Effective Mining of Unstructured Twitter Data for Detecting User Persecution

A. Afiya, Shaik Javed Parvez, S. Arun

Abstract--- With the increase in social media platform there has been heave in user spawn data. These data has generated in remarkable amount on daily basis through all micro blogging sites and most of the data are unstructured. Social media platform like Twitter has become a major platform for people to share their daily thoughts, opinions, suggestions and also pave a way for abusing other users verbally. We propose to investigate user oppression detection on twitter. In the paper we propose data mining techniques for mining unstructured twitter data and apply deep learning concept on tweets. We also present a case revise to exemplify the effectives of proposed system. In this paper we have tried to point the opportunities of future work by providing a expansive perspective on open forum for data mining.

Keywords--- Data Mining, Classification Algorithm, Unstructured Data, Deep Learning.

1. INTRODUCTION

Predominant increase in network connectivity has promised for wide expansion in resource sharing and various alliances. Social platforms like Facebook, YouTube and Twitter has gained much popularity among the people through Web 2.0. Most of 80% data are in unstructured form [1]. This allows people to build strong connection, share different kind of information like pictures, videos and their thoughts on social networks. It is apparent that advent of this real time application has caused people life more convenient also threatening at times due to user oppression. Abusive tweets sharing, commenting and posting cause menace to others users. Social Media source provides immense amount of data. User generates text data on daily basis through tweets and comments. Dataset taken is through Twitter API. Sample amount of twitter words taken for sampling is around 10000 words which contains all kind of tweets. On daily basis huge amount of unstructured data is generated through social mediums and other sources. Most of these data move unused. Information Retrieval (IR) has gained more insight in recent years due to availability of data in digital form and flexibility to access those data. As the name implies it is a process of retrieving significant information from database.

Data mining subset of is a process of retrieving information from dataset and making into a more understandable format. Through data mining interesting hidden patterns can be discovered. The process of getting deeper inside the data is also termed as Data Mining. It involves following six basic classes of task like

1. Anomaly Detection : Identifying the data of interest or identifying odd data
2. Association Rule Learning: Understanding dependencies between variables
3. Clustering: Discovering groups and similar arrangements in data
4. Classification is method of used to predict output based on data analyzes.
5. Regression is process used to predict numbers.
6. Summarization is visualization of results

A single label Naïve Bayes algorithm is used for classification on online discussion data to understand the reflective thinking of teachers [2]. The methodology works by initially collecting the teachers online discussion data, data storage, data sampling and classification modeling and result analyzing. They implemented the initially 2000 post with 70 percent for training and 30 percent of 2000 for tests data. And found Naïve Bayes classifier had improved performance than SVM Classifier.

1.1 Classification

Classification is an important phase of data mining it is data analysis task. It predicts the data model based on given train and test set of data. Classification is a process of defining a group or class of data into one predefined set of data based on each item in dataset. The main aim of classification is predict accurately the target class for each case in data. Before classification there are few steps required to convert the raw data. For example consider a medical student want to analyze the cancer data and predict which one of the treatments must be given to patient among the treatment1, treatment2 and treatment3 is termed as classification and prediction. It is a data analysis process where a classifier is build to predict the data model. It has two steps learning and classifying.

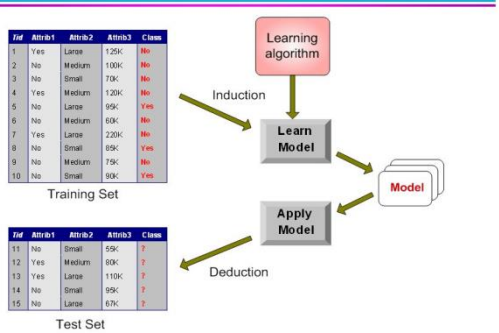

Fig. 1: Illustration of Classification (Source: Google)

Revised Version Manuscript Received on 14 February, 2019.

A. Afiya, Department of Computer Science & Engineering, Vels Institute of Science Technology & Advanced Studies (VISTAS), Chennai, Tamilnadu, India. (e-mail: Afiya.aslam.ar@gmail.com)

Shaik Javed Parvez, Department of Computer Science & Engineering, Vels Institute of Science Technology & Advanced Studies (VISTAS), Chennai, Tamilnadu, India. (e-mail: parvez.se@velsuniv.ac.in)

Dr.S. Arun, Department of Computer Science & Engineering, Vels Institute of Science Technology & Advanced Studies (VISTAS), Chennai, Tamilnadu, India. (e-mail: arun.se@velsuniv.ac.in)

2. RESEARCH METHODOLOGY

The Figure 2 depicts the process of proposed system which comprises of Collection of Data, Labeling, Modeling of Data and Learning Algorithm.

