Analysis of YouTube of Videos: A Literature Survey

Neha Reddy
4th year student, Dept. of CSE
BMS College of Engineering
Bengaluru, Karnataka
1BM16CS057@bmsce.ac.in

Piyush Gupta
4th year student, Dept. of CSE
BMS College of Engineering
Bengaluru, Karnataka
1BM16CS066@bmsce.ac.in

Prasham Mehta
4th year student, Dept. of CSE
BMS College of Engineering
Bengaluru, Karnataka
1BM16CS070@bmsce.ac.in

Puneet Gupta
4th year student, Dept. of CSE
BMS College of Engineering
Bengaluru, Karnataka
1BM16CS072@bmsce.ac.in

Vikranth BM
Asst. Professor, Dept. of CSE
BMS College of Engineering
Bengaluru, Karnataka
vikranthbm.cse@bmsce.ac.in

***Abstract*— Consumption of content from YouTube** (Lanyu Shang, 2019) **and other OTT(over-the-top) platforms is constantly increasing. YouTube** (Lanyu Shang, 2019) **being a source of education, entertainment and promotion, is a very lucrative platform. YouTubers tend to unethically attract viewers into clicking their video by manipulating their title and/or thumbnail. In this paper we present a method to train a model to classify a video as Clickbait** (Lanyu Shang, 2019) **video or non-Clickbait** (Lanyu Shang, 2019) **video.**

***Keywords*—Clickbait, YouTube (Lanyu Shang, 2019) [1], Comments, Title, Thumbnail**

I. INTRODUCTION

YouTube is becoming a major resource for sharing and consuming video content. It is gaining immense popularity and support from viewer community due to its comprehensive repository of videos. Also, it supports diversity by having different facets such as modals, languages, domains and cultures. For a YouTube (Lanyu Shang, 2019) content developer or a YouTuber with various notable channels, (Lanyu Shang, 2019) this is a profession with a lot of monetary potential. The younger generations are recently shifting to YouTube (Lanyu Shang, 2019) and other OTT platforms, away from the traditional television.

A YouTube (Lanyu Shang, 2019) video often consists of a title, thumbnail, video content along with other non-video features. Despite it being unethical, content developers deliberately manipulate the heading and the thumbnail so as to attract more audience and baiting them into viewing their content. There are quite a few instances when the content of the video mismatches with the heading of the video or the thumbnail of the video. This is known as a Clickbait (Lanyu Shang, 2019)Video. Our aim is to classify a video as to whether it is a Clickbait (Lanyu Shang, 2019) or not. This is critically important as a majority of people spend their time on YouTube (Lanyu Shang, 2019) and not getting what they search for is a waste of their precious time. We use sentiment analysis on viewer comments to identify a video as click bait or not.

is increasingly becoming a major resource for sharing

II. DATASET

We are only working with YouTube (Lanyu Shang, 2019) data that consists of viewer comments. The data is collected with the help of YouTube (Lanyu Shang, 2019) API v3. We created a Google Developer account and generated a key to extract all the details of a video in the form of a JSON file. This dataset contains all the details of the trending YouTube videos along with its likes, dislikes, comments, tags and views for each video for a particular year, which comprises a top-level comment and replies, if any exist, to that comment.

III. METHODOLOGY

We implement our project by dividing the process into 4 modules. They consist of: Network Characteristics Analysis, Linguistic Characteristics Analysis, Metadata Characteristics Analysis and Supervised Classification.

A. *Network Characteristics Analysis*

YouTube (Lanyu Shang, 2019) comments have two level thread structure. It has a top-level comment node and another level is replies to that node. This module analyses the comment threads and extracts their semantic and topological structure. We define the network of user comments as a directed graph where we represent each comment as a node and each edge denotes a reply from one comment to another. As a comment can only reply to 1 comment but can have many replies, the comment node can have many incoming edges but only one outgoing edge [1]. The graphical representation of Clickbait (Lanyu Shang, 2019) videos is more structured whereas that of non-Clickbait (Lanyu Shang, 2019) videos is unstructured as the discussion in non-Clickbait (Lanyu Shang, 2019) videos is more focused towards the content of the video.

