

NAME OF THE PROJECT

MALIGNANT COMMENTS CLASSIFICATION

Submitted by: HARSHITA SEHGAL

ACKNOWLEDGMENT

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

Various websites and publish papers helped me solve and understand this problem in depth.

<https://medium.com/towards-artificial-intelligence/understanding-multi-label-classification-model-and-accuracy-metrics-1b2a8e2648ca>

<https://people.iee.ihu.gr/~stoug/odep/papers/Multi-Label%20Classification:%20An%20Overview.pdf>

<https://www.analyticsvidhya.com/blog/2017/08/introduction-to-multi-label-classification/>

INTRODUCTION

Business Problem Framing

Describe the business problem and how this problem can be related to the real world.

Herein we have bunch of comments in train as well as test dataset,our aim is to classify them according to their toxicity,It is multilabel classification problem.In the era of digital world wherein everyone has the freedom to say whatever they wish to ,but often this freedom leads to comments full of hatred that can have a very disturbing impact on one’s mental health .So to limit/curb the hate received on various platforms we have to first identify the different types of hate/toxic comments so a balance can be made

Conceptual Background of the Domain Problem

Describe the domain related concepts that you think will be useful for better understanding of the project.

This is basically a text classification problem,the domain herein is a mix of nlp and classification( machine learning technique),a brief about these domains is given below.

NLP-NLP aims at converting unstructured data into computer-readable language by following attributes of natural language. Machines employ complex algorithms to break down any text content to extract meaningful information from it. The collected data is then used to further teach machines the logics of natural language.

Machine learning-Machine learning is perhaps the principal technology behind two emerging domains: data science and artificial intelligence. The rise of machine learning is coming about through the availability of data and computation, but machine learning methdologies are fundamentally dependent on models.

The emergence of machine learning is closely tied to the emergence of widely available data.

Classification -In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data.This is a multi class classification problem,we have 6 labels/target variables,the methodology and the concepts will vary from a regular classification problem.They are discussed further in this report.

Multi class classification

Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

Firstly the data present is purely text, and 6 labels that have binary values 0 denoting this is not a toxic comment present and 1 denoting it is toxic comment. So to understand the patterns, no of comments for each toxic comment type I did EDA ,I did a research on what visualizations can be done so that I can understand my dataset Next step was to clean my data, again I did some Research what can be removed how it can be removed .Now text cannot be directly processed into the model, tf idf techniques, splitting the data into test and train also required a detailed attention. And the most important step was building the model, selecting the algorithms, So a detailed research was done to understand these algorithms. Also these algorithms were different from the traditional algorithms because it was a multi class classification problem.

Motivation for the Problem Undertaken

Describe your objective behind to make this project, this domain and what is the motivation behind.

Nowadays, the flow of data over the internet has grown dramatically, especially with the appearance of social networking sites. Due to this, an important task now is the development of algorithms to automatically classify the social networks content as "positive" or "negative", in order to prevent possible harm to the society. In recent years there have been many cases in which authorities arrested some users of social sites because of negative (abusive) content of their personal pages. For example, one Man in Thailand was jailed for 35 years for insulting monarchy on Facebook

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem

Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

We will be using multi label classification algorithms but the main issue is that we cannot rely simply on accuracy score ,thereby the use of loss functions in this problem is necessary.

A classifier supposed to be good for solving one of those problems may perform poorly for another problem. In order to facilitate the analysis, we restrict ourselves to two loss functions, namely the Hamming and the subset 0/1 loss. The first one is representative of the single label scenario, while the second one is a typical multi-label loss function whose minimization calls for an estimation of the joint distribution. Our analysis proceeds from the simplifying assumption of an unconstrained hypothesis space, which allows us to consider the conditional distribution for a given x. As such, this theoretical analysis will differ from the experimental analysis reported in Sect. [7](https://link.springer.com/article/10.1007/s10994-012-5285-8#Sec16), where parametric hypothesis spaces are considered. Despite this conceptual difference, our theoretical and experimental results will be highly consistent. While the theoretical analysis mainly provides evidence on the population level, the empirical study also investigates the effect of estimation.

the Hamming loss minimizer and the subset 0/1 loss minimizer will differ significantly. That is, the Hamming loss minimizer may be poor in terms of the subset 0/1 loss and vice versa. In some (not necessarily unrealistic) situations, however, the Hamming and subset 0/1 loss minimizers coincide, an observation that may explain some misleading results in recent Multi label classification problem,But because our predictions were sparse matrixes,so there was trouble calculating loss,other approach is auc roc which will be discussed further

Data Sources and their formats

What are the data sources, their origins, their formats and other details that you find necessary? They can be described here.