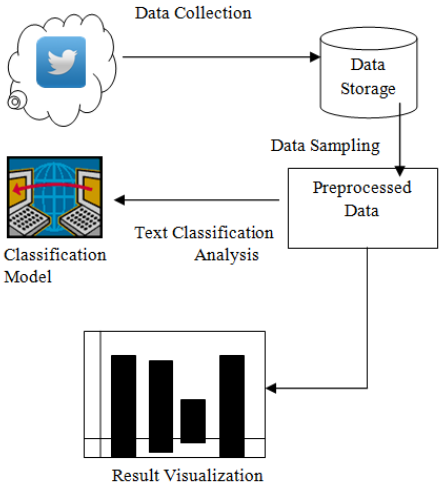

Fig. 2: Research Study Design

Step 1: Data Collection through Twitter API

Step 2: Preprocessing of dataset followed by clean up and Padding

Step 3: Splitting of Dataset training set and test set

Step 4: Classification algorithm implementation LSTM and CNN

Step 5: Result analysis and visualization.

2.1 Data Preprocessing

The process of transforming the raw data into understandable format is termed as data preprocessing. The real time data is incomplete is lack in certain trends and preprocessing solves these issues. In this paper we are using open source tool anaconda python for mining the dataset. The following standard packages have to be downloaded before loading the dataset into the environment.

- numpy
- scipy
- matplotlib
- scikit-learn
- pandas
- keras python library

NLTK is a standard library for processing Natural Language.

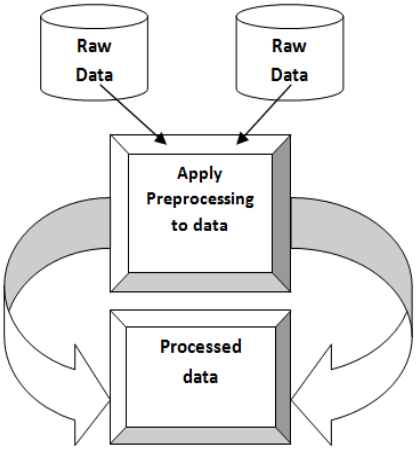

Fig. 3: Data Preprocessing

Step 1: Loading the twitter dataset

Step 2: Validate programmatically to identify valid tweets.

Step 3: Applying NLTK modules for stop word removal like ‘is’, ‘the’ etc

Step 4: Text Normalization is a process used to prepare text for further processing. Lemmatization is used to achieve base form of root word.

2.2. Data Cleaning and Padding

Data Cleanup is a process of removing incorrect data from the dataset or cleaning and filling the missing values from the dataset. There are various methods provided by pandas for cleaning the missing values. Example removal of extra space, extra special characters so on.

```
Hello  world!!!!
Hello     world !!!!
Hello          world !!!!
```

The function pad_sequence() is called for fixing the data at preferred length and the length of sequence is 50 words for this dataset. Traditional Learning faces major accuracy issue due to poor dataset so there is a necessity for data cleanup.

After cleanup the data must be uniform with other data.

Fig. 4: Padding Sequence

2.3. Stratified Shuffle Splitting of Data

In this phase the twitter data is divided into train set and test set with a split of 80:20 ratios on our sample twitter dataset. Before training a model it is necessary to split the data into one for train set and other into one for test set. Consider while working on a model initially the data is trained after training the model has to be tested on different dataset which completely irrelevant to train data. But this could not be always possible in that case the data is split into train and test sets. Sklearn library is imported for splitting dat. Shuffle Split allows randomly splitting test and training data.

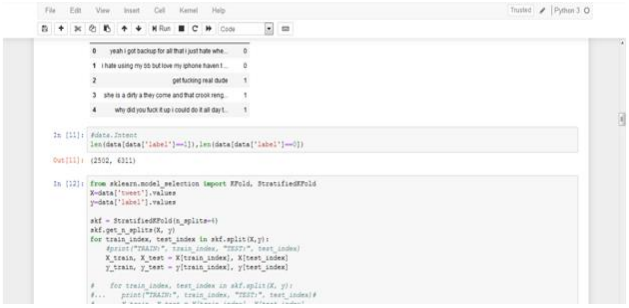

Fig. 5: Data Split

Step 1: Split the Twitter Dataset into two different sets

Step 2: x for independent variables and y for dependent variables.

