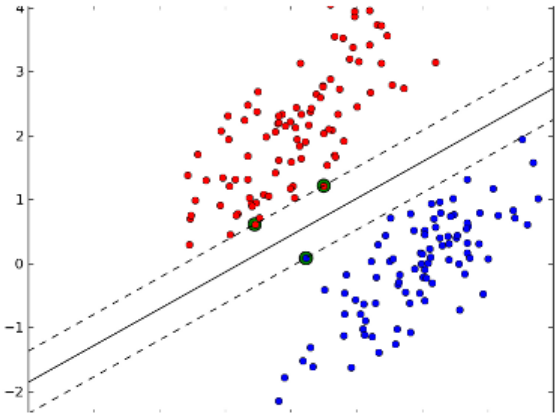

Fig. 1. Hyperplane diving the two classification clusters

K. Dinakar's research work involved detection of hate comments in the YouTube comment section. The system developed bifurcates the comments into various categories such as sexual, body shaming, intellect and cultural background .
Research has also been carried out for online insult detection by use of Fuzzy rule-based system[11]. It is a mathematical tool used to deal with uncertainty and lack of precision. Author B. Sri Nandhini had proposed an approach based on fuzzy inference for online hate detection to help government intervene before the users are subjected to cyberbullying [12] . The developed system makes use of genetic operators like crossover and mutation for optimizing the parameters and obtain precise type of cyberbullying activity[12].

IV. Proposed Approaches

Natural language processing(NLP) capability of machine learning algorithm can be tapped into addressing the issue of social media insults by implementing "Text Classification". The goal of our text classifier will be to segregate text as an insult or not an insult. We would be specifically exploring the scope of Support Vector Machines and Random Forest for the purpose of text classification by performing a comparative study of the same.

Support Vector Machine:

The Support Vector Machine (SVM) was inroduced by Vapnik and ever since, it has been one of the most innovative algorithms in the machine learning research field. [2]. SVM is primarily credited for its performance with regards to classification accuracy when compared to other classification models. Its basic design philosophy is to maximize the classification boundaries and its basic purpose is to maximize the hyper-plane[3].

The working of Support Vector machines can be summarized as below :

- Start with data with relatively low dimension.For eg. One-dimensional data.
- Move the data to a higher dimension. Eg.from One-dimensional data to two-dimensional data
- Find a support vector classifier that classifies the higher dimensional data into two groups.
- Kernel functions are used to systematically find support vector classifier in higher dimension .

Fig.1 depicts the hyper-plane dividing data into 2 clusters.

Random forests:

Random forests are an ensemble method or a classifier that constructs a collection of decision trees during the training phase and outputs the mean prediction of individual trees.
Formally, a random forest is a predictor consisting of a collection of randomized base regression trees[16]:
$(r_n(x, \theta_m, D_n), m \geq 1), where \theta 1, \theta 2$,... are i.i.d.(independent and identically distributed) outputs of a random variable θ. These random trees are combined to form the aggregated regression estimate:
$\bar{r_n}(\mathbb{X}, D_n) = \mathbb{E}_\theta[r_n(X, \theta, D_n)]$,
where $\mathbb{E}_\theta$[16] denotes expectation with respect to the random parameter, conditionally on $\mathbb{X}$ and the data set D_n. The most customary choice for predictors in each leaf is to use average response over the training points which fall in that leaf[4]. Decision trees are very similar to a human brain decision making. Also, over-fitting could be excluded in random forest classifier because it is the mean of all the decision trees. Fig.2 depicts the decision tree formation in random Forest [5].

Below mentioned are the most commonly used feature importance factors based on Random forests:
Gini importance: This is the most commonly used measure in random forests. Gini importance is obtained on the basis of Gini Index on the emerging random forest trees [17]. A bifurcation function called as "Gini Index" is used by random forest classifier inorder to decide the attribute to be split during learning phase of trees. For eg, Under the consideration that binary classification consists of two classes let p be a fraction of positive examples designated to a node k and $(1 - p)$ be fraction negative examples. Then, the Gini index at m is defined as:
$G_k = 2p(1 - p)$
Smaller value of p indicates a purer node.
Permutation based variable importance: Another important method used while performing feature selection using random forest is the random forest permutation importance. It explains the stand alone influence on each variable in parallel with the multivariate interactions with the other features at a given time[14]. It works under the principle of an intuitive permutation strategy and is more widely used than Gini importance.