Provide a proper data description. You can also add a snapshot of the data.

We have data shape of (159571, 8)

159571 rows and 8 columns for train dataset(id,comment text,6 labels)

We have 153164 rows 2 columns in test(id and comment text)

Avg Length of the comments was 400,maximum length 5000.

Data Preprocessing Done

What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

Next step was to clean my data,three major steps in this regard are-

first convert the comments to lower-case

Removal of punctuation-html-tags, punctuation and non-alphabetic characters by df2['comment\_text']=df2['comment\_text'].str.replace(r'[?|!|\'|"|#]', "")

df2['comment\_text']=df2['comment\_text'].str.replace(r'[.|,|)|(|\|/|-|~|\*]', "")

Removal of stopwords-Stop words are basically a set of commonly used words in any language, not just English. The reason why stop words are critical to many applications is that, if we remove the words that are very commonly used in a given language, we can focus on the important words instead

Stemming -Next we do stemming. There exist different kinds of stemming which basically transform words with roughly the same semantics to one standard form. For example, for amusing, amusement, and amused, the stem would be amus.

Tf-id -One technique is to pick the most frequently occurring terms (words with high term frequency or tf). However, the most frequent word is a less useful metric since some words like ‘this’, ‘a’ occur very frequently across all documents.

Hence, we also want a measure of how unique a word is i.e. how infrequently the word occurs across all documents (inverse document frequency or idf).

So, the product of tf & idf (TF-IDF) of a word gives a product of how frequent this word is in the document multiplied by how unique the word is w.r.t. the entire corpus of documents

Data Inputs- Logic- Output Relationships & Approach

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

1 input (comment\_text) and multi outputs-

["malignant","highly\_malignant","rude","threat","abuse","loathe"]

Basically this can be treated as several models for each label, but that will not tell us the type of a particular toxic comment, we will be able to only classify whether it’s toxic or not. But what we actually have to do is predict the type of a comment whether it lies in the above category , so a multilabel classification approach is used here.

State the set of assumptions (if any) related to the problem under consideration

Here, you can describe any presumptions taken by you.

While performing the multi class classification algorithms,most of the algorithms ran out of memory,so In few of my algorithms I used a small dataset limiting the size from the whole dataset.

Hardware and Software Requirements and Tools Used

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

Firstly I saved the excel file as a csv file then uploaded it on jupyter notebook and

LIBRARIES USED

Numpy,pandas,matplotlib,seaborn

Apart from these basic libraries,other which were used

from skmultilearn.problem\_transform import BinaryRelevance

from sklearn.multiclass import OneVsRestClassifier

from skmultilearn.problem\_transform import ClassifierChain

from skmultilearn.problem\_transform import LabelPowerset

from skmultilearn.adapt import MLkNN

Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods)

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

1. OneVsRest

Traditional two-class and multi-class problems can both be cast into multi-label ones by restricting each instance to have only one label. On the other hand, the generality of multi-label problems inevitably makes it more difficult to learn. An intuitive approach to solving multi-label problem is to decompose it into multiple independent binary classification problems (one per category).

In an “one-to-rest” strategy, one could build multiple independent classifiers and, for an unseen instance, choose the class for which the confidence is maximized.

The main assumption here is that the labels are mutually exclusive. You do not consider any underlying correlation between the classes in this method.

For instance, it is more like asking simple questions, say, “is the comment toxic or not”, “is the comment threatening or not?”, etc. Also there might be an extensive case of overfitting here, since most of the comments are unlabeled, i,e., most of the comments are clean comments as observed in visualization out of a lakh comments only around 40,000 were actually some sort of toxic comments,rest were clean

2. Binary Relevance

In this case an ensemble of single-label binary classifiers is trained, one for each class. Each classifier predicts either the membership or the non-membership of one class. The union of all classes that were predicted is taken as the multi-label output. This approach is popular because it is easy to implement, however it also ignores the possible correlations between class labels.

