Detection of Cyberbullying on Reddit Comments

Team Members: Akshya Ramesh, Rathnapriya Gopalakrishnan

Introduction

Cyberbullying is a blanket term that is used to represent online abuse, harassment, sharing or posting hurtful and abusive messages, doxing, and reputation attacks [1]. Cyberbullying is more prevalent on social media platforms like Facebook, Twitter, Instagram, Reddit, Snapchat, and TikTok. There are many difficulties when it comes to curbing cyberbullying on social media platforms. The main one is not being able to identify the source of the offensive comment, which can either be from an individual or from a group of people. These comments can leave a digital footmark on social media platforms which might be difficult to erase. The victims of cyberbullying are subjected to stress, anxiety, low self-esteem, and negative thoughts which can lead them to harm themselves or develop suicidal thoughts. There are various reasons why people choose to cyberbully others, some of them being - jealousy, need for power, trying to fit in the society, boredom, etc. [1].

With the current increase in the use of the latest technologies, 95% of teens in the United States are active on social media and 56% of them are victims of cyberbullying in one way or the other [2]. Of which 42% of teens were called offensive names on social media platforms [2]. The LGBTQ+ communities are often the main victim experiencing a more than average rate of cyberbullying [3].

About 83% of people believe that the social media platform needs to take necessary actions to identify and tackle cyberbullying for the betterment of mental health [3]. We planned to tackle this issue and wanted to contribute to the society by building a model that can be used by social media platforms to identify whether a new comment posted can be considered as cyberbullying comment or not.

Problem Description

Identifying and labeling Reddit comments as cyberbullying or non-cyber bullying using NLP (Natural Language Processing) and predicting the same using Machine Learning

Aim

- Label the unsupervised Reddit comments as cyberbullying or not using Sentiment Analysis and Offensiveness Proportion
- Classify new Reddit comments as cyberbullying or not using Machine Learning models

Data Description

We used two datasets for this project, both taken from online resources. The main dataset consists of 1 million Reddit comments taken from Kaggle (https://www.kaggle.com/datasets/smagnan/1-million-reddit-comments-from-40-subreddits). It has 4 attributes as follows:

- Subreddit – This was a categorical variable that represented the category on which the comment was posted
- Body – This was a string variable that contained the comments

- Controversiality – This was a binary variable that contains the aggregated metric which denotes how controversial the comment was.
- Score – This was numerical variable that calculated the difference in upvotes and downvote.

Of the above-mentioned variables, we use only the body attribute in our project.

The abusive word dataset was taken from GitHub – Profanity_en.csv (https://github.com/surge-ai/profanity/blob/main/profanity_en.csv). This dataset contained a list of 1598 abusive and offensive words from the English language. This dataset also contained attributes like text, canonical_form, category, severity_rating, and severity_description. From the above-mentioned attributes, we only selected the text column and saved it as a new CSV file and used it in our project.

We uploaded both the files to our personal google drive and accessed them Google Colab notebook to perform text processing and machine learning.

Methodology

Our methodology involves data collection, data preprocessing, labeling the unsupervised data, sampling, feature engineering, building machine learning models, and performance evaluation of those models.

Data Collection and storage:

Our Reddit data is collected from Kaggle, and the abusive words dataset is collected from GitHub. They are both stored in google drive.

Data Preprocessing

We access the stored reddit datasets as a Pyspark dataframe and the abusive word dataset as a Pyspark RDD.

1. **Tokenization:**

   We performed lower case conversion on the comments stored in the 'body' column. We then convert the comments to a list of words using a Pyspark dataframe split function and stored them in a new column.

2. **Lemmatization**:

   We first convert the comments to an RDD from a dataframe and we perform mapping operation along with lambda function to speed up the process. We use NLTK library's WordNetLemmatizer to lemmatize the list of words for each comment. This helps us to convert the words to their base form.

3. **Stop-word Removal:**

   We use the English 'stopwords' dataset from the NLTK library' corpus package. For each word in a comment, we checked if it is a stop word or not, if so, we removed the stop word from our list. After this operation, we got a stop-word free list of words for each comment.

Labelling

Our novel contribution to this project is the algorithm used for labeling unsupervised data. In order to classify a Reddit comment as either a cyberbullying or non-cyberbullying comment, we used both the sentiment of that comment along with the proportion of offensive and abusive words in it.