Step 3: Split the x variable into xtrain and xtest

Step 4: Split the y variable in-to ytrain and ytest

Step 5: The data is split into 80-20 ratios

3. ALGORITHM IMPLEMENTATION & RESULTS

Neural Networks are part of Machine Learning Algorithms. Most of Deep Learning uses neural network architecture but conventionally neural network has two to three hidden while deep learning networks has more than hidden layers and it is often termed as deep neural networks.

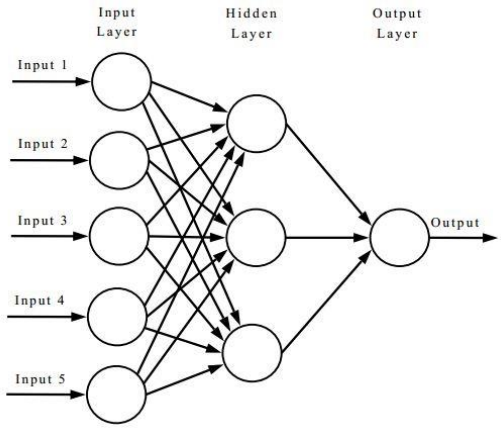

Fig. 6: Example of Neural Networks

The concept behind Deep learning networks is more data, more models leads better results. The neural networks are based on biological neurons. Example of deep learning is voice controller of mobile devices, PC's and other wireless devices.

3.1 LSTM

RNN (Recurrent Neutral Network) works by creating loop in network architecture which serves as memory state. This is modeled to remember what all learned from the beginning.

Vanishing Gradient is the problem that arises due to memory state in the model which means learning from large memory makes network more difficult to remember previous layers and tune parameters of earlier layer. LSTM (Long Short Term Memory) is type of RNN that was created to overcome the gradient issues by using tanh. LSTM is also called as cell state which updates the memory periodically. Looping arrow is recursive and stores the previous data. There are three gates in general the forget gate modifies the cell and input modulation gate.

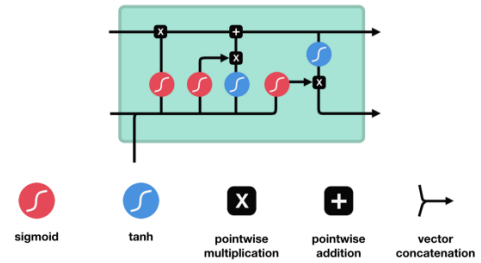

Fig. 7: LSTM Architecture (Source: Medium)

3.1.1 Forget Gate

Sigmoid can output 0 or 1 for forget or remember memory. x(t) and h(t-1) are inputs of sigmoid layers and it decides which information must be kept and deleted. For example consider a person named Peter and statement is about him"he is a good friend of lia" . Now the gender of peter can be removed since the subject is about lia. The Equation of Forget Gate = f(t) * c(t-1).

3.1.2 Input Gate

Hidden state and current input is passed into tanh (-1,1) for regulating network. The output from tanh and sigmoid are multiplied and Sigmoid decides which information must be kept and discarded.

3.1.3 Cell State

Information is collected above to built cell state. The cell state is multiplied by forget vector and drops the non required values. The output from input gate gets added to cell state. Finally the cell state has updated values.

3.1.4 Output Gate

Output Gate determines out of all values which must be forwarded to hidden layers. The new hidden state and cell is carried to next phase. Step 1: Input Layers takes sequence words as input

Step 2: Using LSTM Units compute the output layer

Step 3: The activation of certain neutrons are turned off to prevent over fitting

Step 4: Compute the new best output

Step 5: Run for five epochs

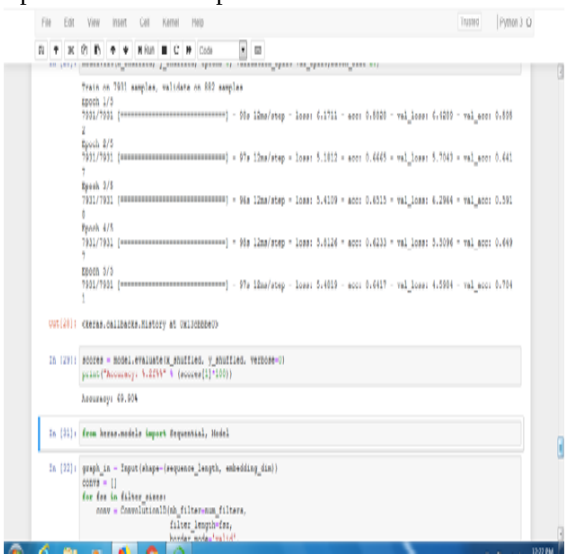

Fig. 8: Implementation of LSTM using Python

3.2 CNN

Convolutional neural network is a class of deep neural networks precisely used in image related problems or image visionary problems. Recently applied to NLP problems [9],[10] and the results are promising. There are three layers in general convolutional layer, pooling layer and classification layer.