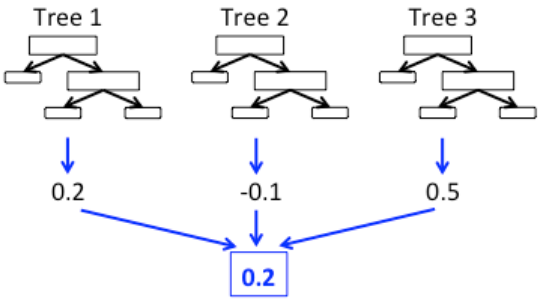

Fig. 2. Tree bifurcation in Random Forest

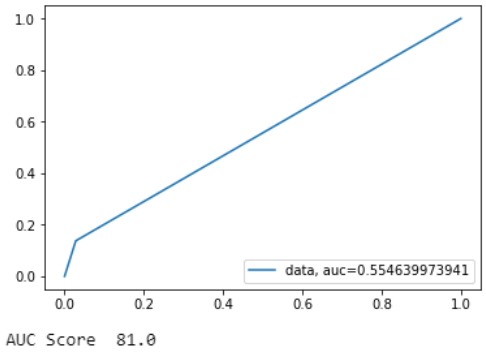

Fig. 3. ROC random Forest

V. Experimental Setup and Implementation Details

The models have been developed using Python language. The various python packages used for analysis are numpy, pandas, sklearn, nltk and matplotlib. Our Experimental set up consisted of the below mentioned steps :

- Collecting data.
- Data Wrangling
- Defining model.

A. Collecting Data

The data set has been procured from Kaggle [6], there are in total 3947 number of training data and 2235 of test data. There are in total 3 columns- ID, Date and Comment. We did our analysis using various Python libraries. The dependent variable has been predicted by assignment of Boolean values. Natural Language Processing techniques as explained below has been used to convert unstructured data into structured data which will eventually help in providing analytical insights.

B. Data Wrangling

The comments contain certain words which are redundant to be used in a Machine Learning Algorithm. Hence, it is essential to wrangle the data into a readable format to be inputted in the learning algorithm. The below mentioned approaches has been used :

- Tokenization: It is the process when the given sentence is broken down into individual words or tokens, and also at times discarding punctuation. Below mentioned is an example of tokenization:
  Input: Tom, Nichole, Assemble here
  Output: | Tom | | Nichole | | Assemble | | here |

- Removal of stop words: A very commonly used approach for information retrieval in natural language processing is the so-called bag-of-words model. A basic approach is the removal of uninformative or redundant words, commonly referred to as stopwords [7].

- Stemming and Lemmatization :Stemming is used to reduce related forms of a word to its basic form. For instance:
  he, his, him ⇒ he
  house, houses, housing, ⇒ house.
  Lemmatization is the process of converting words into their dictionary format.

C. Defining and Training Model

We have used a linear classification model Support Vector Machines and a non - linear classification model Random forest for training the data in Python using Scikit-Learn. We have used accuracy and 10-fold cross validation to arrive at the primary evaluation metrics. The training accuracy has been utilized to analyse how well the models are fitting with the testing data.

VI. Results and Discussions

After training the model, the test data metrics used for evaluation of the model is the Precision ,Recall , Accuracy and AUC obtained from the confusion matrix. Fig.3 depicts the ROC curve and AUC obtained from random forest. We have done a comparative study between Support vector machine and Random Forest. The metrics obtained before performing 10-fold cross validation from both the methods is as follows:

SVM metrics
Precision: 81.1
Recall: 85.2
Accuracy: 82.5

Random Forest metrics
Precision: 85.0
Recall: 86.2
Accuracy: 82.5

Post the 10-fold cross validation, the results obtained are as follows:

Technique	Normal Accuracy	Cross Validation
SVM	83.54	82.71
Random Forest	83.54	82.23

