Identifying COVID-19 Instagram Behaviour Patterns via a Novel Network Analysis Pipeline

Abstract

Instagram has become one of the most widely used social media platforms globally. The volume of information generated by Instagram users makes it an excellent data mine to study human behavior. Recently, researchers have turned to Twitter data to observe the flow of (mis)information surrounding the COVID-19 pandemic. However, there is little-to-no publication on coronavirus-related behaviours on Instagram. To address this knowledge gap, an extensive analysis of COVID-19-related big data from Instagram was conducted. Hashtags obtained from 5,300 global Instagram posts collected this year between January 5 and March 20 were processed using a novel, highly extensible Social Network Analysis pipeline. Eigenvector centrality and weight cluster indexing partitioned the hashtag data into clusters, three of which were studied to reflect public behaviour. It was concluded that contributions to the mental health cluster consisted mainly of Instagram Business Profiles, and that these were associated with positive sentiments. It was also found that a niche cluster containing conspiracy-related content was primarily created by genuine Instagram users, and was associated with very negative sentiments. Lastly, hashtags related to comedy and memes were classified by a supervised machine-learning approach to be bot-generated content. Some suggestions with promising directions for future investigations are provided. It is recommended that a large-scale behavioral analysis of bots and their effects on pandemic information be performed. Progress will be made through public health research focusing on the intersectionality of mental health and COVID-19, as well as the rise of conspiracy theorists amidst the infodemic.

Keywords: Social network analysis, machine learning, eigenvector centrality, Instagram, Covid-19

Introduction

It is undeniable that the internet has become a core pillar of the modern information society. Given a global digital population of nearly 4.66 billion in October 2020, 4.14 billion are active users of social media [1]. Since the declaration of the COVID-19 pandemic [2], public health challenges have surfaced alongside economic and social repercussions [3, 4]. To understand this multifaceted issue, Instagram—a social media platform with over one billion active users per month—can be used to analyze business and individual behaviors during the pandemic [5]. According to a systematic review published in April 2020, researchers are currently using text mining to monitor the growth, nature, and spread of COVID-19 [6]. This allows for population and public health insights to be drawn from digital conversations, a source of information that may differ from media reports and epidemiological findings. In the face of public health challenges, as many inputs as possible should be considered before arriving at a policy decision. However, only Twitter data has been used for this purpose to-date [6]. This paper characterizes COVID-19-related open big data from Instagram to identify patterns and create recommendations for public health authorities to manage the pandemic. Since only Twitter data has been explored for this purpose, our study fills the knowledge gap for published COVID-19 analyses using Instagram datasets. This paper aims to leverage network analysis to categorize public behaviours into themes, use machine learning to identify different user groups associated within the Instagram dataset, and utilize sentiment classification to gauge how positive or negative the conversations are around COVID-19.

Of the many features within an Instagram post, hashtags are the simplest starting point to identify social media interactions. Each hashtag represents a topic of discussion, and can be analysed using account type, post-associated captions, and post-associated comments. Using Social Network Analysis (SNA), COVID-19-related Instagram hashtags are clustered to characterize public discussions. We partitioned hashtags into clusters and identified user types such as Instagram Business Accounts. We employed ensemble methods in machine learning to detect potential bot behaviour and utilized sentiment classification to gain awareness of various populations' moods around the pandemic. In all, we created a novel and extensible social analysis pipeline that can be used to guide personal and public health decisions.

Methods

Instagram Data Set

COVID-19 related Instagram data was obtained from the Hyper Articles en Ligne, through the Instagram Application Program Interface's hashtag engine [7]. Approximately 5300 public Instagram posts from January 5, 2020 until March 20, 2020 containing hashtags such as #coronavirus, #covid19, #covid_19, #corona, and #stayhome were collected. Approximately 5,300 global Instagram posts and their associated comments were included from January 5, 2020 until March 20, 2020 [7]. COVID-19-related posts were purposely targeted to gain insight into public behaviour, generating data that reflected sentiments provoked by the ongoing pandemic.

Thematic Clustering

Utilizing the Instaloader library in Python 3.7, information from each Instagram post was extracted. A function, hashtags_count, was created to rank the hashtags based on their frequency of occurrence. Classification of hashtags into thematic groups was performed using an adjacency matrix, M (Appendix 1). To understand the relationship between hashtag nodes, eigenvector centrality was implemented to define node importance based on its relationship to other nodes.

This relationship, known as the node connectivity, ranks nodes higher based on how many connections a given node has to other high importance nodes. In practice, this would represent hashtag utilization frequency and how often they are used with other high-traffic hashtags. Using the Perron-Frobenius theorem, positive eigenvalues were guaranteed for the largest eigenvalue. From this method, the hashtags that were most connected with other well-connected nodes were deemed the most important. In this analysis, only the first eigenvector of the sociomatrix was utilized, as this eigenvector captured most of the variance found in a network [8].

The cluster coefficient of nodes was evaluated as another primary metric to establish thematic groupings of the hashtag content. The cluster coefficient provides values that indicate the strength of association to its neighbours [9]. This translates into the ability for easy identification of specific, key hashtags that are less connected than others.

The Girvan-Newman algorithm was also used to distinguish hashtag clusters. The edge removal process within the algorithm is based on calculating the betweenness of edges and removing the edges with the greatest betweenness value. Betweenness is calculated by counting the number of shortest paths from one node to all other notes present. The higher the value, the more distant the nodes are from each other. This repetitive process provides a hierarchical decomposition of the complex network into distinct clusters.

To account for over- or under-fitting of the data, the weighted density was incorporated such that the density contribution for a cluster was proportional to the number of nodes present. Through this metric, underfitting would be represented by a low weighted density that increased quickly as the number of clusters increased. Overfitting would be shown by high densities that stayed constant regardless of the number of communities. The weighted density of the network was then calculated and plotted (Figure 1) to determine that 11 clusters would be the most representative of the data. As this was a global dataset, several clusters were in non-English languages including Persian and Portuguese. Therefore, only the three clusters, in English, were selected for further analysis.