In other words, if there’s q labels, the binary relevance method create q new data sets from the images, one for each label and train single-label classifiers on each new data set. One classifier may answer yes/no to the question

3. Classifier Chains

A chain of binary classifiers C0, C1, . . . , Cn is constructed, where a classifier Ci uses the predictions of all the classifier Cj , where j < i. This way the method, also called classifier chains (CC), can take into account label correlations.

The total number of classifiers needed for this approach is equal to the number of classes, but the training of the classifiers is more involved

4. Label Powerset

This approach does take possible correlations between class labels into account. More commonly this approach is called the label-powerset method, because it considers each member of the power set of labels in the training set as a single label.

This method needs worst case (2^|C|) classifiers, and has a high computational complexity.

However when the number of classes increases the number of distinct label combinations can grow exponentially. This easily leads to combinatorial explosion and thus computational infeasibility. Furthermore, some label combinations will have very few positive examples.

Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing.

In continuation to the above possible methods,these 4 algorithms are basically problem transformation methods and algorithm adaptation methods, Problem transformation methods transform the multi-label problem into a set of binary classification problems, which can then be handled using single-class classifiers.

Whereas algorithm adaptation methods adapt the algorithms to directly perform multi-label classification. In other words, rather than trying to convert the problem to a simpler problem, they try to address the problem in its full form

Another algorithm used was-

Algorithm adaptation methods for multi-label classification concentrate on adapting single-label classification algorithms to the multi-label case usually by changes in cost/decision functions.

Here we use a multi-label lazy learning approach named ML-KNN which is derived from the traditional K-nearest neighbor (KNN) algorithm.

The [skmultilearn.adapt](http://scikit.ml/api/api/skmultilearn.adapt.html" \l "module-skmultilearn.adapt) module implements algorithm adaptation approaches to multi-label classification, including but not limited to ML-KNN.

Run and Evaluate selected models

Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

Binary Relevance results

Final score = > 0.9679

precision recall f1-score support

0 1.00 0.87 0.93 971

1 1.00 0.58 0.74 101

2 0.99 0.87 0.93 527

3 1.00 0.48 0.65 33

4 0.97 0.82 0.89 494

5 0.98 0.49 0.65 84

micro avg 0.99 0.83 0.90 2210

macro avg 0.99 0.69 0.80 2210

weighted avg 0.99 0.83 0.90 2210

samples avg 0.09 0.08 0.08 2210

2. OneVsRestClassifier

\*\*toxicity type- malignant comments...\*\*

Test accuracy is 0.8926666666666667

\*\*toxicity type- highly\_malignant comments...\*\*

Test accuracy is 0.986

\*\*toxicity type- rude comments...\*\*

Test accuracy is 0.9353333333333333

\*\*toxicity type- threat comments...\*\*

Test accuracy is 0.9966666666666667

\*\*toxicity type- abuse comments...\*\*

Test accuracy is 0.946

\*\*toxicity type- loathe comments...\*\*

Test accuracy is 0.992

3. ClassifierChain

Accuracy = 0.8886666666666667

4. LabelPowerset

Accuracy = 0.892

Accuracy = 0.892

results with naive bayes

5.mlknn(adapted algorithm)

Accuracy = 0.892

Key Metrics for success in solving problem under consideration

label-powerset method scores best, followed by the one-against-all method,Also binary classifier gives us good results but while calculating test accuracy and train accuracy I encountered overfitting.

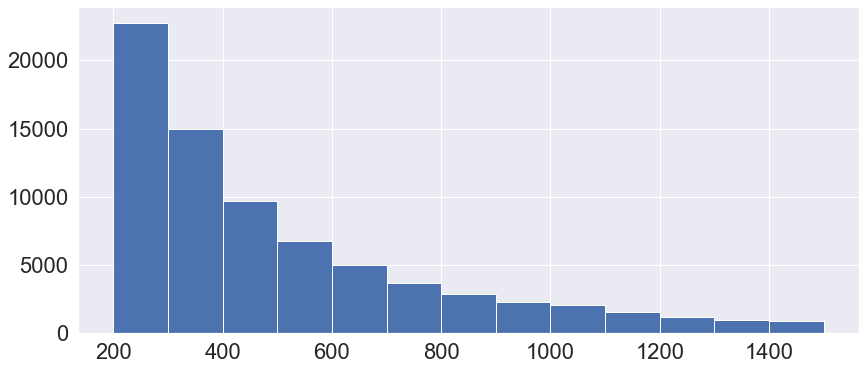
What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

