Social Media Comments Network Analysis

Surendhar Muthukumar
Institute of Computer science
University of Goettingen
surendhar.m@stud.uni-goettingen.de

Abstract - This study explores the extent of opinion-based homogeneity in YouTube through a combined approach of Sentimental Analysis (SA) and Social Network Analysis (SNA). The topic of interest is the "Future of Artificial Intelligence (AI) and its impact on future jobs". The study aims to understand how people approach improvements in AI and how their opinions are influenced by opinion clouds. YouTube is used to collect relevant opinions of individuals through video comments. The results of this study show that there is a lack of significant opinion clouds in the global network, indicating that the network is heterogeneous. The study also found that users tend to interact with individuals of opposite perspectives, implying a higher degree of heterogeneity. However, subnetworks with opposite E-I Index values tend to cancel each other out, resulting in a network free from opinion clouds and heterogenic interactions. It is important to note that the accuracy of the results is subject to the topic and user base, and the study is dependent on Natural Language processing.

1 Introduction

The proliferation of social media such as Twitter, Facebook, Instagram, and YouTube has significantly altered the way of interaction and thinking among individuals around the world. After the boom of smartphones and rapid internet developments, exposure to social media has increased among individuals allowing them to simultaneously develop knowledge as well as probabilistically adhere to certain ideologies. Among these platforms, YouTube is a big contributor to the internet usage of people, allowing them to both view and upload videos of their interest. It has a user base of almost 2 billion people with access to a wide variety of videos and the ability to comment their thoughts on the same. The vast amount of data generated by these comments are a rich source of information that can be effectively used to analyze various influential aspects and opinions of people.

In this study, the extent of opinion-based homogeneity of YouTube users is analyzed using a combined approach of Sentimental Analysis (SA) and Social Network Analysis (SNA). Instead of traditional ideological homogeneity, the opinion-based homogeneity study concentrates on individual topics than a common ideology, hence the opinions are highly concentrated on the selected research topic and are not biased by the common ideology. The topic of interest of this study is "Future of Artificial Intelligence (AI) and impact of AI on Future jobs". In the past decade, rapid advancements in the field of Artificial Intelligence have elevated its usability. The influence and demand for AI are exponentially increasing among both the general public and industries/businesses. For example, people are more used to the AI applications like AI assistants, AI home automation systems, etc... Certain industries rely on AI to analyze the user base and provide services based on user preference and interests. This rapid growth has paved way for mixed opinions among individuals as AI has slowly gained an ineluctable position in individual and work environments. Hence analysis of opinions on this topic provides a better understanding of how people approach these improvements and their opinions are influenced by opinion clouds if there are any. YouTube is used as a medium in this study to collect relevant opinions of individuals through video comments.

The combined approach of SA and SNA has been exceptionally viable in past considerations done by Daniel Rochert et al [6], examining the degree of homogeneity among users' political opinions around 3 controversial political subjects through the related YouTube video comments. Sentiment analysis (SA), also known as opinion mining is a branch of natural language processing used to determine the emotional tone of texts such as reviews, and comments and classify them as positive, negative, or neutral. Social Network Analysis (SNA) provides a powerful tool for investigating the patterns and relationships among individuals. The examination of network

structures that emerge from user interactions through comments collected from relevant videos provides valuable insights and results.

To decide the degree of homogeneity, a combined approach of Sentiment Analysis (SA) and Social Network Analysis (SNA) is adjusted. The combination of Opinion Investigation and Arrange Examination permits us to explore the connections between people and the passionate setting of their comments within the setting of the bigger comment arrangement. The results of this study have important implications for our understanding of online communities and their influence on social discourse. As this study concentrates on analyzing a specific topic of interest on YouTube, the results and observations are subject to the topic of interest and may not be the same for all users across YouTube. By examining the sentiment and network structures of YouTube comments, we aim to analyze the extent of opinion-based homogeneity and inspect the presence of any opinion clouds in the network that influence the users and expose them to similar opinions of their interest rather than diverse opinion base.