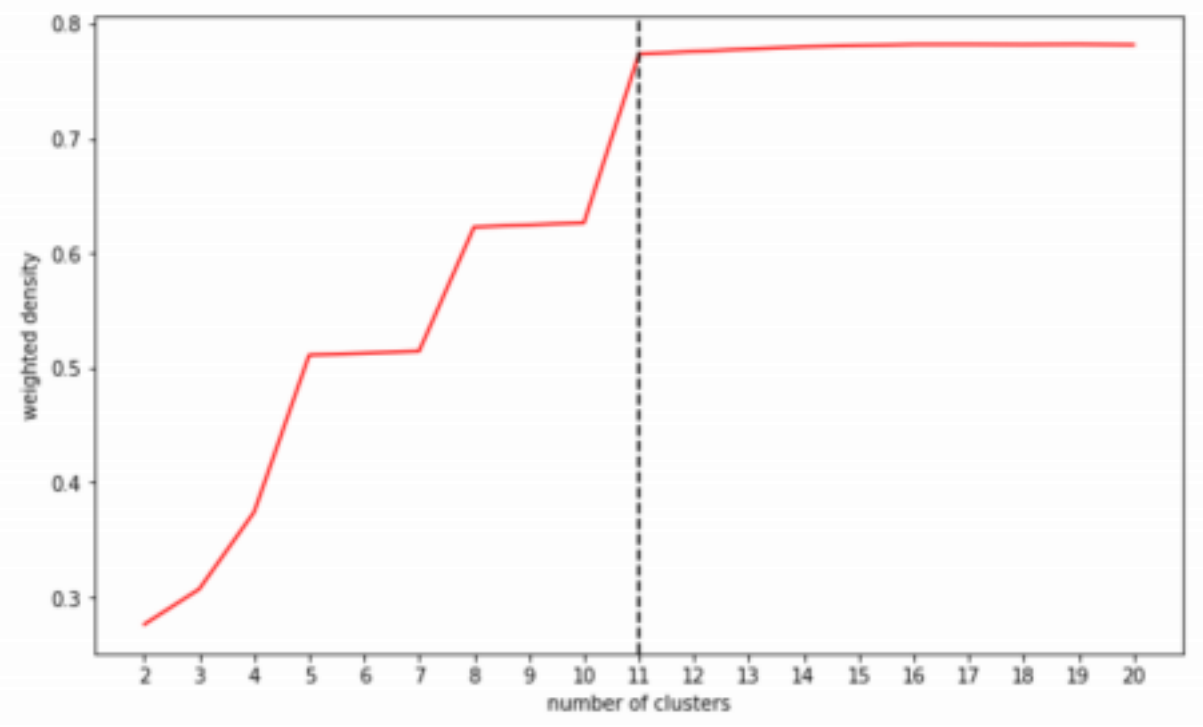

Figure 1. To prevent underfitting and overfitting of data, the weighted density was calculated for each node and averaged in a given cluster. From the analysis, it was determined that 11 clusters would be the most representative of the specific dataset.

Bot Detection & Account Type

To identify bot accounts, a program was created to analyze associated account data from the posts. A supervised learning approach was utilized as a pre-labelled test and training data set distinguishing human and bot accounts based on several characteristics, including profile picture presence, ratio of numerical characters in the username and number of followers, was used for the development of the program. Given the small size of the training dataset, different ensemble learning methods were evaluated as they tend to overfit less on small datasets than other methods, according to the "wisdom of crowds" principle [10].

The methods tested were: Decision Tree Classifier with Adaboost, Extra Trees Classifier, Random Forest Classifier, and Gradient Boosting Classifier. Each model was fitted with many hyperparameters, with 10-fold cross validation acting as the performance metric to isolate the best hyperparameters. A gridsearch determined the most representative model. Using this information for the training and cross-validation sets, learning curves (Figure 2) were created to visually identify the most effective model. The Extra Trees Classifier was found to be the best performing machine learning model for the bot-detection program and also had the highest test accuracy at 92%. This was used to identify bot accounts.

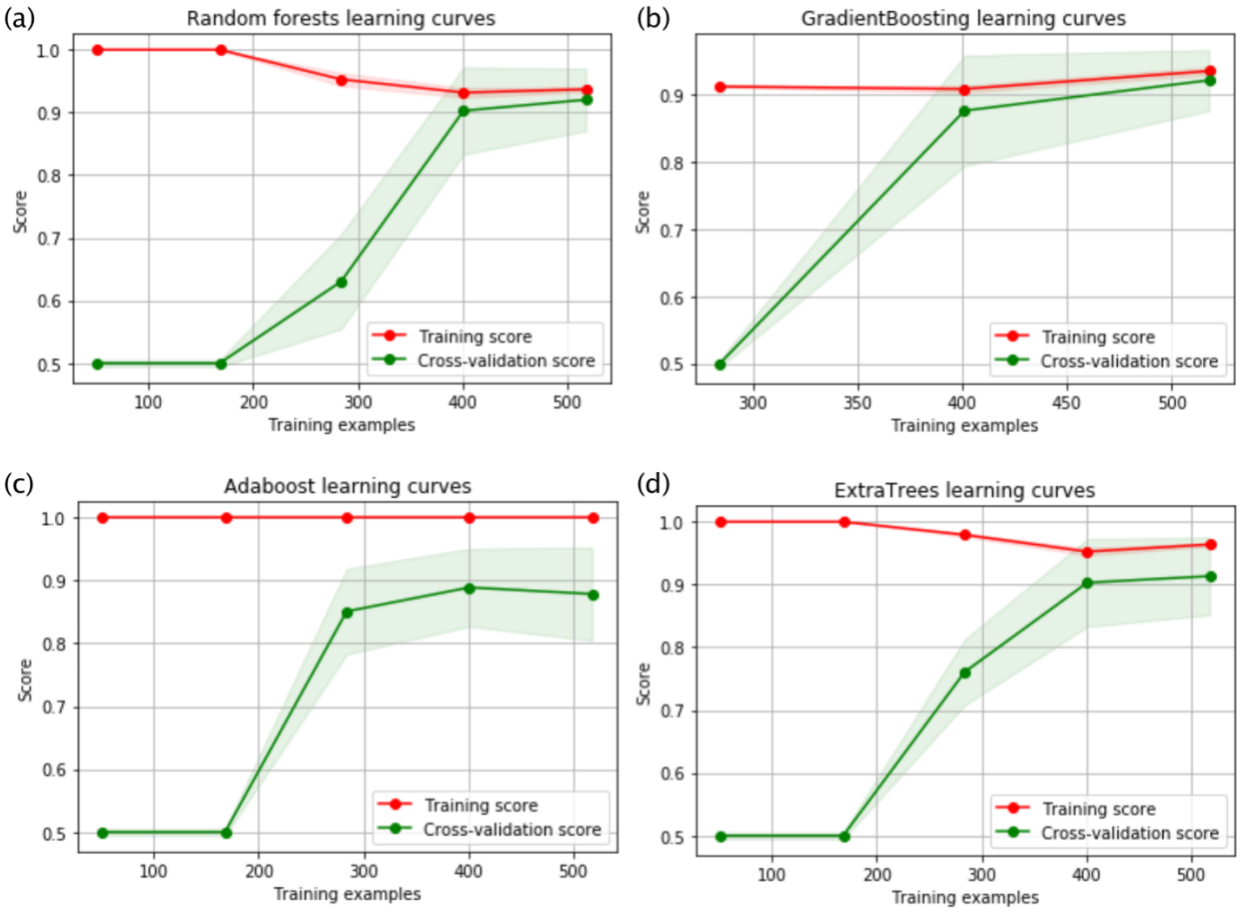

Figure 2. The learning curves for the four machine learning models tested: (a) Random Forest Classifier, (b) Gradient Boosting Classifier, (c) Decision Tree Classifier with Adaboost, and (d) Extra Trees Classifier. An indication of bias was present in the Random Forest and Gradient Boosting classifiers as the maximization of the training-set data had not yet occurred. The Decision Tree Classifier presents 100% training set accuracy with only a 10% differential in cross-validation accuracy—indicative of overfitting. Overall, the use of the Extra Trees Classifier was deemed most representative of the data present.

