Exploratory Data Analysis of YouTube Comments

Our project is filtering spam comments from YouTube videos. We will be using two datasets. Dataset1 is extracted from [1] while Dataset 2, is extracted through **YouTube API V3** using a JavaScript program. We will be using dataset 1 as training dataset and dataset 2 as testing dataset. The first dataset contains five csv files, as shown in the table 1.

CSV FILE	YOUTUBE ID	INVALID COMMENTS	VALID COMMENTS
PSY	9bZkp7q19f0	175	175
KATY PERRY	CevxZvSJLk8	175	175
LMFAO	KQ6zr6kCPj8	236	202
EMINEM	uelHwf8o7_U	245	203
SHAKIRA	pRpeEdMmmQ0	174	196

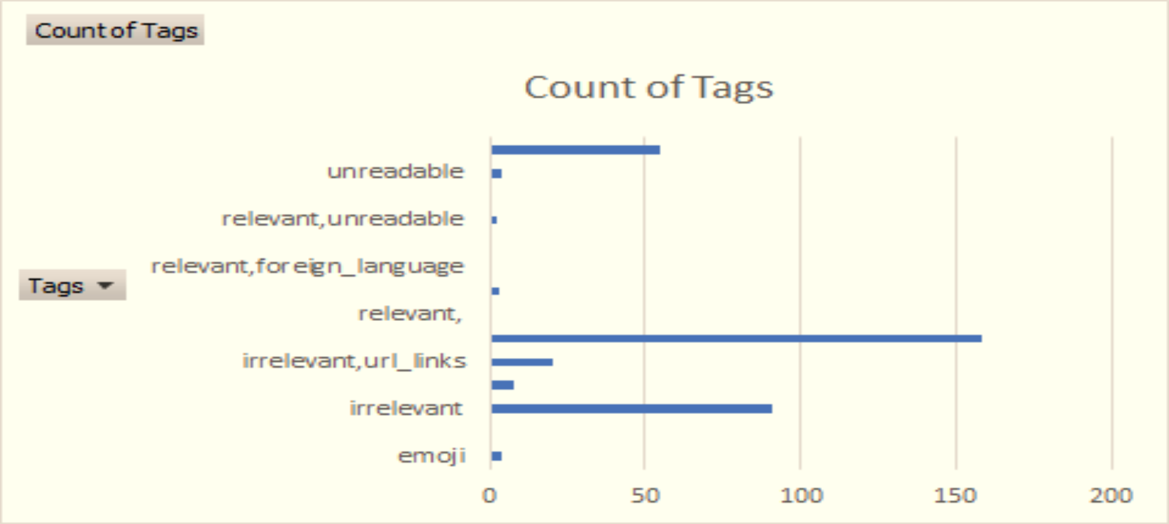
The structure of all the csv files that we have downloaded is as follows:
a)Comment ID- the unique Id of the comment,
b)Author-username of vlogger,
c)Date- Date when the comment was posted,
d)Content- The content of the comment(text),
e)Class- Whether it is valid comment or invalid.

Apart from this we have added one more column -
f)Tags- where we have added tags to those comments such as Irrelevant, relevant, emoji, url links, html tags ,hashtag, unreadable or combinations of these tags.

We have created bar-graphs for all the csv files in Microsoft excel so that data visualization is easier to understand.