2 Methodology

2.1 Data Set

The data for this study is collected from YouTube videos that are related to the topic of interest: "Future of AI and AI in future Jobs". To obtain all the comments that are related to the topic, custom-written Python code that integrates YoutubeData API is used. Every video available on YouTube is intrinsically associated with a unique video ID, acting as the reference text in the YouTube database. As the first step, the video IDs of videos corresponding to the topic are obtained using the search.list() and video.list() methods, predefined in API. A list of the most relevant videos is only considered for the study, hence the top 100 most relevant videos are queried. When requesting for the videos through the API methods, the parameter **"Category ID"** is set to **Category 28** to collect videos that are classified as Tech and Science related. The **"relevantLanguage"** parameter of the API request is set to **'en'** to get videos primarily in the English language. The request result is further sorted intrinsically using the parameter **"order"**, where the value is set to **"relevance"**, sorting the requested output in the order of videos that are most relevant to the study topic. Such filters make the data set more technically sound and reliable to the context of the study purpose. The output of any query associated with the API is a dictionary of key-value pairs. The conversion program is written to obtain the data in tabular format with heading rows and value rows (row, column format). The obtained video IDs are stored in local storage as an excel file. With the available video IDs as input for the API methods: comments.list(), commentthreads.list(), every individual comment and the associated replies of relevant videos are collected and stored in local storage as an excel file. The data set was acquired on Jan 19, 2023. In total 155200 comments and replies from 81 unique videos were scrapped for analysis. The data set collected, comprises of fileds: VideoID, Comment/Reply, Author, and Replied_to for the respective comments. The significance of each field is mentioned in table 3. Statistical values of dataset is provided in table 1. The highest number of comments received by a video was 43405, while the lowest number being 12. In total 102526 unique users have contributed to the entire comments set, providing a statistical average of 1.5 comments per user. The corpus is comprised of 4519869 words in total, where on average a comment is composed of 29 words.

Category	Value
Total videos	81
Max comments	43405
Total comments	155200
Unique Users	102526
Average comments	1.5
Total words	4519869
Average words	29

Table 1: **Comments Dataset Statistics**

2.2 Pre-processing Data Set

Figure 1 provides a detailed schematic pipeline of the sentimental analysis process. The comment/reply are texts that comprise lexical words along with noises like emojis, URLs, special

symbols, stopwords, and usernames (in case of specific replies). Hence as the first step to proceed with sentiment analysis, the data pre-processing was performed as mentioned in the pipeline in Figure 1[5][4]. An extensive number of steps are applied to standardize the dataset and reduce its size. Some general data massaging on comments are done as follows:

- Convert the comments into lowercase alphabets
- Strip spaces from the ends
- Remove videos with number of comments less than 10

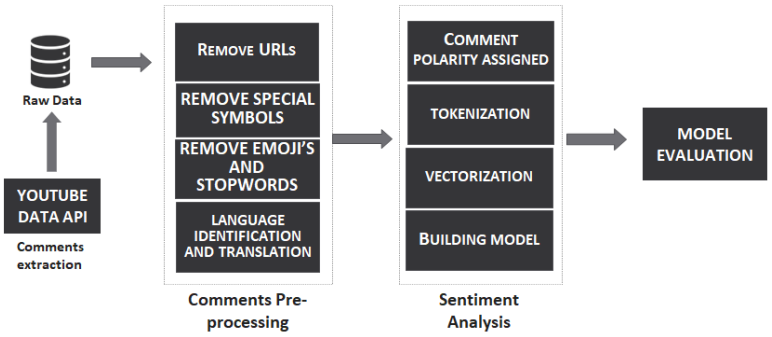


Figure 1: **The sentimental Analysis Pipeline.**

Other noises are handled as follows:

2.2.1 Uniform Resource Locator URL

Comment authors often include links for other web pages as hyperlinks in their comments for reference. These URLs do not contribute to the sentiment analysis and are removed. The URLs are removed with blank space[6][4]. Regex expression to match URL is used: "http\S+"

2.2.2 Username

As YouTube allows addressing replies to specific users, it is common that the comments will have a username attached. The username can be present in any part of the comment. These texts are mere mentions and don't contribute to the analysis, hence are removed using regex. The regex expression to match the username is "`@[^\s]+\\s`". The matched usernames are replaced with blank spaces.

2.2.3 Emoji

Users often use emojis in comments. These provide a fun and quick way to convey a message without the need for lengthy explanations or texts. The emojis may be pictorial or through special characters like punctuation symbols. These are significant noises in the sentiment analysis and prediction, hence are removed from the individual comments in the corpus. Pictorial emojis are identified using respective Unicode and are replaced with blank spaces. Individual emoji sets have predefined Unicode ranges. Emojis generated through special characters (using '`#`.:;,()/\") are regex matched and replaced with blank spaces.

2.2.4 Stop-words

Stop words are common words in a language used in texts. These words do not contribute a specific meaning at the contextual level but are generally used for connecting more important words or to provide grammatical structure. Python **NLTK** (Natural Language Tool Kit) library has a predefined cluster of stop-words for individual languages, that is used to identify and remove the stop-words making the corpus clean for the machine learning model to consume and map the words to sentiment contribution.