The data allowed for the detection of specific user identities (Instagram Business Profile, bot account or regular user) associated with a given hashtag. Business Profile information was explicitly given in the data. For each individual hashtag, the percentage of Business Profiles and bot accounts was determined with any remaining accounts being classified as personal accounts of regular users.

For further reporting of the model's performance, confusion, precision, and recall matrices were plotted (Figure 3).

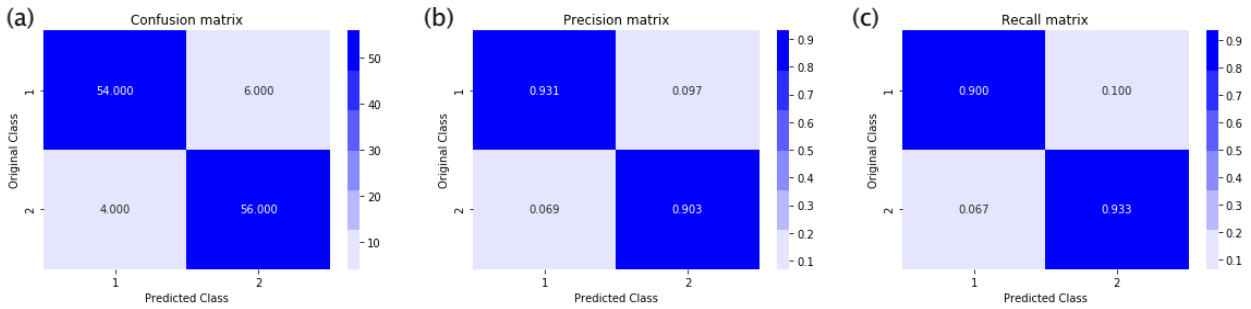

Figure 3. (a) Confusion matrix describing the performance of the Extra Trees Classifier on the bot-detection dataset, showing good model performance concerning identification of True Positives and Negatives. (b) Precision matrix showing the fraction of relevant instances among retrieved instances. (c) Recall matrix showing the fraction of the total amount of relevant instances actually retrieved.

Sentiment Classification

In order to monitor the general sentiment or mood of populations engaging in certain conversations, a Bidirectional Encoder Representations from Transformers (BERT) model was fine-tuned. For the analysis, pretrained weights of the BERT base model were utilized. This model, with 12 transformer layers, has 12 self-attention heads and uses the Hugging Face library [11]. A binary classification layer was then added to the model and adjusted for sentiment classification tasks to apply transfer learning. The Sentiment-140 dataset from Twitter was chosen to train the model. This dataset contains 1.6 million examples of Tweets, rated as either positive or negative, that train the BERT model to accurately detect the sentiments behind a statement. The BERT model was adjusted for 2 epochs using the Adam optimization algorithm with weight decay and resulted in a final test accuracy of 87%.

Results

Frequency of Hashtag Appearance

Altogether, 21,913 hashtags were isolated from the dataset. At first glance, the most utilized hashtags within this criterion included “art” and “love”. While the frequency of hashtag occurrence provided an analytical starting point in determining public behaviour, the grouping of hashtags into broad, thematic categories allowed for insights to be drawn on a macroscale. The adjacency matrix allowed for observation of how strongly-related certain hashtags were to each other.

Table 1. Most Used Hashtags

Rank	Hashtag	No. of Uses
1	#love	334
2	#art	265
3	#fashion	142
4	#explore	137
5	#fitness	134
6	#beautiful	133
7	#nature	129
8	#life	126
9	#memes	124
10	#tiktok	124

Eigenvector Centrality, Cluster Coefficient & the Girvan Newman Algorithm

The top 200 hashtags were to generate the 11 key clusters, where the 3 multi-node, english clusters were chosen for further analysis (Appendix 2). Each hashtag in every cluster was analyzed given a title representing the overarching theme (Table 2). Cluster A represented mainstream trends and encompassed a variety of genres typical to the average Instagram user. Cluster B presented terms related to “mental health” and Cluster C was representative of terms related to conspiracies. The Girvan-Newman algorithm shows this identification of key communities (Figure 4).

Table 2: Chosen clusters from the co-hashtag network barring low weighting & foreign languages

Category	Cluster A	Cluster B	Cluster C
Number of Nodes	112	25	23
Number of Edges	4163	300	253
Average Clustering Coefficient	0.7377	0.8407	0.9481

