DATA ANALYSIS

MITESH J PANCHAL

PG number: 20134275
mp20944@essex.ac.uk
dept. School of Computer Science and Electronic Engineering
University of Essex
Colchester campus, UK

***Abstract*—The public now a days more active on social media rather than the past. Today's generation obtain information from the networks rather the old scenario's of news paper,magazine and television. There for the famous organization moving towards the network side for advertising of their product, so they are connected with the public. In particular,organizers use the social media and evaluating/analyzed the data received from the community by the mode of tweeter so they can improve their image,provide proper facility to community and raise their business. We are going to take three different benchmark from TweetEval-Hate speech detection,Offensive language identification and Sentiment analysis. The data import from the tweets,checking the data type and further evaluation for null values and dropping the null values.**

I. Introduction

Twitter is a great source for Natural Language Processing researchers. It has sufficiently large size of data, along with outstanding qualities - which include of real-life conversations, uniform length, lots of variety,and real-time data stream. The data from the TweetEval of Hate speech detection, Offensive language identification and sentiment analysis were taken for the data prepossessing. The Hate speech mainly defined on the basis of characters like gender, color,religion and so on. Hate speech detection measure on the tweets received for hateful and not-hateful nature of public.

The Offensive language identification- In today's era offensive language is the base platform like Facebook and Tweeter, so the manual detection of the review is not possible. The task is generally done by supervised classification examples.The data is compile by its offensive or non-offensive criteria.[3]

The Sentiment analysis is the text type analysis with the social media platforms like twitter. The sentiment analysis task generally done on the base as the text expression is a Positive,a Negative or a Neutral sentiments by a person or an event.[4]

The Data from the TweetEval of Hate speech detection, Offensive language identification and Sentiment analysis were importing and data were compiling for checking the nulls values weather present in the data or not. After checking the nulls values, dropping the nulls values in the formats so we can analyzed it for further process in data cleaning activity. In Data cleaning, shorting the raw in the base of longer in length and complete the analysis on this data sets. It is necessary to clean it thoroughly so the further data preprocessing will be done on the basis of it.[2]

II. Literature Review

Examples of offensive content studied in previous work include hate speech(Davidson et al., 2017; Malmasi and Zampieri, 2017, 2018),and aggression. Moreover,[5][6] given the multitude of terms and definitions used in the literature, some recent studies have investigated the common aspects of different abusive language detection sub-tasks (Waseem et al., 2017; Wiegand et al., 2018)[7]
The Sentiment Analysis in Twitter task has been run yearly at SemEval since 2013 (Nakov et al., 2013; Rosenthal et al., 2014; Nakov et al., 2016b)[8], with the 2015 task introducing sentiment towards a topic (Rosenthal and the 2016)[8] task introducing tweet quantification and five-point ordinal classification

The Sentiment Analysis in Twitter task has been run yearly at SemEval since 2013 (Nakov et al., 2013; Rosenthal et al., 2014; Nakov et al., 2016b),[8] with the 2015 task introducing sentiment towards a topic (Rosenthal et al., 2015)[]8 and the 2016 task introducing tweet quantification and five-point ordinal classification .

III. Methodology

The 1St step in the data preprocessing is to understanding the data and according to that we should choose the proper data processing method.

The basic steps as :

A. Call the data

The TweetEval data sets have their raw text data set. The data sets from the tweets data call by df-train=df.read-fwf(") function.

Identify applicable funding agency here. If none, delete this.

B. Import the Data

The called data is further importing and the raw and columns size can be count.

C. Check for Nulls values

The data now having much nulls values so checked for nulls values available in the data.

D. Dropp the Nulls values

The nulls values is not be prefer for further analysis,so dropping the nulls values from the data sets.

E. Data cleaning

The data contains much longer raw length so we analyzed its thoroughly.

IV. Result

.

- The Nulls values available as:

  1.Hate speech detection is @user nice new signage. Are you not concerned by Beatlemania -style hysterical crowds crongregating on you… 0

  2. Offensive language identification is ibelieveblaseyford is liar she is fat ugly libreal snowflake she sold her herself to get some cash !! From dems and Iran ! Why she spoke after JohnKerryIranMeeting ? 0 Unnamed: 1 851 Unnamed: 2

  3. Sentiment analysis is ”QT @user In the original draft of the 7th book, Remus Lupin survived the Battle of Hogwarts. HappyBirthdayRemusLupin” 0 Unnamed: 1

V. Discussion

- The Data sets from the TweetEval found in raw and columns form.

- The data sets of Hate speech detection,Offensive language identification and Sentiment analysis are as follows:
  1. Hate speech detection are 8992 raw’s and 1 column shape.
  2. Offensive language identification are 859 raw’s and 3 columns shape.
  3. Sentiment analysis are raw’s 45614 and 2 columns shape.

VI. Plan

1. Data convert in to lower class.

2. Data should be handles and URLs free.

3. Data Normalization.

4. Data Vectorization.

5. Own model training.

6. Validate the model.

VII. Refences

1.Multi-class Classification of Tweets Based on Kindness Analysis Charles Han, Wanzi Zhou, Xinyuan Huang hcs@stanford.edu, wanziz@stanford.edu, xhuang93@stanford.edu

2. SemEval-2019 Task : Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter

3.SemEval-2019 Task :Identifying and Categorizing Offensive Language in Social Media (OffensEval)

4. SemEval-2017 Task : Sentiment Analysis in Twitter

5.Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM)

6. Shervin Malmasi and Marcos Zampieri. 2017. Detecting Hate Speech in Social Media. In Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP).

7. Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. arXiv preprint arXiv:1705.09899.

8. Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 task 9: Sentiment analysis in Twitter. In Proceedings of the 8th International Workshop on Semantic Evaluation. Dublin, Ireland, SemEval ’14, pages 73–80.