3 Sentiment Analysis

The pre-processed comments must be classified into sentimental classes as positive, negative, or neutral for further proceeding into the sentimental analysis and sentiment prediction [6]. The comments were classified using human annotators[6] and lexicon-based tools[5][4] in previous studies. In this study, the comments are classified using a lexicon based tool. The sentiments are classified in context to their stance towards the topic and not to the video content. The videos are used as a support mechanism to find appropriate comments on the topic.

3.1 Sentiment Classification

The comments are classified into positive, negative, or neutral with the help of a lexicon-based annotating tool as mentioned in the previous section. NLTK - Vader (Valence Aware Dictionary and Sentiment Reasoner) is a rule-based tool available through the python NLTK package. This tool is very effective for social media comments and texts. Vader assigns a polarity score between the range of -1 to 1 for the comment texts. The negative range of the polarity score expresses the comment's degree of affiliation to negative sentiment, while the positive range expresses the degree of affiliation to positive sentiment. For this study, the classification is done based on a custom polarity score threshold as shown in Table 2. The number of comments associated and the unique users in individual sentiment class is stated in Table 2. The final dataset with sentiment classified comments have 6 columns and 155200 rows, where each row represent a comment. The significance of the columns/fields are stated in Table 3

Polarity Score	**Sentiment Classification**	**Classified Comments Count**	**Classified Unique Users**
-1.0 to -0.5	Negative	13700	11498
-0.5 to 0.5	Neutral	113387	69040
0.5 to 1.0	Positive	28113	21988

Table 2: **Threshold definition for Sentiment Classification**

3.2 Tokenization and Vectorization

As mentioned in the sentiment analysis pipeline in Figure 1, the annotated (sentiment-classified) corpus is converted into individual lemmatized tokens in the tokenization process. The tokens are converted into a weighted sparse matrix of order m X n, where m is the number of total comments and n is the total number of unique tokens (words). Vectorizers help in the matrix generation process. Two different vectorizers namely, Count Vectorizer and TF-IDF Vectorizer are used and compared in this study. The output matrix from the vectorizer is fed to the machine learning algorithm. In this study the dimension of the matrix was **155200 X 74531**, where 74531 is the number of unique lemmatized tokens/words of the corpus. Since the corpus is natural language oriented the results differ from the nature of the corpus, hence two vectorizers are used and compared for the best suitable option identification through model prediction accuracy as the model prediction depends on the input. The comparison is expressed in table 4.

Field Name	**Datatype**	**Significance**
VideoID	VARCHAR	Unique video ID assigned by YouTube intrinsically
Comment/Reply	STRING	Comment or reply associated to a video
Commentor	STRING	Author of the comment/Reply
Replied to	STRING	To whom the reply is addressed
Polarity Score	FLOAT	Sentiment score assigned by VADER
Sentiment	STRING	Sentiment classification

Table 3: **DataSet Statistics**

3.3 SVM

SVM (Support Vector Machine) machine learning model is used. The model has been proven to be the best option for studies related to social media analysis in [6][4]. The matrix from the vectorizer and the Sentiment field in the corpus is split into train and test sets[2]. Different proportions of train

and test sets provided further details on the dependency of test set significance to the prediction accuracy, mentioned in Table 4. The model was fed with the train set to extract the features from the corpus and develop a best fit of data, desirably used for further sentiment prediction. The accuracy of the prediction was compared using the test set. Examining table 4, Model prediction with count vectorizer with a train, test sets of 50% each has provided the best prediction results. Thus the sentiment analysis is accomplished, providing annotated user comments for the entire corpus.

Train Set %	**Test Set %**	**SVM Model Accuracy**	
		Count Vectorizer	**TF-IDF Vectorizer**
50	50	0.836	0.820
40	60	0.830	0.817
30	70	0.825	0.810
20	80	0.815	0.802
10	90	0.796	0.785

Table 4: **SVM Model Evaluation**

4 Social Network Analysis

Unlike other social media networks like Facebook, Twitter, and Instagram, where individual people are connected directly through friend requests and follow request mechanisms, the social network structure of YouTube users is not visible[6]. Individual users can interact with others on YouTube through only means of commenting and replying to the videos' comments section. Hence the approach to constructing a network from these interactions varies significantly from other social networks. In this study, the interactions between comments, respective replies of users, and channel owners who have uploaded videos are established through a network diagram. The extent of homogeneity of the network is examined through the computation of Krackhardt E-I ratio[6] to the generated global network.