Average Eigenvector Centrality	0.0892	0.0231	0.0070

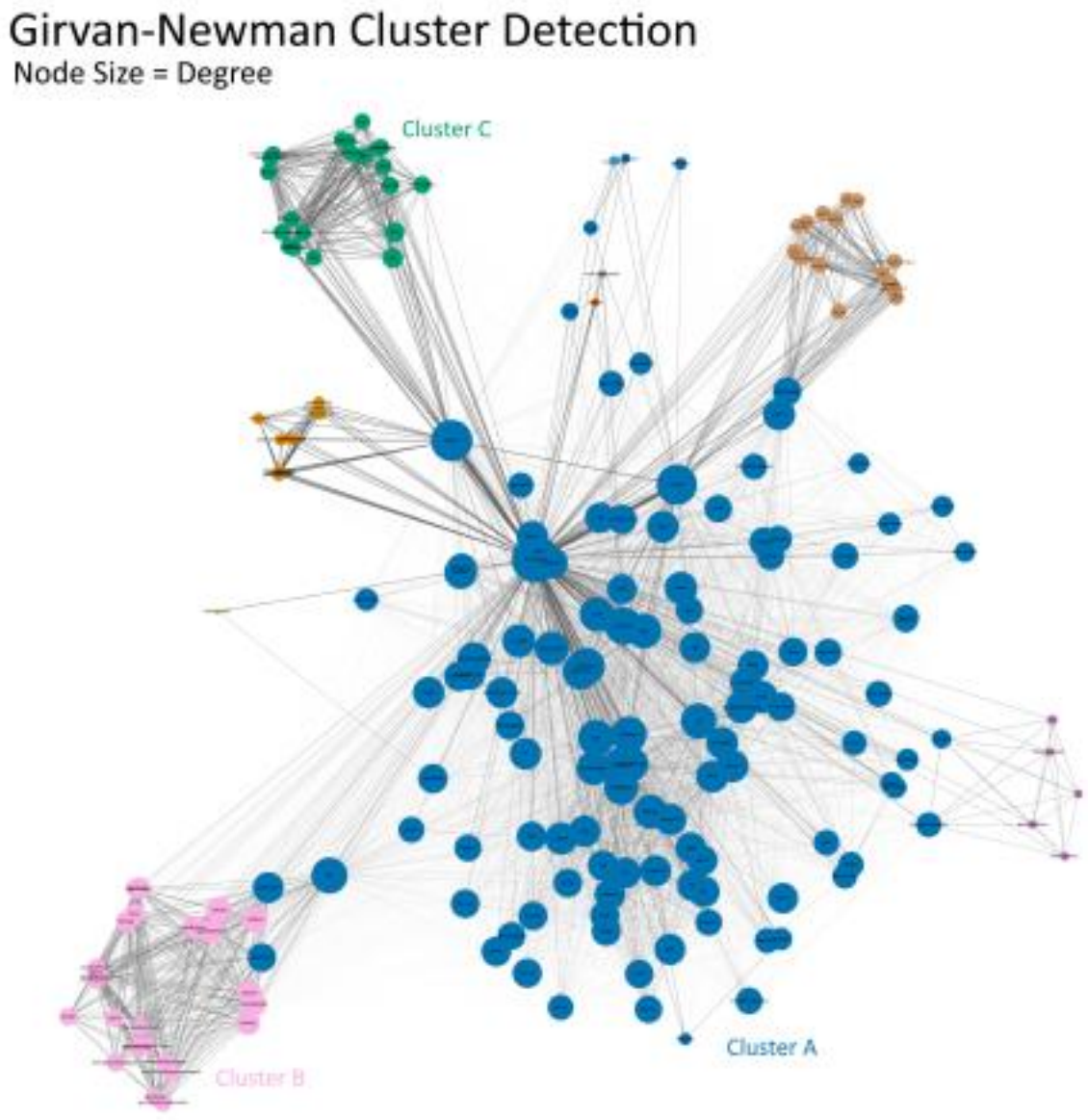

Figure 4. The Girvan-Newman algorithm identified 11 communities within the utilized Instagram dataset. Using the "spring layout" of the plotting function, it was observed to be an excellent visual partition that further confirmed findings.

Through Social Network Analysis, interactions between hashtags were mapped onto undirected graphical representations. Eigenvector centrality was chosen over other centrality analyses since it reflected how widely used and versatile any given hashtag was (Figure 5). A

graph was created to show nodes with greater clustering coefficient values to allow for the quantification of how often specific hashtags are used outside of mainstream trends (Figure 6).

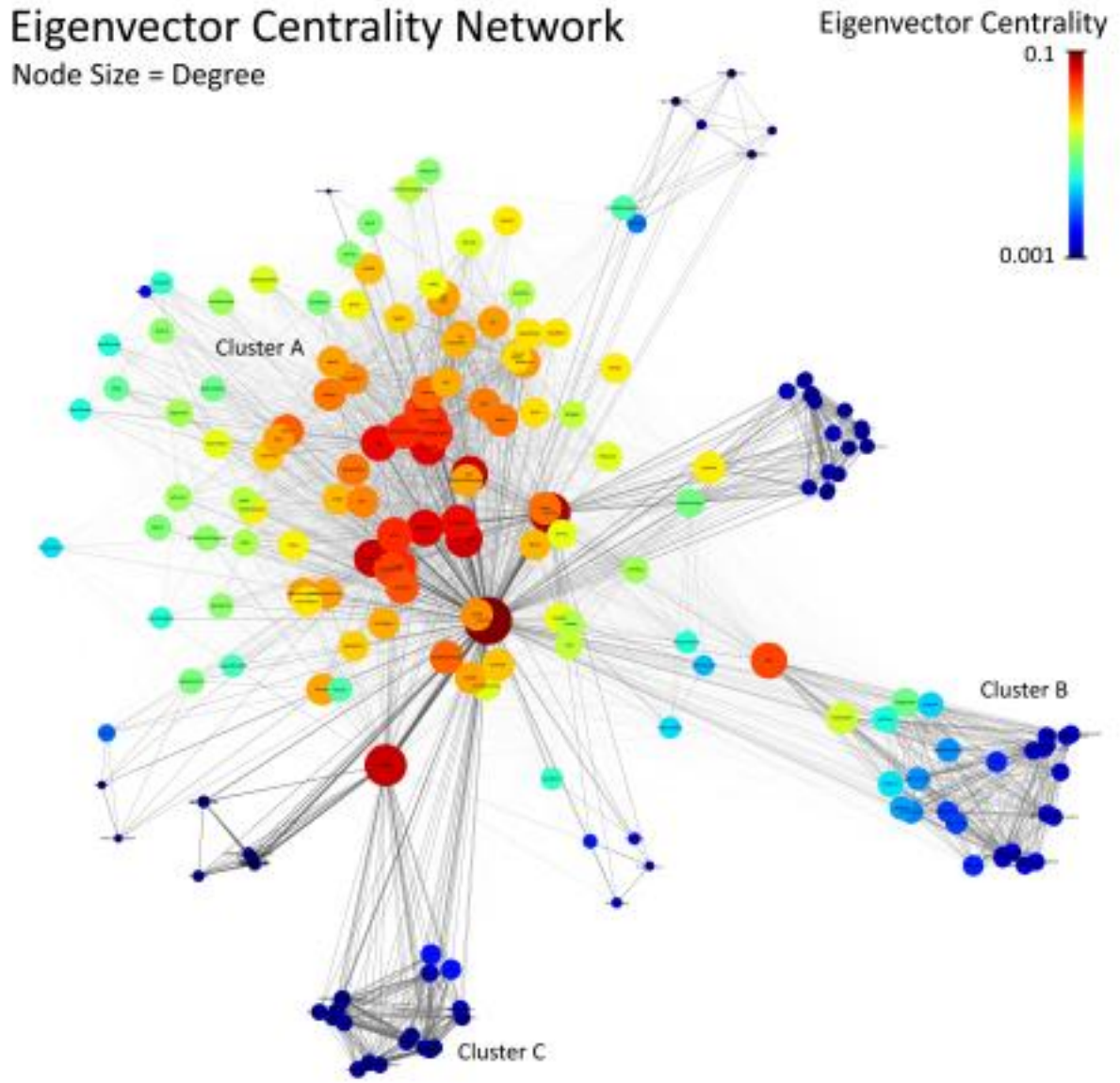

Figure 5. The co-hashtag network using the eigenvector centrality matrix measures hashtag versatility. Hashtags with higher eigenvector centrality values (red) are well-connected to a broad variety of topics. Hashtags with low eigenvector centrality values (blue) showcase specific, niche conversations in the network.