Visualizations

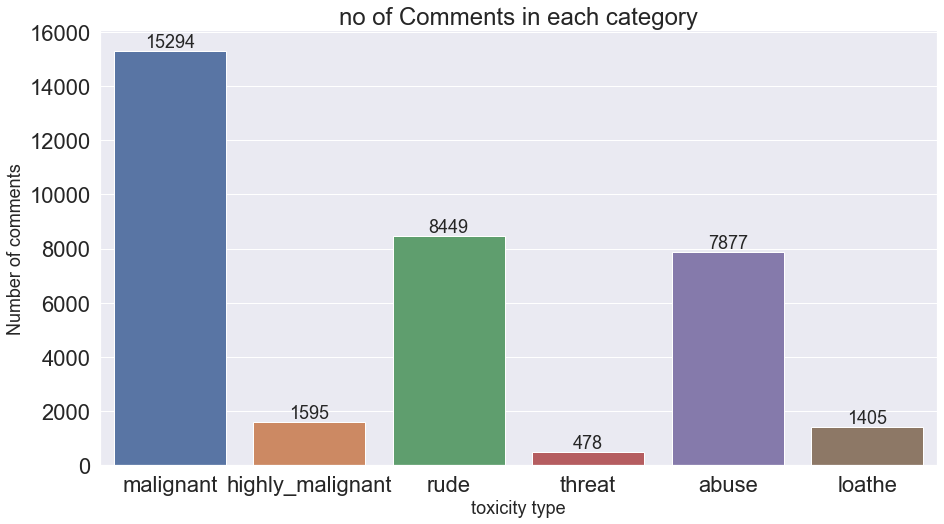
Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

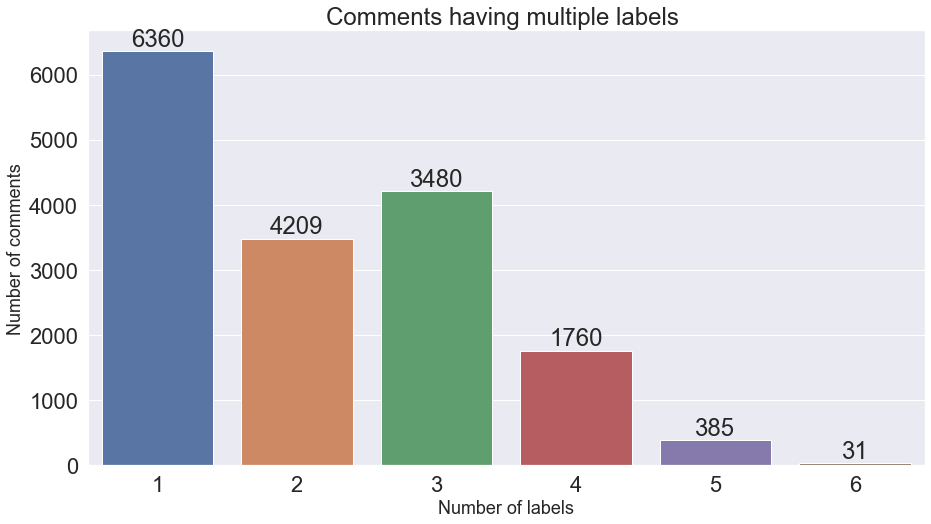
If different platforms were used, mention that as well.