The SNA is performed in 2 parts in this study. As the first step, the network is created using YouTube data to visualize the interactions between users and related videos. The second part deals with the computation of the E-I Index and the analysis of homogeneity.

4.1 Network Creation

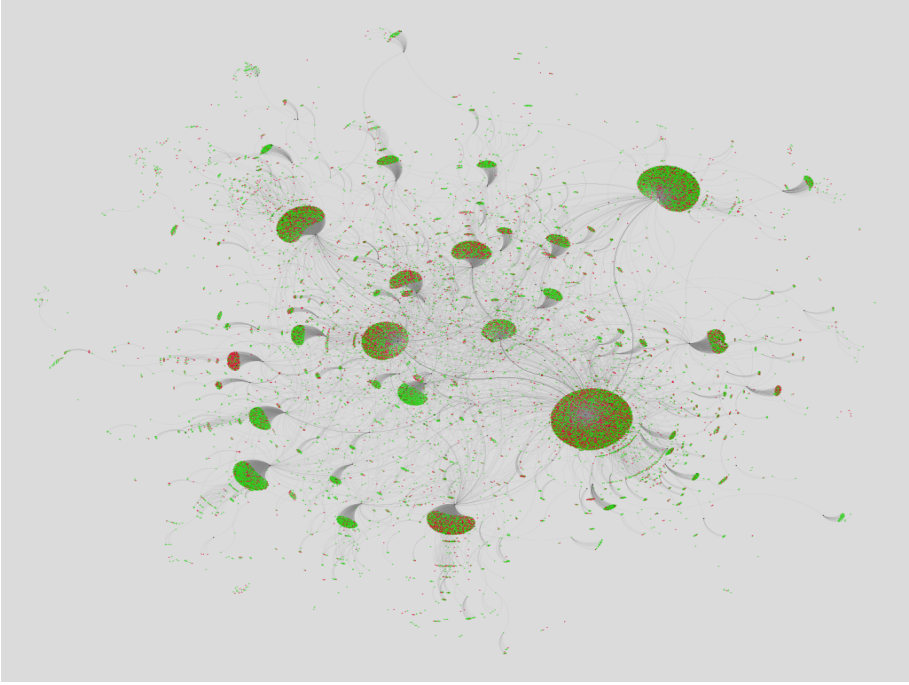


Figure 2: **YouTube Comments Network**

The network is constructed through Gephi. The nodes of the network represent unique users with their associated sentiment class. For users with two or more comments, the dominant sentiment class is considered. The homogeneity intrinsically depends on the presence of opinion clouds, where positive and negative classes are two extreme opinions. Hence only positive and negative sentiment class users are considered for node definition. The video owners are users, whose sentiment class is taken neutral. The edges are typically the interaction between users i.e., how the users connect. Figure 2 provides the generated network. The created network is cleaned by removing isolators. For further understanding of the network deeply and gaining more insights various statistics are calculated and reported in table 5.

Network Parameter	**Value**
Nodes	31789
Edges	35534
Average degree	1.118
Diameter	4
Modularity	0.869

Table 5: **Sentiment distribution of user comments**

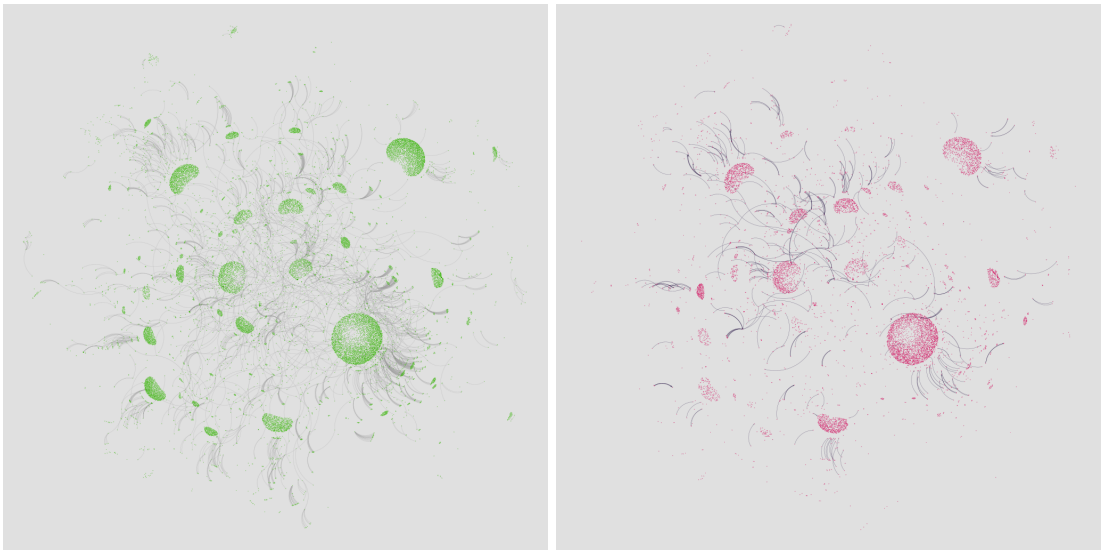
The average degree of the directed network is slightly greater than one, suggesting that an average user has connected to approximately one another user. The dense bundles with higher in-degree in the network represent the users that have uploaded videos. The density of the network is of the order of negative power of 6, explaining the fact that the data originates from a real network.