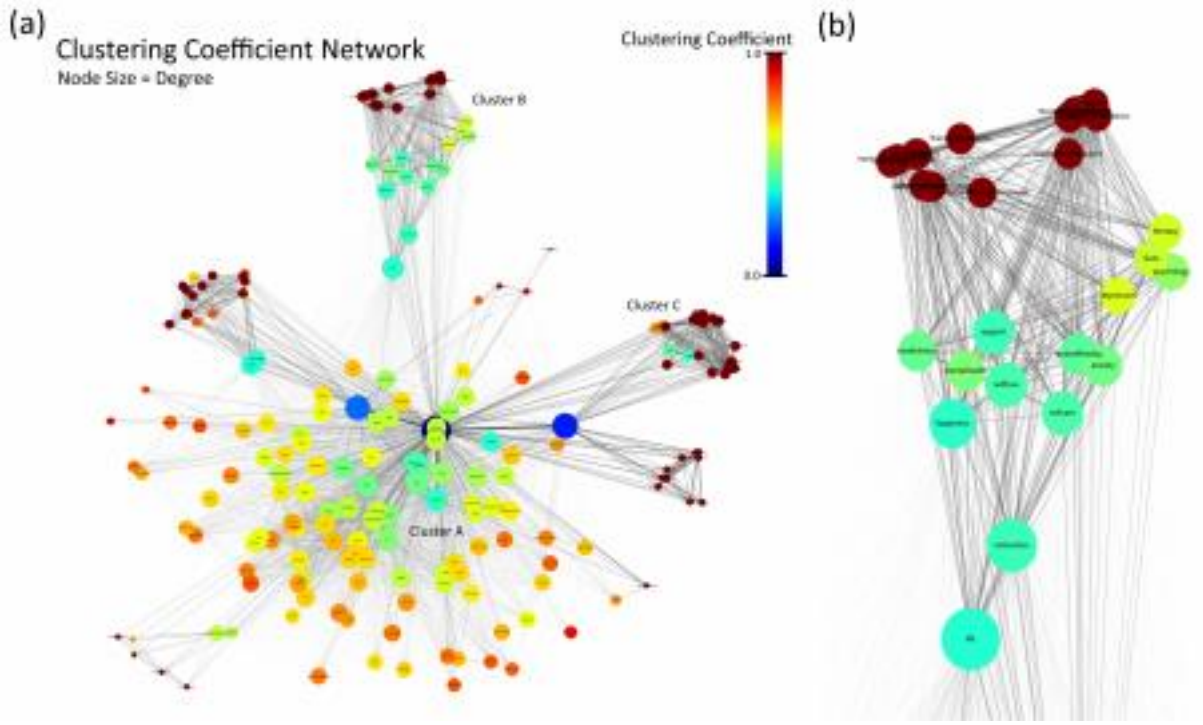

Figure 6. (a) Clustering index network categorizing nodes based on the clustering coefficient and characterizing node connectivity within the cluster. (b) A focused view of the clustering index network, showcasing the variations within the clustering coefficient values of Cluster B.

Hashtag Distribution of Business and Bot Accounts

The bot-detector program focused on how bot-defining characteristics interact with one another. Figure 7 displays a heat-map and pairwise plot of the obtained data distinguishing bot and human activities to better understand the correlations that lie between bot definers.

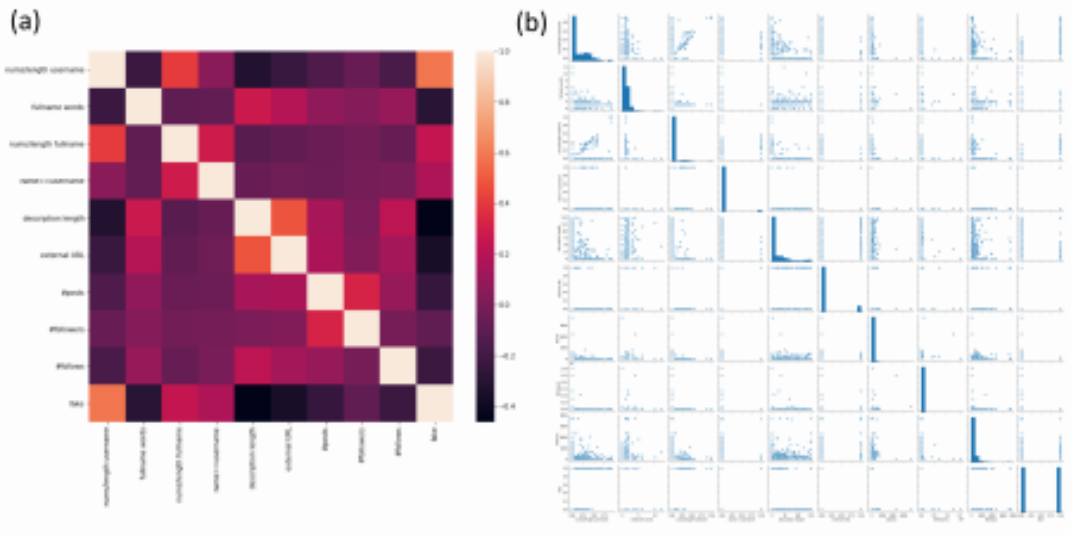

Figure 7. (a) Heat map plotting various bot-defining characteristics, such as profile picture presence, username composition, etc. from the dataset with the lighter boxes showing characteristics with greater correlation with one another. (b) A pairwise plot of titular dataset to gain intuition on different correlations between the same characteristics as outlined in a), but with a greater emphasis on visualizing trends and the distribution of variables between two features.

Focusing on the three clusters of interest, Table 3 identifies what percentage of hashtags are associated with which type. The "mainstream trends" shows a heavy presence of bot accounts. Bot accounts are completely absent from the mental health cluster, instead Instagram Business Profiles make up the majority of mental-health-related hashtag use. The "conspiracy" cluster shows minimal involvement from both. Data from Table 3 was further visualized in Figures 8 and 9. Figure 8 has a fairly distributed representation of hashtags with large business presence and others without, Figure 9 displays uniformity in that most hashtags have little contribution by bots, with a select few hashtags being highly associated with fake accounts.