1. **Sentiment Analysis:**

   We utilized the Vadar Sentiment library for calculating the sentiment of each comment. The SentimentIntensityAnalyzer's polarity_scores() give us four values as follows. The neg (negative), pos(positive), neu (neutral) and compound values. The compound value helps us identify if the comment was addressed in a positive, negative, or neutral way.

   - If the score is between –0.05 and 0.05, the sentence is categorized as having a neutral sentiment.

   - If the score is above 0.05, the sentence is categorized as having positive sentiment.

   - If the score is below –0.05, the sentence is categorized as having negative sentiment.

   We only need the compound value, so we extracted it using regular expressions for each of the original Reddit comments.

2. **Offensive Proportion:**

   We first counted the total number of offensive words in each comment by comparing it to the abusive word dataset. We extracted the abusive words in an RDD and calculated their count through len() function. This gives us the count of offensive and abusive words for each comment. We calculated the offensive proportion by dividing the number of offensive words by the total number of words in the comment after stop-word removal.

3. **Labeling algorithm:**

   We consider a comment as cyberbullying in two scenarios:

   1) If the overall sentiment of the comment is negative and the proportion of abusive words is greater than a certain threshold.
   2) If the overall sentiment is neutral, the proportion of abusive words is greater than a 0.5

   We consider a comment as non-cyberbullying in the following scenarios:

   1) If the overall sentiment of the comment is positive or neutral and the proportion of abusive words is less than a certain threshold.
   2) If the overall sentiment of the comment is positive but the proportion of abusive words is greater than a certain threshold.
   3) In all other scenarios

   We used various threshold values and found 0.25 to be effective for this reddit dataset.

Sampling:

We found that the number of cyberbullying records is very low in comparison with the non-cyber bullying records. In order to tackle this class imbalance, we performed under-sampling so that we get an even ratio of cyberbullying and non-cyberbullying records to feed the machine learning models.

Machine Learning:

1. **Train-Test split**

   We split the sampled data into 80% training and 20% test data. The independent variable is the comment, and the target variable is the class (cyberbullying/ non-cyberbullying)

2. **Feature Engineering**

   In feature engineering, the independent variable is tokenized using the Tokenizer from Pyspark's ML package. We vectorize the feature through hashing using Pyspark ML's HashingTF(). We perform IDF calculation which identifies the importance of a word in a document using Pyspark ML's IDF ().

3. **Model Building**

   We built three machine learning models to predict whether a comment is cyberbullying or not. We built Logistic Regression, LinearSVC and RandomForestClassifier model from Pyspark ML's classification package. We fit the model to the training data and tested it on the test data. We calculated the evaluation metrics of these models by using MulticlassClassificationEvaluator from Pyspark ML Evaluation package. The following metrics were calculated – accuracy, F1 score, Precision and Recall.

4. **Performance Evaluation**

   Due to the heavy text processing operations, running the entire 1 million records was time-consuming, the below results were obtained from models built and evaluated on 50k records.

Evaluation	**Logistic Regression**	**SVM**	**Random Forest**
Precision	0.333	0.375	0.333
Recall	0.5	0.5	0.5
Accuracy	0.438	0.5	0.438
F1 Score	0.444	0.508	0.444

From the above evaluation results, we can see that Random Forest and Logistic Regression perform the same, and SVM has a slightly higher performance rate when considering precision, accuracy, and F1 score.

Problems Encountered

Our initial idea for labeling the unsupervised data was to use only the offensiveness proportion but later we realized that such an algorithm does not consider sarcasm and will not be able to identify comments that are addressed in a positive way but containing a few commonly used inappropriate words.

In order to overcome this issue, we used both sentiment analysis and offensiveness proportion. So that the algorithm will consider the sentiment behind the comments and correctly label them.

Limitations

Since our project contained heavy text processing operations which takes up a lot of time and resource our personal google cloud was not able to handle such high load. So instead of running the entire 1 million records we ran 50K records to check the efficiency of our model.

Results

We were able to label our unsupervised data using both Sentiment Analysis and offensive proportion. We constructed three machine learning models – Logistic Regression, SVM, and Random Forest and calculated the performance evaluation metrics such as Accuracy, Precision, Recall, and F1 for all the three models. We plan to improve our model's performance by feeding data from other social media platforms like Facebook, Twitter, Instagram, YouTube, etc.