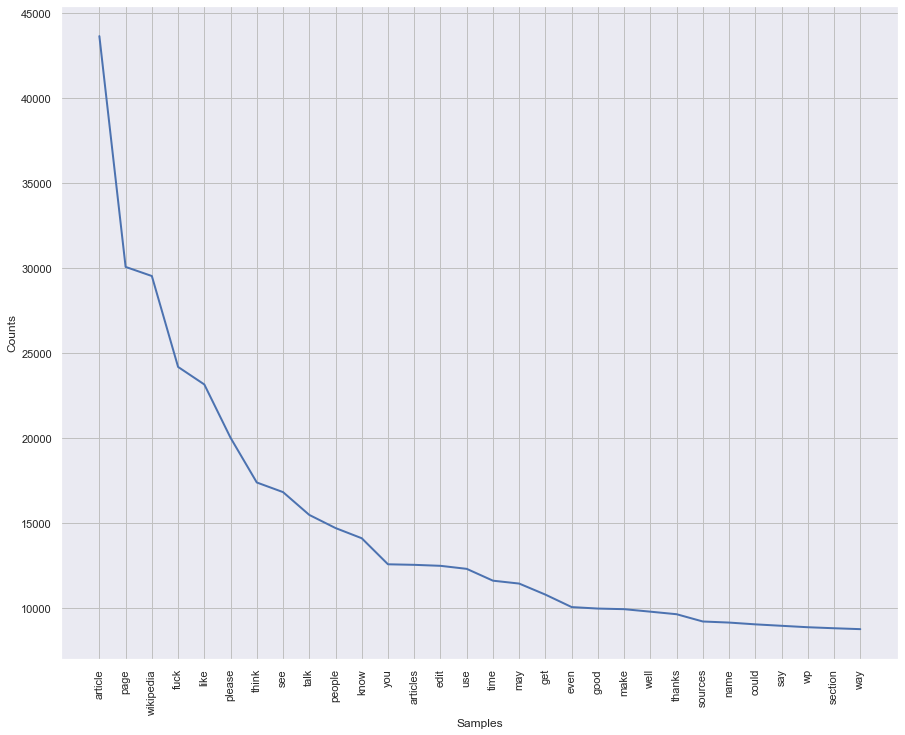
I plotted a histplot to see the length of comments



Then I plotted subplots for word cloud for each category of comment

Then using sum function across each label column I got to know the no of comments in each category

Then doing counts across the rows, I observed the labels having multiple columns

FDIST of most frequent words from test dataframe comment\_text column

Interpretation of the Results

Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

I observed that an average length of comment is short ,some comments even have length of 2,5 .considering the whole dataset average length was 400

Then while plotting word clouds,I observed that there was no significant word which would make a clear distinction among the categories,all the categories had cuss words and infact majority had similar cuss words,but in case of threat there were words like die die so that was the only difference I observed among all the categories

Next I observed there are many rows which have no toxic comment,around 4000 are toxic comments Rest are clean

Also while plotting the fdist I observed words like page,article, Wikipedia .

CONCLUSION

Key Findings and Conclusions of the Study

Describe the key findings, inferences, observations from the whole problem.

Problem transformation methods transform the multi-label problem into a set of [binary classification](https://en.wikipedia.org/wiki/Binary_classification) problems, which can then be handled using single-class classifiers.

Whereas algorithm adaptation methods adapt the algorithms to directly perform multi-label classification. In other words, rather than trying to convert the problem to a simpler problem, they try to address the problem in its full form.

In an extensive comparison with other approaches, label-powerset method scores best, followed by the one-against-all method.

Learning Outcomes of the Study in respect of Data Science

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

Visualization observations

I observed that an average length of comment is short ,some comments even have length of 2,5 .considering the whole dataset average length was 400

Then while plotting word clouds,I observed that there was no significant word which would make a clear distinction among the categories,all the categories had cuss words and infact majority had similar cuss words,but in case of threat there were words like die die so that was the only difference I observed among all the categories

Next I observed there are many rows which have no toxic comment,around 4000 are toxic comments Rest are clean

Also while plotting the fdist I observed words like page,article, Wikipedia .

Here are the roc auc scores for the algorithms used,

LABEL POWERSET

0.5

ClassifierChain

0.4987570551782572

BINARY RELEVANCE

0.5005978018404166

Limitations of this work and Scope for Future Work

What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.

Time and memory issues

Majority of algorithms took considerable amount of time when run on this dataset, so experimentation was done on a random sample of the train data.

Further improvements:

The same problem can be solved using LSTMs in deep learning.

For more speed we could use decision trees and for a reasonable trade-off between speed and accuracy we could also opt for ensemble models