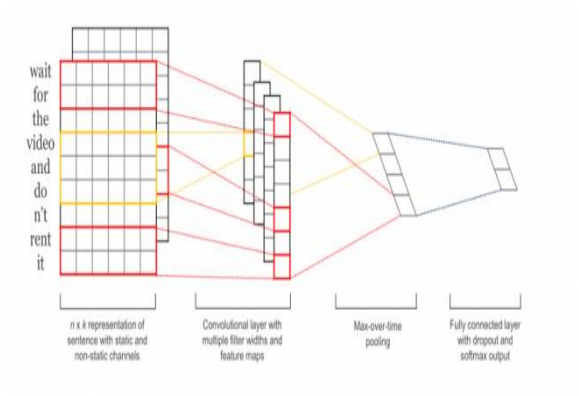

Fig. 9: Illustration of CNN (Source: Wildml)

Step 1: Input test is represented in matrix form n x k where maximum number of words refers to n and length of embedding refers to k

Step 2: Apply convolution multiple times with k as width along different height

Step 3: Apply Max pooling across the output filter and select one output from each filter (for detecting any feature presences)

Step 4: Last is fully connected layer and output of softmax. Softmax is used for probability distribution.

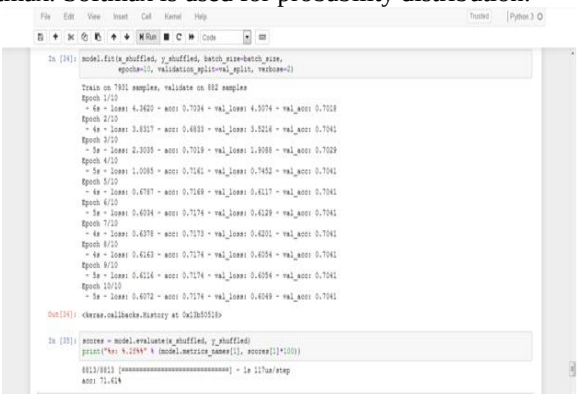

Fig. 10: Implementation of CNN using Python

Method	Accuracy Rate
LSTM	69.04
CNN	71.61

Fig 11: Accuracy of Classification Algorithms

4. CONCLUSIONS

In data mining the traditional classification method learns from the data and makes decision based on the learned data. While Deep Learning classification algorithms which are detachment of ML but functioning capabilities are different and precise. LSTM and CNN create neural networks and make an intelligent decision. The main aim of classification algorithm is to generate accurate and better results. Henceforth there are different classification algorithms and the each works for different data. There are more deep algorithms for different requirement and may work well for text mining concepts.

REFERENCES

1. T.K.Das , P.Mohan Kumar, "BIG Data Analytics: A Framework for Unstructured Data Analysis" International Journal of Engineering and Technology (IJET) Feb 2013
2. Qingtang Liu, Si Zhang, Qiyun Wang, and Wenli Chen. "Mining Online Discussion Data for Understanding Teachers Reflective Thinking" on IEEE transaction 2017
3. Ravi. Parikh and M. Movassate, "Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques", Final Report, CS224N.,2009
4. Akshi Kumar and Teeja Mary Sebastian "Sentiment Analysis on Twitter "IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012
5. Natural language processing. Gobinda G Chowdhury. Annual Review of Information Science and Technology. 2005
6. Jiarui Zhang, Yingxiang Li1, Juan Tian1, Tongyan Li1 "LSTM-CNN Hybrid Model for Text Classification" IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) 2018
7. G.Kesavaraj and Dr. Sukumaran "A Study On Classification Techniques in Data Mining" ICCCNT 2016
8. Smitha .T, V. Sundaram "Comparative Study Of Data Mining Algorithms For High Dimensional Data Analysis"IJEAT 2012
9. Y. Kim, "Convolutional Neural Networks for Sentence Classification", Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751.
10. Y. Zhang, B. C. Wallace, "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification", 2015
11. www.analyticvidhya.com for Algorithm Learning.
12. Implementation understandings and analysis Medium.com