References

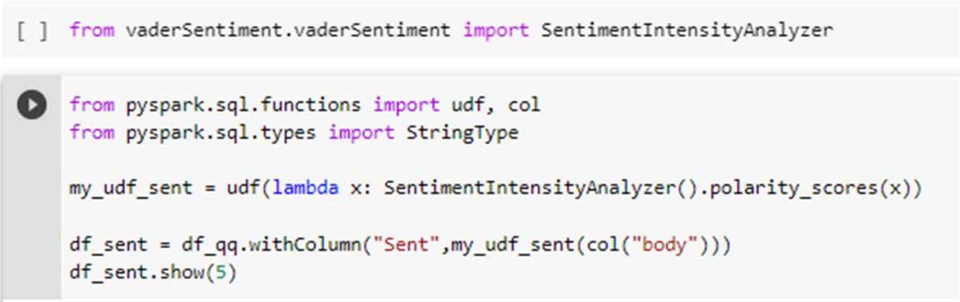
1. https://www.cybersmile.org/advice-help/category/what-is-cyberbullying
2. https://www.pewresearch.org/internet/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/
3. https://www.dosomething.org/us/facts/11-facts-about-cyber-bullying
4. https://www.statista.com/topics/1809/cyber-bullying/#topicHeader__wrapper
5. https://enough.org/stats_cyberbullying
6. https://medium.com/@junwan01/oversampling-and-undersampling-with-pyspark-5dbc25cdf253
7. https://stackoverflow.com/
8. https://www.tutorialkart.com/apache-spark/spark-mllib-tf-idf/
9. https://www.kaggle.com/datasets/smagnan/1-million-reddit-comments-from-40-subreddits
10. ttps://github.com/surge-ai/profanity/blob/main/profanity_en.csv

Appendix 1

Team Member	Contributions
Akshya Ramesh	Text processing operations, Sentiment Analysis, ML – Logistic Regression, Model Evaluation
Rathnapriya Gopalakrishnan	Data Collection, Text processing operations, Offensiveness Proportion calculation, Machine Learning – SVM, Random Forest, Model Evaluation

Appendix 2

Sentiment Analysis



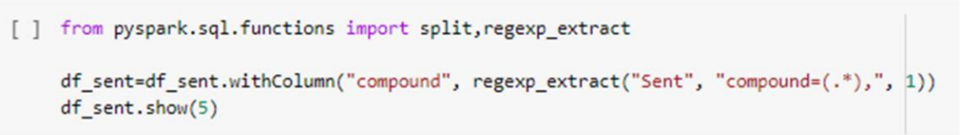
```
[ ] from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
```

```
from pyspark.sql.functions import udf, col
from pyspark.sql.types import StringType

my_udf_sent = udf(lambda x: SentimentIntensityAnalyzer().polarity_scores(x))

df_sent = df_qq.withColumn("Sent",my_udf_sent(col("body")))
df_sent.show(5)
```

Figure 1: Calculating the polarity scores using Vadar Sentiment Analysis

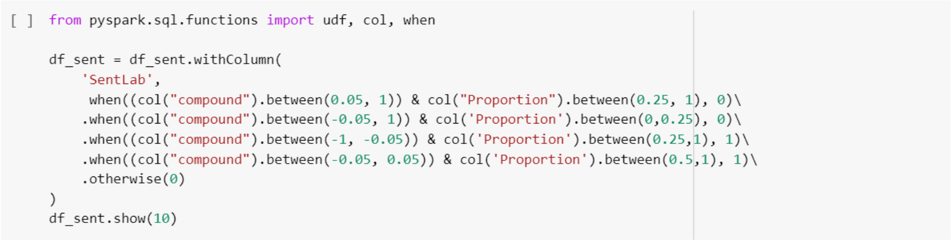


```
[ ] from pyspark.sql.functions import split,regexp_extract

    df_sent=df_sent.withColumn("compound", regexp_extract("Sent", "compound=(.*),", 1))
    df_sent.show(5)
```