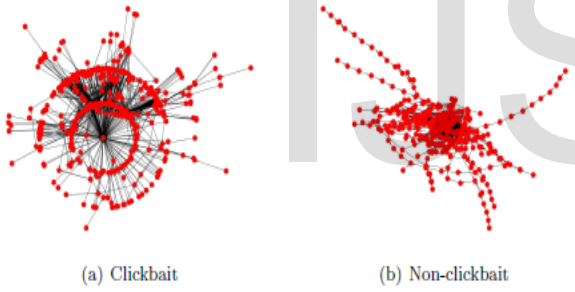

Fig. 1 Graphical structure of comments of Clickbait (Lanyu Shang, 2019)and non-Clickbait (Lanyu Shang, 2019) videos

The semantics features of comments consist of two attributes: Sentiment attribute and Endorsement attribute [1]. A random walk is done across both the attributes. A random walk across sentiments produces a path that represents how polarized a comment thread is, that is how positive or negative it is. Clickbait (Lanyu Shang, 2019) videos have more polarization than non-Clickbait (Lanyu Shang, 2019) videos. A random walk across endorsement produces a path which states how much likes or dislikes a comment got. Clickbait (Lanyu Shang, 2019) videos have more frequency of endorsement on particular selected negative comments whereas non-Clickbait (Lanyu Shang, 2019) videos have a random and diversified likes/dislikes count.

B. *Linguistic Characteristics Analysis*

In this second module we analyses the linguistic characteristics of the comments. The type and frequency of word clouds in Clickbait (Lanyu Shang, 2019) and non-Clickbait (Lanyu Shang, 2019) videos comment threads are different. Clickbait (Lanyu Shang, 2019) videos have more hate words such as 'Clickbait' (Lanyu Shang, 2019), 'fake', 'scam', 'thumbnail' whereas non-Clickbait (Lanyu Shang, 2019) videos have diverse words relevant to the content of the video.

Fig. 2 Word Cloud for Clickbait (Lanyu Shang, 2019) and non-Clickbait (Lanyu Shang, 2019)videos

C. *Metadata Characteristics Analysis*

In this module we extract the metadata of the videos. This data gives us information which cannot be extracted from network and linguistic characteristics.

Features	Description
Comment Count	Total # of comments
Dislike Count	Total # of dislikes
Like Count	Total # of likes
View Count	Total # of views
Like to Dislike	The ratio of like count to dislike count
Daily View Count	Avg. # of daily views
Like to View	The ratio of like count to view count
Duration	Length of video in minutes
Description	Length of video in minutes
Description URL count	Avg. # of URLs in description
Like Count per Comment	Avg. # of likes in each comment
Word Count per Comment	Avg. # of words in each comment
Clickbait Count	Avg. # of words related to clickbait in each comment.

Table 1 Metadata Features

D. *Supervised Classification*

This module integrates results of all the above modules and performs a binary classification on it. We compare different classification techniques such as Logistic Regression, Random Forest and Adaboost classifier.

IV. EVALUATION

We evaluate the accuracy and precision of all the methods against each other and also plot a Receiver Operating

Characteristic (ROC) curve for all to evaluate the robustness of their performance.

V. ACKNOWLEDGEMENT

We are highly indebted to Mr. Vikranth B M for his guidance and constant supervision as well as for providing necessary information regarding the project and also for his support in completing the project. The work reported in this paper is supported by college through the Technical Education Quality Improvement Program (TEQIP – III) of the Ministry of HRD, Government of India.

VI. REFERENCES

[1] Towards Reliable Online Clickbait Video Detection: A Content-Agnostic Approach: Lanyu Shang, Daniel Zhang, Michael Wang, Shuyue Lai, Dong Wang

[2] Peter Adelson1, Sho Arora2, and Jeff Hara3 - Clickbait; Didn't Read: Clickbait Detection using Parallel Neural Networks

[3] Kennedy Ogada, Waweru Mwangi, Wilson Cheruiyot - N-gram Based Text Categorization Method for Improved Data Mining

[4] Hanif Bhuiyan, Jinat Ara, Rajon Bardhan - Retrieving YouTube Video by Sentiment Analysis on User Comment

[5] Muhammad Zubair Asghar1, Shakeel Ahmad2, Afsana Marwat1, Fazal Masud Kundi1 - Sentiment Analysis on YouTube: A Brief Survey

[6] Tyler West - Going Viral: Factors That Lead Videos to Become Internet Phenomena

[7] M. M. U. Rony, N. Hassan, M. Yousuf, Diving deep into clickbaits: Who use them to what extent in which topics with what effects?

[8] M. B • artl, Youtube channels, uploads and views: A statistical analysis of the past 10 years.

[9] D. Wang, T. Abdelzaher, L. Kaplan, Social sensing: building reliable systems on unreliable data, Morgan Kaufmann, 2015 (2015).

[10] A. Agrawal, Clickbait detection using deep learning, in: 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), IEEE, 2016.

[11] E. Alpaydin, Introduction to machine learning, MIT press, 2014