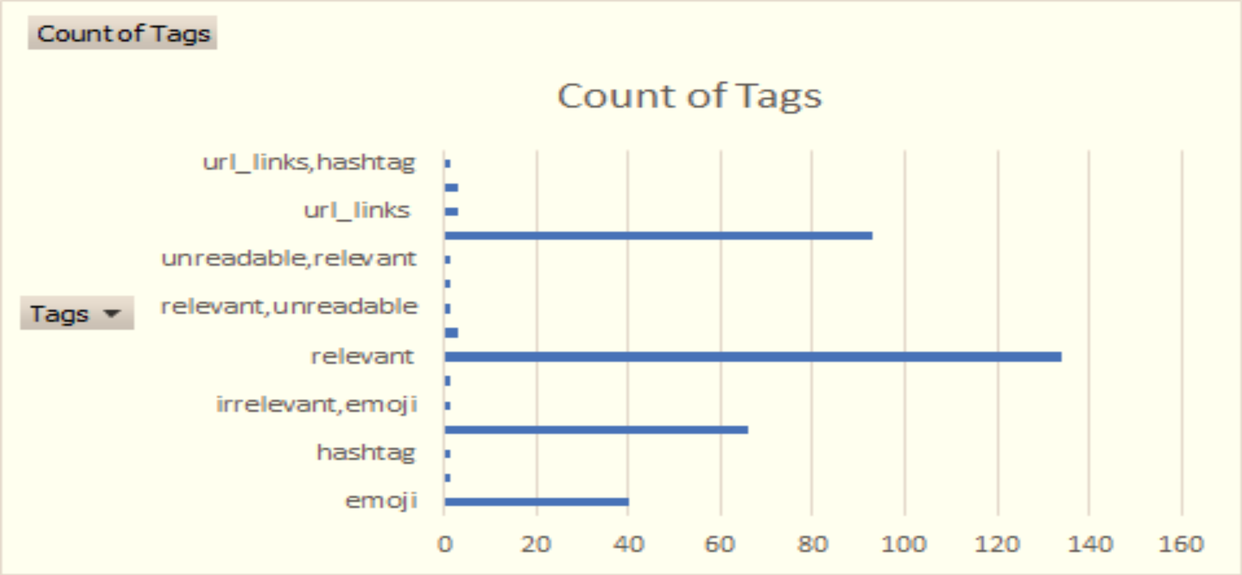
1) PSY.csv

This file has 350 comments. Following is the bar chart of the count of unique tags-



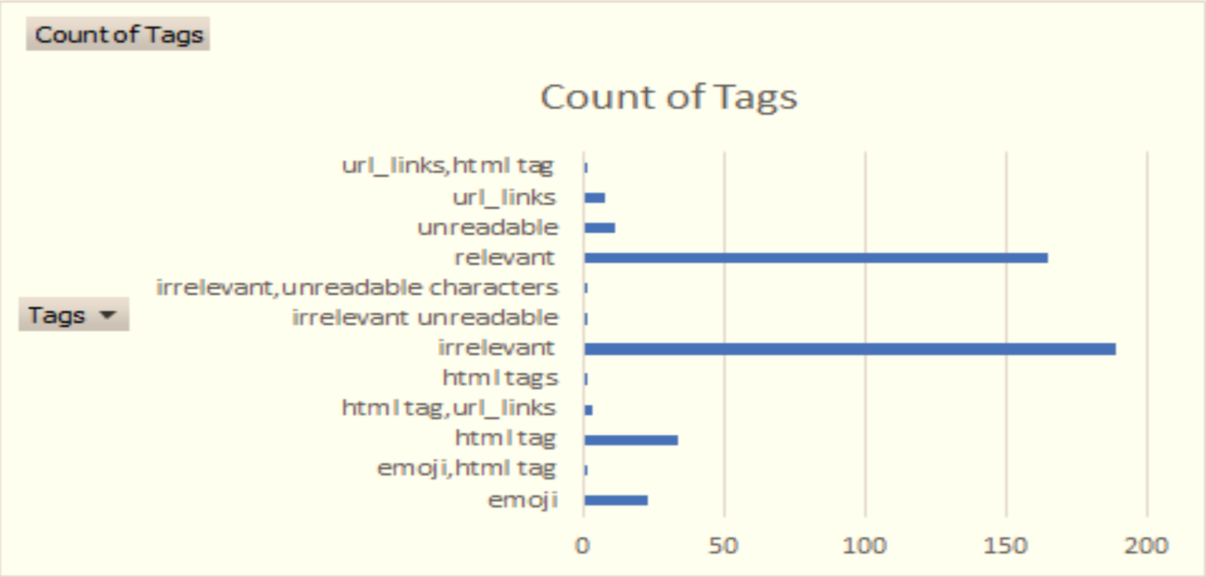
2) KatyPerry.csv

This file has 350 comments. Following is the bar chart of the count of unique tags-