Figure 2: Code for extracting the compound from the Vader Sentiment analysis

Labeling Algorithm

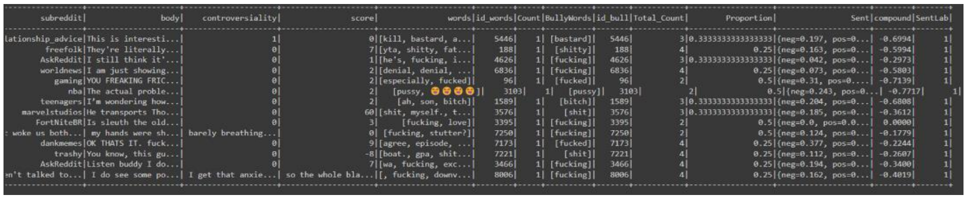


```
[ ] from pyspark.sql.functions import udf, col, when

    df_sent = df_sent.withColumn(
        'SentLab',
         when((col("compound").between(0.05, 1)) & col("Proportion").between(0.25, 1), 0)\
        .when((col("compound").between(-0.05, 1)) & col('Proportion').between(0,0.25), 0)\
        .when((col("compound").between(-1, -0.05)) & col('Proportion').between(0.25,1), 1)\
        .when((col("compound").between(-0.05, 0.05)) & col('Proportion').between(0.5,1), 1)\
        .otherwise(0)
    )
    df_sent.show(10)
```

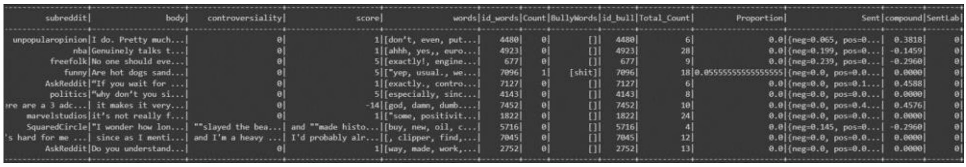
Figure 3: Labelling algorithm that classifies as cyberbullying or non-cyberbullying

Output



```
--------------------+--------------------+--------------------+--------------------+--------------------+--------+-----+----------+-------+-----------+------------------+--------------------+--------+-------+
           subreddit|                body|    controversiality|               score|               words|id_words|Count|BullyWords|id_bull|Total_Count|        Proportion|                Sent|compound|SentLab|
--------------------+--------------------+--------------------+--------------------+--------------------+--------+-----+----------+-------+-----------+------------------+--------------------+--------+-------+
   lationship_advice|This is interesti...|                   1|                   0|[kill, bastard, a...|    5446|    1| [bastard]|   5446|          3|0.3333333333333333|{neg=0.197, pos=0...| -0.6994|      1|
            freefolk|They're literally...|                   0|                   7|[yta, shitty, fat...|     188|    1|  [shitty]|    188|          4|              0.25|{neg=0.163, pos=0...| -0.5994|      1|
           AskReddit|I still think it'...|                   0|                   1|[he's, fucking, i...|    4626|    1| [fucking]|   4626|          3|0.3333333333333333|{neg=0.042, pos=0...| -0.2973|      1|
           worldnews|I am just showing...|                   0|                   2|[denial, denial, ...|    6836|    1| [fucking]|   6836|          4|              0.25|{neg=0.073, pos=0...| -0.5803|      1|
              gaming|YOU FREAKING FRIC...|                   0|                   2|[especially, fucked]|      96|    1|  [fucked]|     96|          2|               0.5|{neg=0.31, pos=0....| -0.7139|      1|
                 nba|The actual proble...|                   0|                   2|   [pussy, 😍😍😍😍]|    3103|    1|   [pussy]|   3103|          2|               0.5|{neg=0.243, pos=0...| -0.7717|      1|
           teenagers|I’m wondering how...|                   0|                   2|    [ah, son, bitch]|    1589|    1|   [bitch]|   1589|          3|0.3333333333333333|{neg=0.204, pos=0...| -0.6808|      1|
       marvelstudios|He transports Tho...|                   0|                  60|[shit, myself., t...|    3576|    1|    [shit]|   3576|          3|0.3333333333333333|{neg=0.185, pos=0...| -0.3612|      1|
          FortNiteBR|Is sleuth the old...|                   0|                   3|     [fucking, love]|    3395|    1| [fucking]|   3395|          2|               0.5|{neg=0.0, pos=0.0...|  0.0000|      1|
: woke us both...| my hands were sh...| barely breathing...|                   0| [fucking, stutter?]|    7250|    1| [fucking]|   7250|          2|               0.5|{neg=0.124, pos=0...| -0.1779|      1|
           dankmemes|OK THATS IT. fuck...|                   0|                   9|[agree, episode, ...|    7173|    1|  [fucked]|   7173|          4|              0.25|{neg=0.377, pos=0...| -0.2244|      1|
              trashy|You know, this gu...|                   0|                  -8|[boat., gpa, shit...|    7221|    1|    [shit]|   7221|          4|              0.25|{neg=0.112, pos=0...| -0.2607|      1|
           AskReddit|Listen buddy I do...|                   0|                   7|[wa, fucking, exc...|    3466|    1| [fucking]|   3466|          4|              0.25|{neg=0.194, pos=0...| -0.3400|      1|
en't talked to...| I do see some po...| I get that anxie...| so the whole bla...|[, fucking, downv...|    8006|    1| [fucking]|   8006|          4|              0.25|{neg=0.162, pos=0...| -0.4019|      1|
--------------------+--------------------+--------------------+--------------------+--------------------+--------+-----+----------+-------+-----------+------------------+--------------------+--------+-------+
```