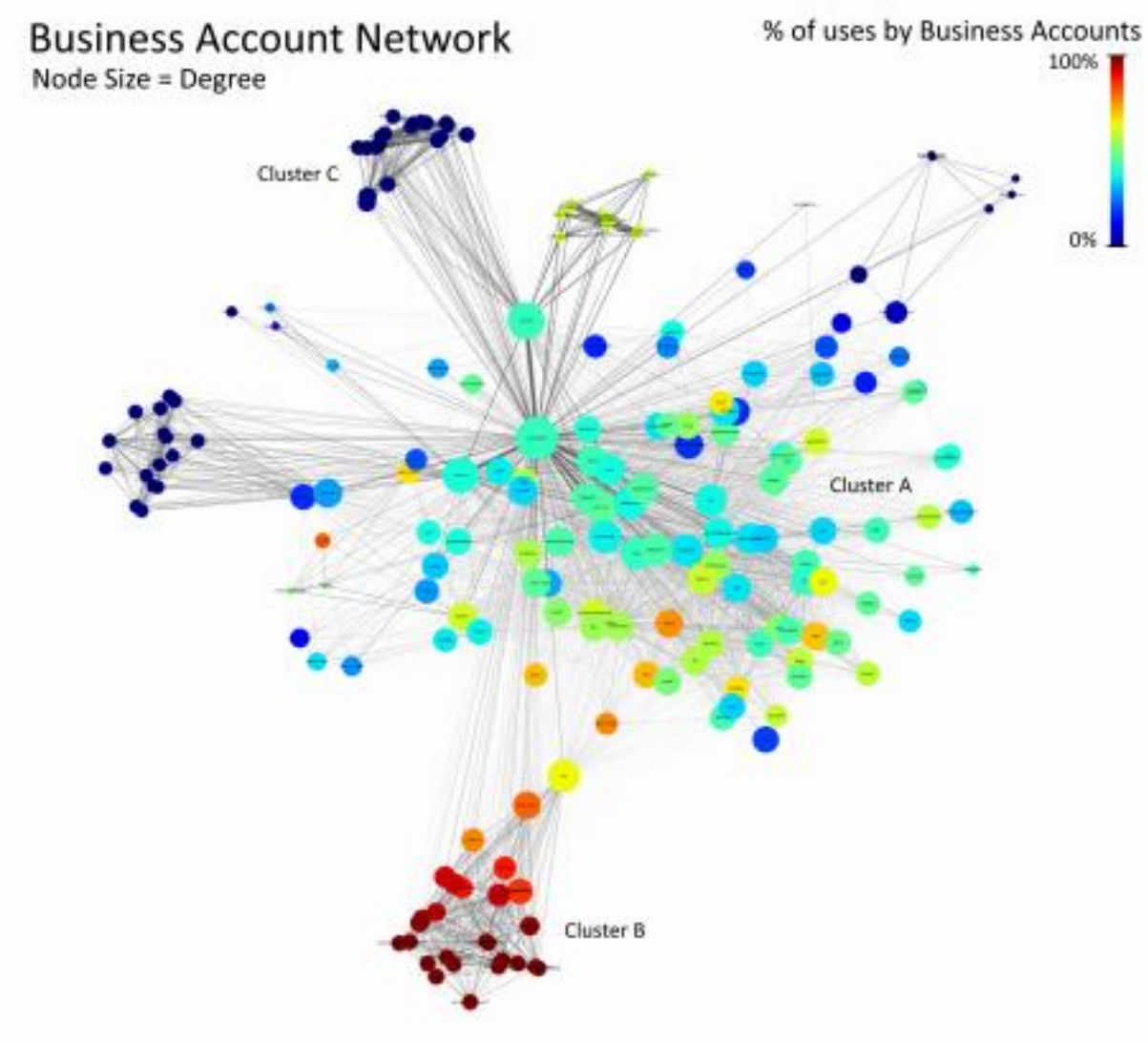

Figure 8. The undirected network displaying frequency of hashtag usage by Instagram Business

Profiles. Darker reds indicate stronger use of the given hashtag by Business Profiles, while blues indicate infrequent utilization. This plot has a wide spectrum of colours, with more uniform clusters.

Table 3: Business- and Bot-Account Presence

	Business Profiles (%)	% of Bot Accounts (%)
Cluster A (Mainstream)	42.2	1.7
Cluster B (Mental Health)	96.2	0
Cluster C (Conspiracy)	0.1	0.1

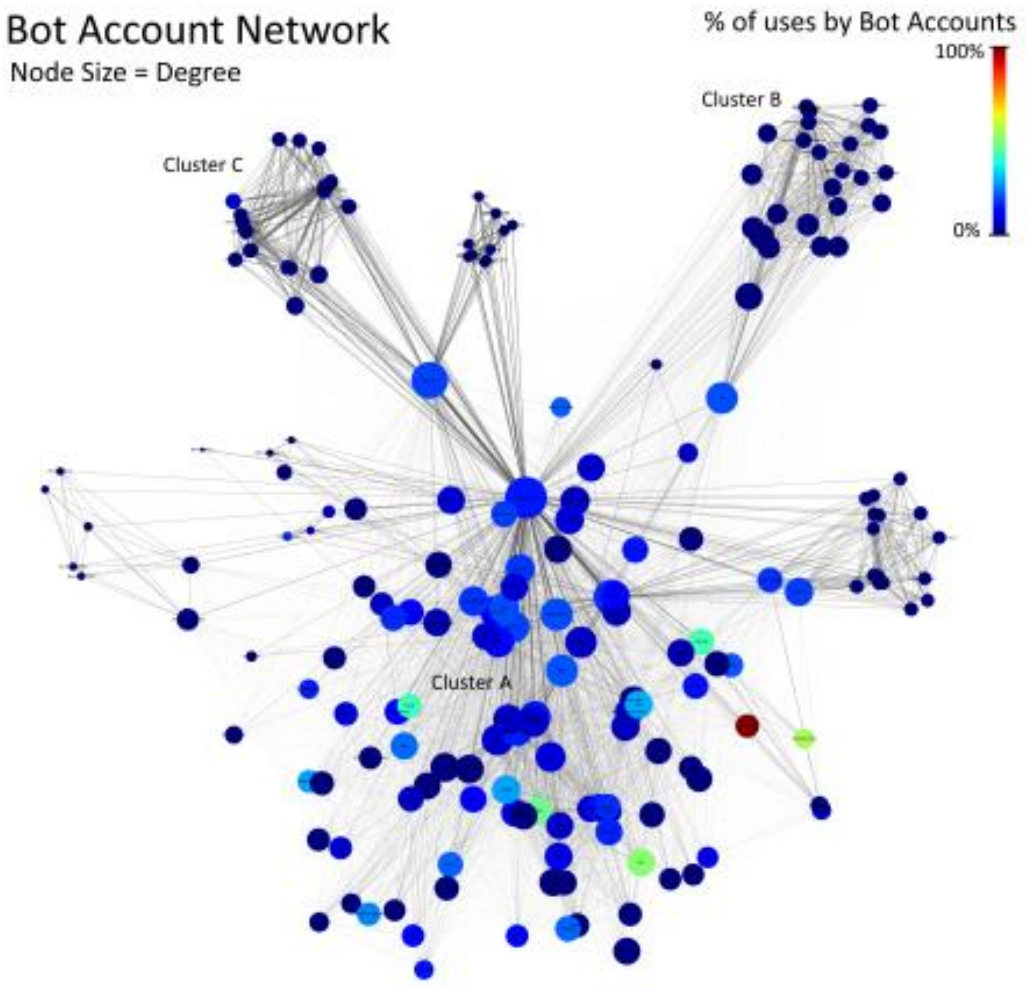

Figure 9. The undirected network displaying frequency of hashtag usage by bot accounts. Darker blues represent little association with fake accounts. Most posts have very little bot account presence, but certain hashtags (e.g., #funnymemes, #masks) have a large bot-account presence.

Sentiment Classification with the BERT Neural Network

The sentiment classification process only considered captions and comments written in the English language due to the limitations of the BERT platform. Figure 10 shows the co-hashtag network for posts. Posts containing mental health related hashtags have generally more positive sentiment in their captions. In contrast, conspiracy-related hashtags show very negative post captions. The "mainstream" hashtag generated both positive and negative sentiments.