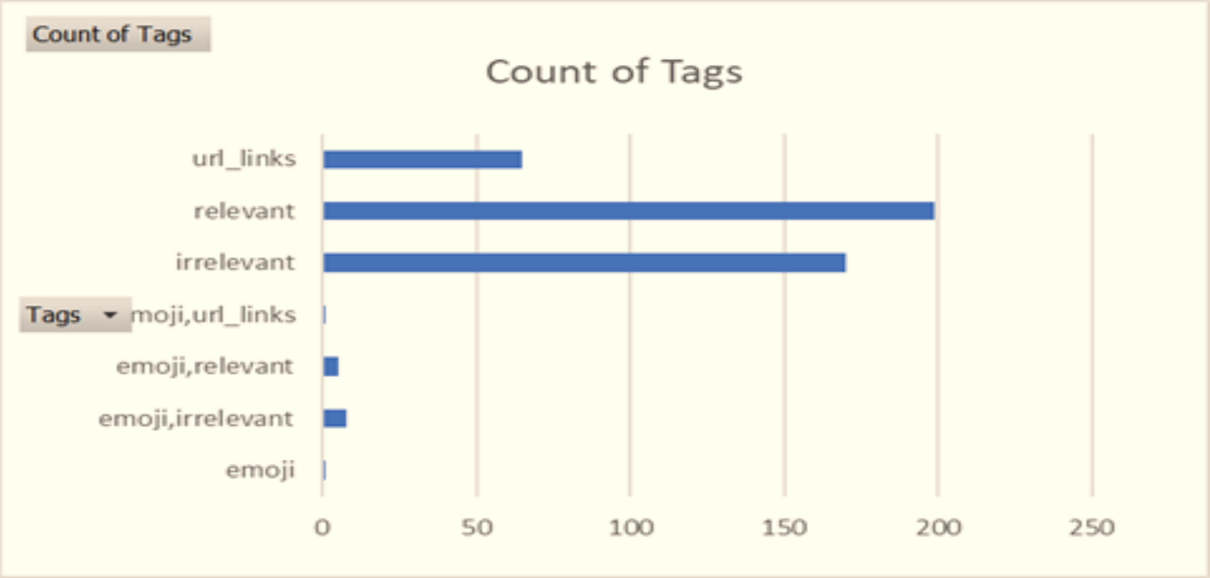
3) LFMAO.csv

This file has 438 comments. Following is the bar chart of the count of unique tags-



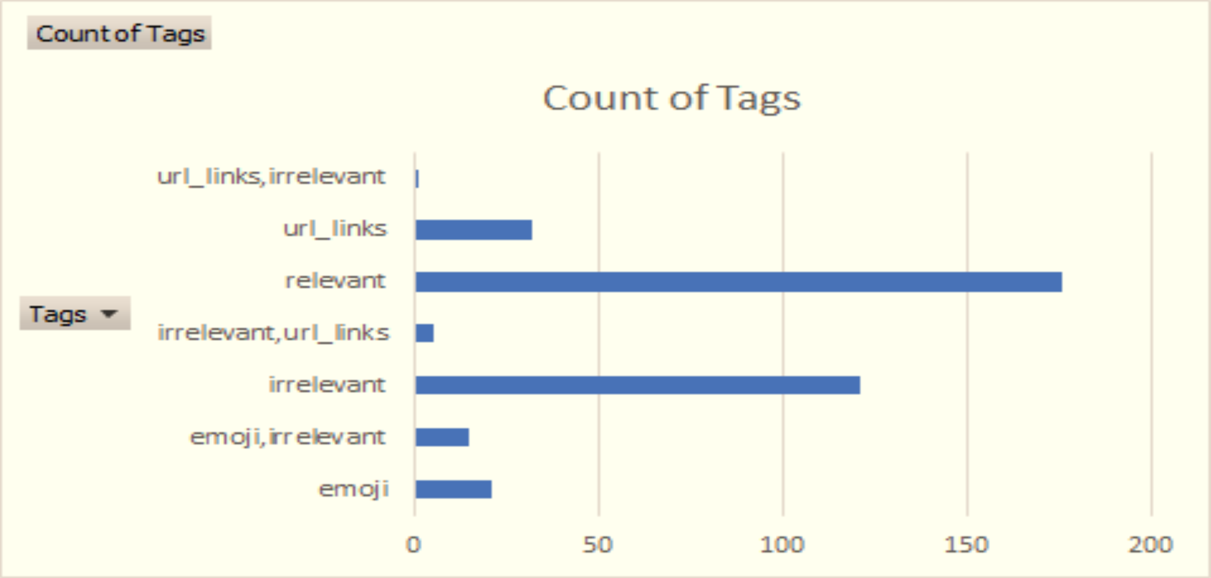
4) Eminem.csv-

This file has 449 comments. Following is the bar chart of the count of unique tags-