Figure 4: Output of comments labeled as cyberbullying (1)



```
--------------------+--------------------+--------------------+--------------------+--------------------+--------+-----+----------+-------+-----------+------------------+--------------------+--------+-------+
           subreddit|                body|    controversiality|               score|               words|id_words|Count|BullyWords|id_bull|Total_Count|        Proportion|                Sent|compound|SentLab|
--------------------+--------------------+--------------------+--------------------+--------------------+--------+-----+----------+-------+-----------+------------------+--------------------+--------+-------+
    unpopularopinion|I do. Pretty much...|                   0|                   1|[don’t, even, put...|    4480|    0|        []|   4480|          6|               0.0|{neg=0.065, pos=0...|  0.3818|      0|
                 nba|Genuinely talks t...|                   0|                   1|[ahhh, yes,, euro...|    4923|    0|        []|   4923|         28|               0.0|{neg=0.199, pos=0...| -0.1459|      0|
            freefolk|No one should eve...|                   0|                   5|[exactly!, engine...|     677|    0|        []|    677|          9|               0.0|{neg=0.239, pos=0...| -0.2960|      0|
               funny|Are hot dogs sand...|                   0|                   5|["yep, usual., we...|    7096|    1|    [shit]|   7096|         18|0.0555555555555555|{neg=0.0, pos=0.0...|  0.0000|      0|
           AskReddit|“If you wait for ...|                   0|                   1|[exactly., contro...|    7127|    0|        []|   7127|          6|               0.0|{neg=0.0, pos=0.1...|  0.4588|      0|
            politics|“why don’t you si...|                   0|                   5|[especially, sinc...|    4143|    0|        []|    4143|          8|               0.0|{neg=0.0, pos=0.0...|  0.0000|      0|
ere are a 3 adc...| it makes it very...|                   0|                 -14|[god, damn, dumb....|    7452|    0|        []|   7452|         10|               0.0|{neg=0.0, pos=0.4...|  0.4576|      0|
       marvelstudios|it’s not really f...|                   0|                   1|["some, positivit...|    1822|    0|        []|   1822|         24|               0.0|{neg=0.0, pos=0.0...|  0.0000|      0|
       SquaredCircle|"I wonder how lon...| ""slayed the bea...| and ""made histo...|[buy, new, oil, c...|    5716|    0|        []|   5716|          4|               0.0|{neg=0.145, pos=0...| -0.2960|      0|
's hard for me ...| since as I menti...| and I'm a heavy ...| I'd probably alr...|[, clipper, find,...|    7045|    0|        []|   7045|         12|               0.0|{neg=0.0, pos=0.0...|  0.0000|      0|
           AskReddit|Do you understand...|                   0|                   1|[way, made, work,...|    2752|    0|        []|   2752|         13|               0.0|{neg=0.0, pos=0.0...|  0.0000|      0|
--------------------+--------------------+--------------------+--------------------+--------------------+--------+-----+----------+-------+-----------+------------------+--------------------+--------+-------+
```

`

Figure 5: Output of comments labeled as non-cyberbullying (0)