Offensive Language Detection Using Textual Dataset

-by BATCH 11C

Domain:

- Machine learning
  - Machine learning is the study of computer algorithms that improve automatically through experience.
  - It is seen as a part of artificial intelligence.
  - Machine learning enables analysis of massive quantities of data.
  - Increases the computational power.

Offensive Language:

- Offensive language is the offence of using language in a way which could cause offence to a reasonable person.
- Offensive language usage has grown as an important societal challenge.
- Most of the Offensive language usage was through textual data.
- Offensive Language usage leads to cyber bullying.
- Cyber bullying may lead to deep psychiatric and emotional disorders for those affected.

Platforms for Offensive language

Input:

- Two datasets are being given as the input to the program. The data is in csv format.
- Each dataset consists of 3 columns
  - id
  - comments
  - label
- The two datasets are
  - Train Dataset
  - Test Dataset

Output:

- Labelling of harsh words in comments
- Frequency of harsh words in the comments.
- Calculation of f1-score.

Process:

- Collection of datasets from resources.
- Data preprocessing
  - Removing of @handles
  - Removing of punctuations and special characters
  - Removing of short words
  - Tokenization
  - Stemming

Feature Extraction:

- A process by which an initial set of data is reduced by identifying key features of the data
- Feature Extraction can be done with two methods
  - Bag-of-Words
  - TF-IDF
- TF-IDF can be applied to each training document at once.

Using TF-IDF

- WHAT IS TF-IDF?
  - It is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.
  - TF : term frequency.
  - IDF : inverse document frequency.

- **Term Frequency (tf)**: gives us the frequency of the word in each document in the corpus.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

- **Inverse Data Frequency (idf):** used to calculate the weight of rare words across all documents in the corpus.

$$idf(w) = log(\frac{N}{df_t})$$

Tf-idf score:

- Combining these two we come up with the TF-IDF score (w) for a word in a document in the corpus.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

Classification:

- Machine Learning algorithms are used for classification. Some of them are
  - Support Vector Machine
  - Naive Bayes
  - Decision Tree
  - Logistic Regression
- Logistic Regression can be best used as the obtained values are continuous between the range of 0 and 1.

Logistic Regression:

- The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth.
- We obtain a S-shaped curve ranging the values between 0 and1.
- logistic regression equation:

  y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))

  Where y is the predicted output,

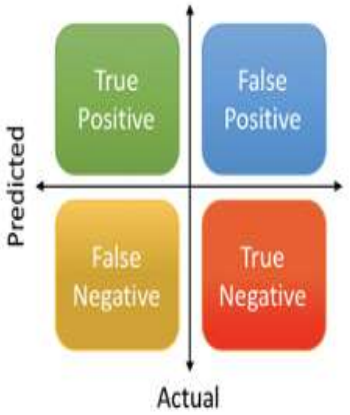
  b0 is the bias or intercept term and b1 is the coefficient for the single input value (x)

F1-Score:

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

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

Accuracy = $\frac{\text{True Positive + True Negative}}{\text{Total}}$



F1-score helps us to understand the accuracy of the model we construct.

$$\text{F1} - \text{score} = 2 * \frac{Recall \ * Precision}{Recall + Precision}$$

Conclusion:

- We implemented a program to identify the harsh and offensive words in the comments mentioned.
- We can know the frequency of the offensive words.
- Labelling of harsh words and classifying them as positive and negative words.

References:

- Hariani, Imam Riadi, "Detection of Cyberbullying on Social Media Using Data Mining Techniques", International Journal of Computer Science and Information Security (IJCSIS) Vol. 15, No. 3, March 2017
- Cole, Cornell, Dewey, Sheras, "Identification of School Bullies by Survey Methods", Professional School Counselling, Vol. 9, April 2006
- Machine Learning-Blog, "The Logistic Regression Algorithm", https://machinelearningblog.com/2018/04/23/logistic-regression-101/, Accessed 13 March 2020
- Gandhi Rohith, June 7 2018, Towards Data Science, "Support Vector Machine — Introduction to Machine Learning Algorithms", https://towardsdatascience.com/support-vectormachine-introduction-to- machine-learningalgorithms-934a444fca47

Thank You!!!