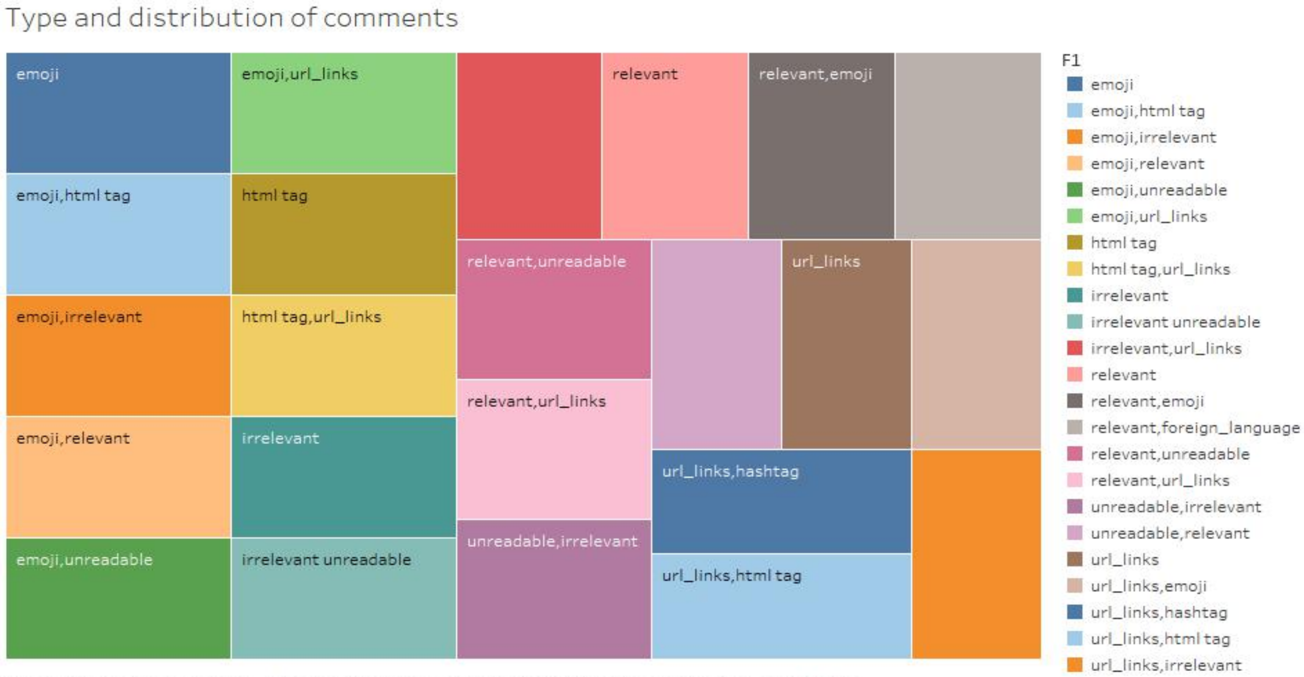


5) Shakira.csv-

This file has 371 comments. Following is the bar chart of the count of unique tags-



Finally ,the following visualization summarizes overall distribution of type of comments on YouTube videos:



F1. Color shows details about F1. Size shows distinct count of Number of Records. The marks are labeled by F1.

Works Cited

1. Alberto, T., J.V., L., & Almeida, T. (2015). TubeSpam: Comment Spam Filtering on YouTube. *Proceedings of the 14th IEEE International Conference on Machine Learning and Applications (ICMLA'15).* FL. From http://dcomp.sor.ufscar.br/talmeida/youtubespamcollection/