Spam Detection in YouTube Comments

ABSTRACT

Social media networks have become a place for people to share their views, opinions and daily activities. However, social spam such as links to low quality content, malicious content, negative and vulgar comment can also be found on these platforms. This content can harm the experience of users such as content producers and other followers. So, it became necessary to filter such content using efficient spam detection techniques. In this work, we chose YouTube, a popular video sharing platform, as our study target because YouTube offers limited tools for comment moderation, and this increases the spam volume which lead vloggers (channel owners) of famous channels to disable their comments. Even for established classification methods it is a challenge since the messages are very short and often contains symbols, slangs and abbreviations. We have decided to evaluate the comments that are present in most viewed videos on YouTube channels and will use machine learning techniques such Naïve Bayes and Decision tree to identify spam comments and filter them.

General Terms

Decision Tree, Naïve Bayes, Stemming, Tokenization, Java, You Tube, Spam classification, Social Network analysis, YouTube.

1. INTRODUCTION

The popularization of internet and smart devices has commercialized online social networking sites such as Facebook, Twitter, Instagram, YouTube etc. among users. It became commonplace action to post daily activities, social gatherings, photos and videos on these social networks. YouTube has become popular over the years due to its content-publication platform [1], as it allows users and video producers to share their ideas and views and provide a stage to interact with each other using various features such as comments, Like/Dislike and share.

YouTube is a leading video sharing platform and is well ahead of its competitors such as Amazon Prime Video, Netflix and Hulu. Consumers spent nearly 9.5 billion hours on YouTube mobile app [2]. Due to the growing popularity of its platform, YouTube started monetization program to reward its video producers to encourage them to generate more good quality content. This lead to YouTube being inundated by videos having low quality content known as spam. Users began exploiting this program by posting spam videos and sharing their own video links in comments section of most popular videos in the hope of increasing the views. Due to large user, base millions of advertisers use YouTube for promoting their products and advertising is major revenue source for YouTube. The problem caused by social spam hurts the quality of content and platform’s community.

The established techniques for automatic spam filtering were not efficient enough since comments on You Tube are generally very short and widespread with slangs, idioms, emoticons, abbreviations and symbols which make even tokenization a challenging job.

It is necessary to establish an automatic spam filtering system for eliminating such spam comment on YouTube and enhance the quality of content shared on this platform.

2. PROBLEM STATEMENT

YouTube is one of the most popular social media platform. Millions of users visit this website to post, find and watch videos of different genres. The social video sharing platform provided by the YouTube allow online users to interact, to express their opinions and to read opinions of other users. Many users post comments on YouTube videos to express their views, share ideas about the video and to interact with other users. Even though the comment section in videos is a vital part of the YouTube community, it is also a place used for spamming, trolling and cyber-bullying of users by posting negative comments. Reading such spam comments is a waste of time for online users and can also cause damage to the reader. YouTube provides users with tools to review comments posted on their videos, flag such comments, and mark whether they are spam or not. However, it might not be possible for a user to review all the comments. But, it is possible to identify such spam comments and eliminate them by using machine learning methods and topic detection. Our goal is to develop a system which can detect spam or irrelevant comments and filter them. This system would help to improve the quality of comments posted in YouTube community.

3. RELATED WORK

The initial research on spam detection was in email spam detection [3] in late nineties. In this research work, authors show that by considering domain-specific features of spam problem, they can produce accurate filters. The work which motivated us is [4] , this research work talks about detecting spam comments from YouTube and other blogs by mean of some indicators, like a discontinuous text flow, inadequate language. Their work was based on machine learning algorithms such as decision trees and support vector machine. The authors of this paper have used a very old and small dataset by [5]. Another similar research [1], presented a new public datasets of spam words but these data sets were manually labelled as spam or legitimate words, there was no pre-processing to identify those words, so it is not completely relevant. Also, this work provides an online tool which can automatically detect spam comments posted in YouTube videos but efficiency of the tool cannot be justified.

In our research project, we are trying to prove that spam comments can be detected from YouTube comments with proper feature selection. This will help to identify and filter spam comments from similar online social networks

4. PROJECT GOAL

The goal of this project is to identify spam comments on YouTube videos based on various features such as number of links in the comments, number of white spaces, word duplication, number of non-ASCII characters in the comments etc.