4.2 Measuring Opinion-Based Homogeneity

The extent of opinion-based homogeneity is measured using the E-I index value. The formula to calculate the E-I index is defined below:

$$\textbf{E-I Index} = \frac{E - I}{E + I} [6][3]$$

E is the number of external links to a given subnetwork (sentiment class) and I is the number of internal links to or between nodes within the same subnetwork (sentiment class)[6]. The index value ranges between -1.0 to +1.0. A resultant value in the negative range(-1.0 to 0) indicates that the network is homogeneous with the degree of homogeneity decreasing with the value approaching zero. Likewise, the resultant value in the positive range(0 to 1.0) indicates that the network is heterogeneous with the degree of heterogeneity decreasing with a value approaching zero. The network is partitioned[1] into two subnetworks of positive and negative sentiment class to determine the E-I Index value of the network.



(a) **Subnetwork of positive users** (b) **Subnetwork of negative users**

Figure 3: **Network Partition**

5 Results

Table 6 provides the E-I Index value of the comments network. To obtain the E-I Index value of the entire network, two subnetworks corresponding to individual sentiment classes (positive and negative) are segmented as shown in Figure 3. The positive and negative subnetworks comprise only users with a positive, and negative stance towards the topic respectively.

Sentiment	**Internal ties (I)**	**External ties (E)**	**Class E-I Index**	**Global E-I Index**
Positive	3266	2419	-0.15	0.345
Negative	468	5217	0.84	

Table 6: **Individual subnetwork and Network E-I Index**

The lack of significant opinion clouds in the global network can be understood from the positive E-I Index value. Hence the network is heterogeneous. The extent of interaction between users of dissimilar perspectives is comparably higher than the interaction between like-minded users in the global network. This implies users' tendency to engage in conversation with individuals of opposite perspectives to either explain their stance or to understand the other. Digging deep into the E-I Index value of subnetworks, the presence of quantifiable differences among the users' interactions are identified.

The positive subnetwork with a negative E-I Index value exhibits a higher degree of homogeneity. Users in the positive subnetwork interact more within themselves than with negative perspective users. This behavior can be understood as the tendency, that people with positive opinions to connect within the positive circle to prove their opinion correctness to themselves. On the other hand individuals of the negative subnetwork, with a positive E-I Index, who are highly heterogeneous connect outside of their network to seek justification for other opinions. Though the subnetworks have opposite E-I Index values, the differences tend to cancel one another, resulting in a network free from opinion clouds.

This study may not be taken as applicable to all YouTube users because of the topic constraint, the results may vary based on the topic and user base. Additionally, the accuracy of the results is not absolute due to the fact that NLTK itself is not 100% percent accurate in understanding/interpreting human context, and hence this study highly dependent on Natural Language processing shares the same effect.

References

[1] https://gephi.org/users/.

[2] https://medium.com/@vasista/sentiment-analysis-using-svm-338d418e3ff1.

[3] https://osf.io/e92n3/?view_only=95ece274e9b74cc29dcadb49a06062fb.

[4] I.Hemalatha, Dr. G.P.S.Varma, and Dr. A.Govardhan. Automated Sentiment Analysis System [u]sing Machine Learning algorithms. *International Journal of Research in Computer and Communication Technology,*, 3(3), 2014.

[5] Shaunak Joshi and Deepali Deshpande. Twitter Sentiment Analysis System. *International Journal of Computer Applications (0975 – 8887)*, 180(47), 2018.

[6] Daniel Röchert, German Neubaum, Björn Ross, Florian Brachten, and Stefan Stieglitz. Opinion-based Homogeniety on YouTube. *Computational Communication Research*, 2(1), 2020.