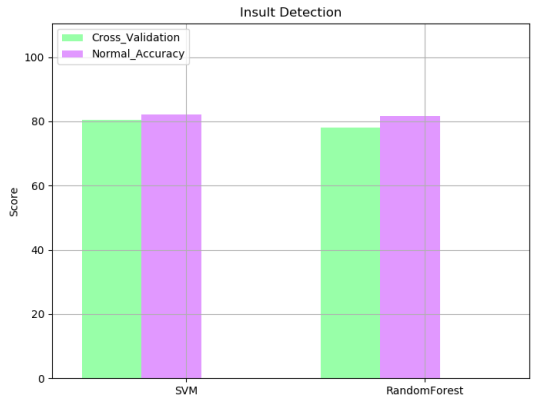

Fig. 4. Normal vs 10-fold cross validation accuracy comparison

Technique	Precision	Recall
SVM	0.85	0.81
Random Forest	0.82	0.82

Fig.4 depicts the improvement in accuracy in random forest after we applied 10-fold cross validation. Based on the results we observed that after the 10-fold cross validation, random forest yielded better result with:

- High Classification Accuracy - 83.5%.
- High AUC Score - 81.0%
- Balance Precision / Recall values : 82%-82%.

VII. Conclusion

The evaluation of the results suggested random forest to be more robust for the data set, although SVM is also a good classification method . As a part of future work, it would be interesting to extend this research by use of fuzzy logic or deep neural networks. social media insults are one of the most unpleasant offering of the digital world and as always, addressing real world issues more efficiently is the real gratification for the scientific community.

References

[1] Smith, Peter K.; Mahdavi, Jess; Carvalho, Manuel; Fisher, Sonja; Russell, Shanette; Tippett, Neil . "Cyberbullying: its nature and impact in secondary school pupils". The Journal of Child Psychology and Psychiatry. 49 (4): 376–385,(2008).
[2] V. Vapnik. The Nature of Statistical Learning Theory. NY: Springer-Verlag. 1995.
[3] Yujun Yang , Jianping Li , Yimei Yang,The research of the fast SVM classifier method, IEEE,2015
[4] Misha Denil ,David Matheson,Nando de Freitas1,Narrowing the Gap: Random Forests In Theory and In Practice
[5] https://databricks.com/blog/2015/01/21/random-forests-and-boosting-in-mllib.html,Last accessed 11/06/2020
[6] https://www.kaggle.com/c/detecting-insults-in-social-commentary
[7] Martin Gerlach, Hanyu Shi Luís A. Nunes Amaral, A universal information theoretic approach to the identification of stopwords, 2019
[8] Olweus D. Bullying at School: What We Know and What We Can Do. 2nd ed. Wiley; 1993.
[9] Cynthia Van Hee, Gilles Jacobs, Chris Emmery,Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter DaelemansAutomatic detection of cyberbullying in social media text,2018
[10] Joni Salminen, Maximilian Hopf, Shammur A. Chowdhury, Soon-gyo Jung, Hind Almerekhi Bernard J. Jansen ,Developing an online hate classifier for multiple social media platforms, 2020
[11] Victoria Lopez, Alberto Fernandez, Maria José del Jesus , Francisco Herrera, "A hierarchical genetic fuzzy system based on genetic programming for addressing classification with highly imbalanced and borderline data-sets", Elsevier, Knowledge-Based Systems 38 (2013) 85–104
[12] B. SriNandhini, J.I.Sheebab,Online Social Network Bullying Detection Using Intelligence Techniques, Volume 45, 2015, Pages 485-492
[13] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," in Proc. IEEE International Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 2011
[14] Chen, X., Wang, M., Zhang, H.: The use of classification trees for bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1(1), 55–63 (2011)
[15] Yanjun Qi,Random Forest for Bioinformatics,http://www.cs.cmu.edu/ qyj/papersA08/11-rfbook.pdf,Last accessed 12/06/2020
[16] Gerard Biau, Analysis of a Random Forests Model,Journal of Machine Learning Research 13 (2012)
[17] Breiman, L.: Random forests. Mach. Learn. 45, 5–32 (2001). DOI 10.1023/A:101093340432