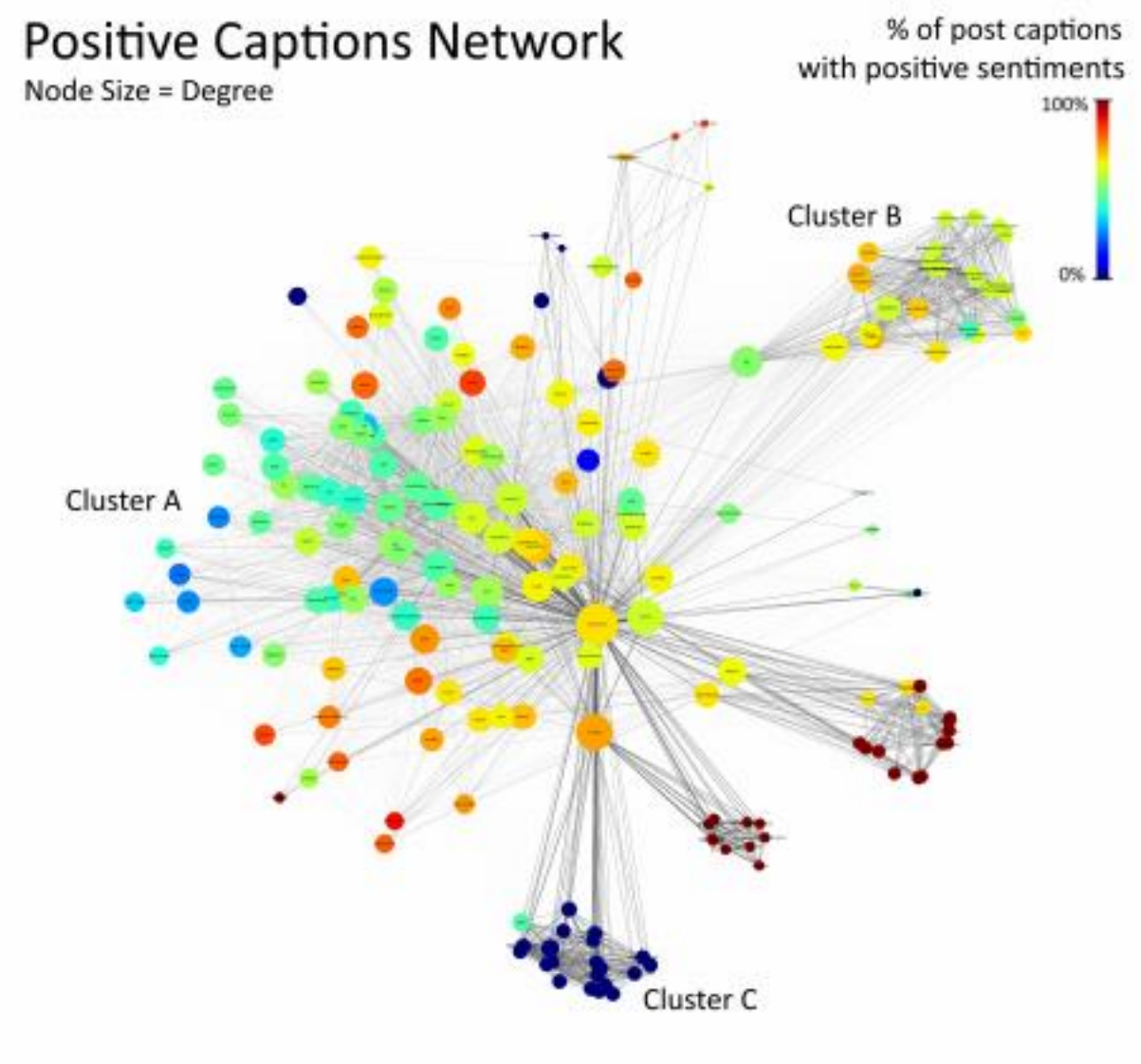

Figure 10: The undirected network displaying percentage of positive captions. Darker red nodes indicate that the captions associated with the given hashtag are generally more positive, while blues indicate general negative sentiment.

Figure 11 shows the co-hashtag network for comments. Posts containing mental-health related hashtags have more positive sentiment in their comments, in addition to their captions. Again, conspiracy-related hashtags show generally negative post comments, as seen by the blue community at the upper-right of Figure 11. “Mainstream” hashtags are also, again, used in posts with both positive and negative sentiments–this time in their comments–as was the observation for these posts’ captions.

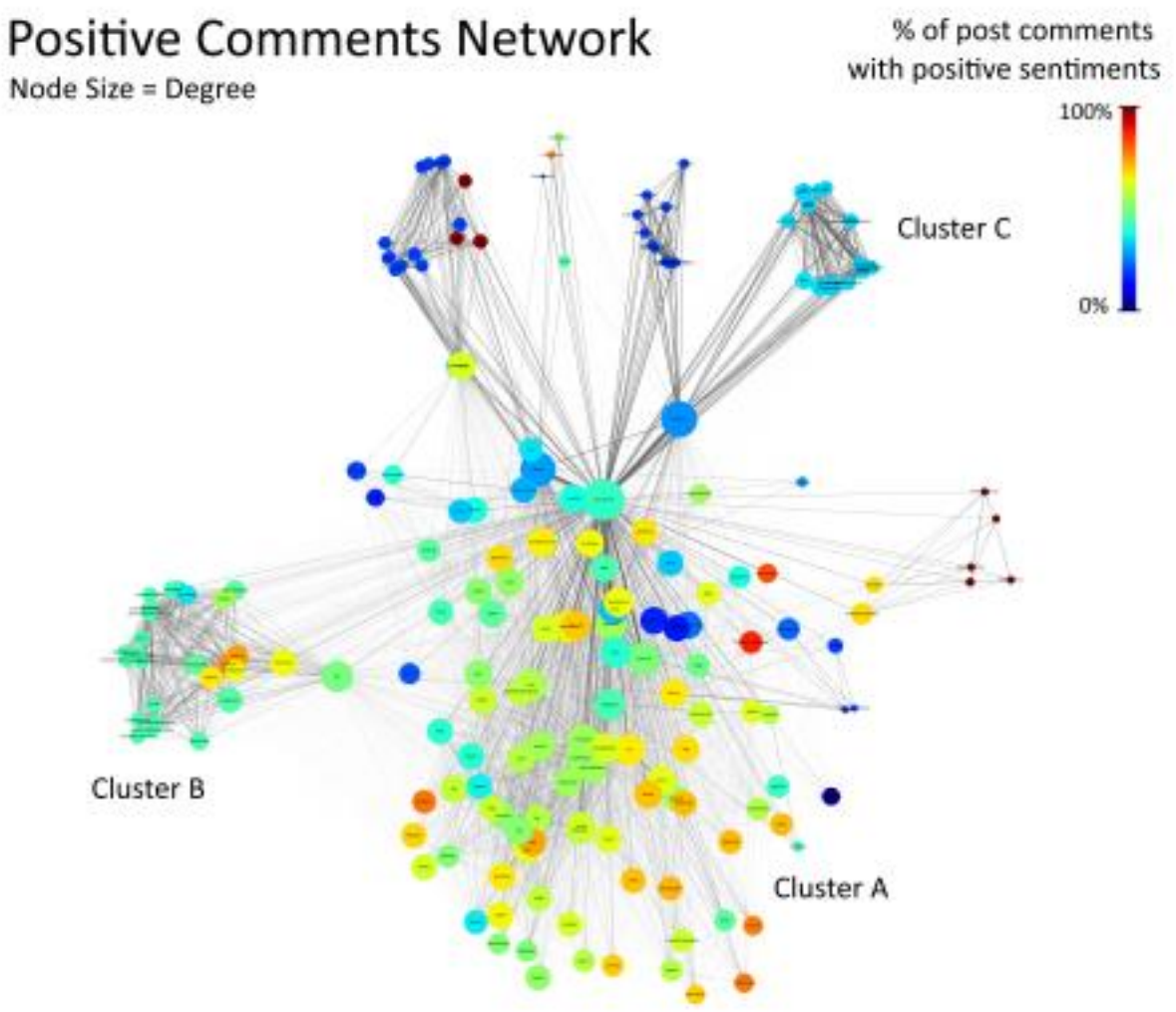

Figure 11: The undirected network displaying percentage of positive comments. Darker red nodes indicate that the comments associated with the given hashtag are generally more positive, while blues indicate general negative sentiment.