5. PROJECT PLAN

We are planning to build a model out of labeled dataset to detect spam comments. The labeled dataset is in natural language. The experiments we will perform are on YouTube comments by extracting them from YouTube through its API [https://developers.google.com/youtube/v3/]. The extracted comments will be stored in csv files for further processing. The high-level architecture of our system will have two parts as shown in the figure1:

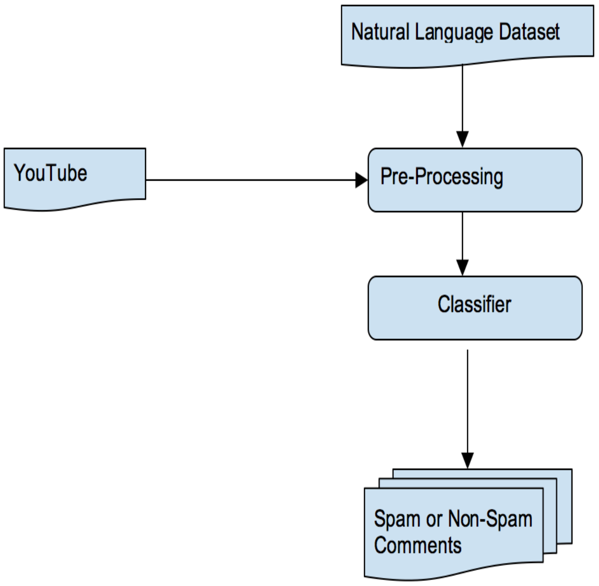

Figure 1: High-level architecture of proposed YouTube spam detection system

5.1 Data Pre-Processing

In the first step, we will pre-process our dataset to remove stop words, number of links, white spaces, non-ASCII Characters, Word duplication. We will use the following methods for data pre-processing:

a) Tokenization
b) Stop word removal
c) Stemming
d) Normalization

5.2 Classifier

In the second step, we will feed our training dataset to the classifier. The classification algorithms we are planning to use are Decision Tree and Naive Bayes. We will be using Python nltk, scikit-learn packages for building the classifier. We will classify comments as spam based on sentence structure, vulgar language and words irrelevant to the post. We will use F1-measure and precision to evaluate classification performance.

5.3 Web application

We are planning to build a web application using HTML, CSS and JavaScript which will be used for searching and retrieving YouTube video posts, comments and filtering providing option for user to filter spam comments.

6. CHALLENGES

The challenges we have evaluated are:

a). Extracting data in csv format from YouTube API.

b). We assume that the YouTube comments will be in English language. It will be a challenge to identify spam comments if they are in any other language.

c). Most of the comments posted by users do not follow proper rules of English grammar. Users often use language used for instant messaging for posting comments. Many comments use words in abbreviated forms or emoticons for expressing views regarding the video. Some comments are very short and are not proper sentences. It will make tokenizing such comments a challenging task.

7. LIMITATIONS

The limitations will be:

a). We can extract only limited data from YouTube channels due to YouTube quota limits [3].

b). For this project, we will only classify comments which are in English language.

8. Deliverables

We will deliver a Web Application – A user interface to search and display YouTube video posts and filtered comments along with source code and documentation.

9. REFERENCES

[1] Alberto, T. C., Lochter, J. V., & Almeida, T. A. (2015). TubeSpam: Comment Spam Filtering on YouTube. Machine Learning and Applications (pp. 138-143). Miami, FL, USA: IEEE.

[2] S. Perez, "YouTube’s app is dominating mobile video by monthly users, time spent", TechCrunch, 2017

[3] Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. In AAAI-98 Workshop on Learning for Text Categorization (pp. 98-105). Madison, Wisconsin: AAAI.

[4] Rădulescu, C., Dinsoreanu, M., & Potolea, R. (2014). Identification of Spam Comments using Natural Language Processing Techniques. Intelligent

Computer Communication and Processing (ICCP) (pp. 29-35). Cluj Napoca, Romania: IEEE.

[5] Mishne, G., Carmel, D., & Lempel, R. (2005). Blocking blog spam with language model disagreement. irst International Workshop on Adversarial Information Web (AIRWeb) (pp. 1-6). Chiba, Japan: IEEE.