Table 4 summarizes the results by showing what percentage of comments and captions. For each cluster, the percentage of negative captions was more prevalent than the percentage of negative comments, by a factor of three. Of the three clusters, Cluster C showed the greatest proportion of negative comments and captions.

Table 4: Caption and comments sentiments

	Negative comments (%)	Negative captions (%)
Cluster A (Mainstream)	14.7	44.1
Cluster B (Mental Health)	13.7	38.0

Cluster C (Conspiracy)	29.8	96.2

Lastly, a summary dataframe was created to compare the correlation between all the metrics analysed for the co-hashtag network (Figure 12). A weak negative correlation (r = –0.2) exists between the percentage of negative comments and the percentage of fake users, implying that comments of negative sentiment are not always made by fake users. There was also a weak negative (r = –0.18) correlation between the percentage of negative captions and the percentage of Business Profiles, meaning that Business Profiles may slightly tend to produce captions with a positively-classified sentiment. Interestingly, there was a moderate positive correlation (r = +0.5) correlation between fake users and the eigenvector centrality of hashtags. This suggests that some number of accounts suspectedly held by fake users generate content with influence on Instagram, as eigenvector centrality is a measure of a node's influence within a network.

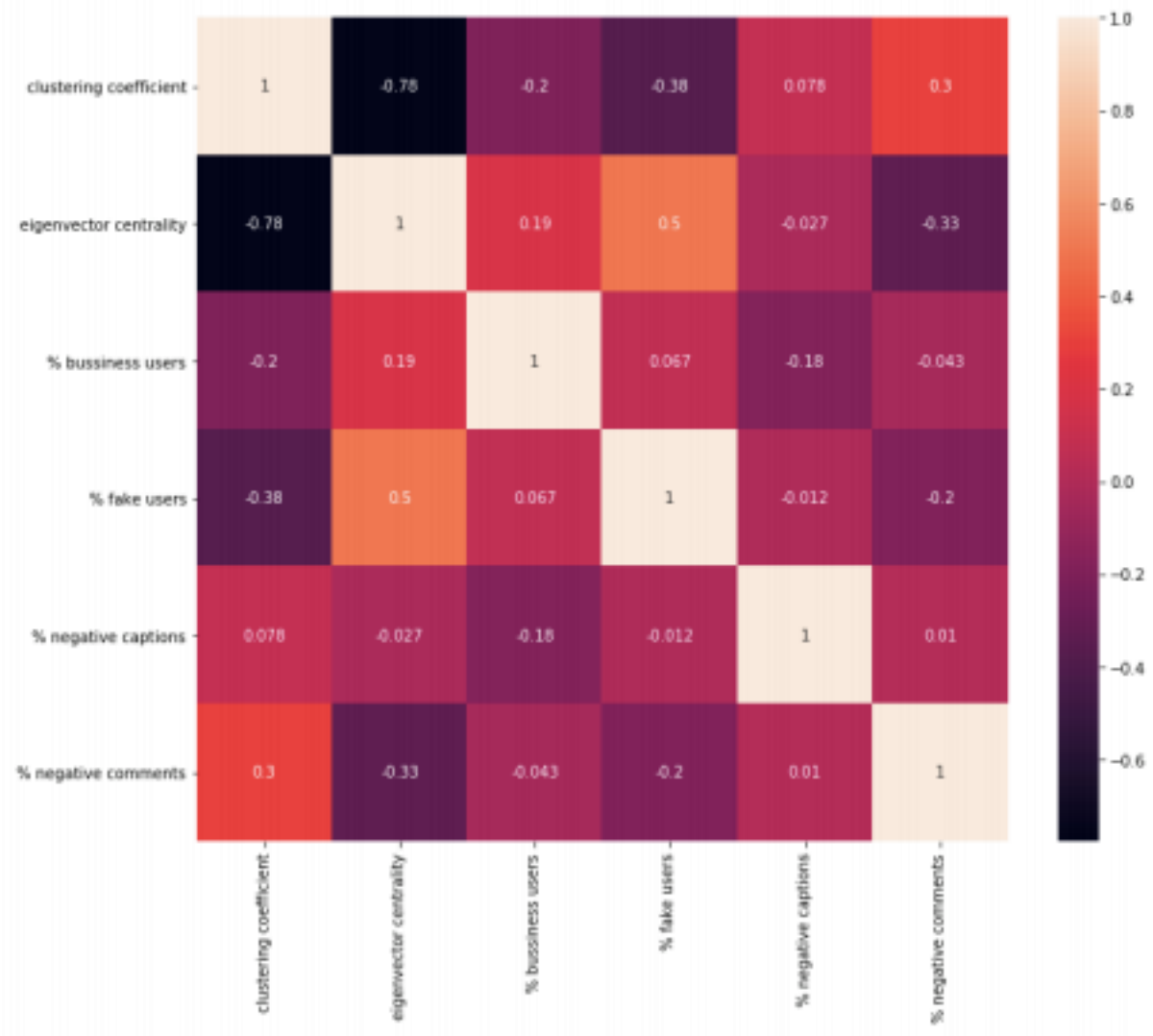

Figure 12: Correlation matrix with heat-map displaying the correlation coefficient between each

metric used and another given metric used, on the co-hashtag network.

Discussion

This study explored the digital realities of the coronavirus pandemic by leveraging big data to investigate social media behaviour. Social network analysis was used to characterize COVID-19-related Instagram activity. Hashtags, captions, and comments were chosen as features to study within the Instagram ecosystem. A co-hashtag network was created, from which edges and nodes were analyzed to indicate the degree of commonality and nicheness of Instagram-user activities. Supervised machine learning and sentiment analysis were used to analyse the possibility of bot-generated content, as well as to map general user content to positive and negative connotations.

The Girvan-Newman algorithm and clustering index analysis grouped nodes into 11 clusters with various degrees of interconnectivity. A visual representation via an undirected co-hashtag network confirmed the precision of the partition (Figure 5). Three clusters were chosen for further analysis: "Mainstream Trends", "Mental Health", and "Conspiracy Theories". The chosen attributes from these nodes include the percentage of Instagram Business Profiles and the percentage of "fake" users (i.e., bots) in the network.

To quantitatively analyse these clusters, their respective clustering coefficients and eigenvalue centrality metrics were compared. A higher eigenvalue centrality indicates the presence of more mainstream topics. A high clustering coefficient alludes to the community being popular amongst niche users only (i.e., echo-chamber). A closer look at the three clusters of interest shows that:

I. Hashtags grouped in Cluster A originated from genuine users, rather than bots (Table 3). Additionally, 42.4% of accounts using these hashtags were Instagram Business Profiles. Nodes within this cluster include #art, #music, #nature, #fashion, #photography, and #love.

II. The mental health hashtag cluster contained the highest percentage of Instagram business account users, at 96.2% (Table 3). It could be that businesses that do not generate revenue primarily from mental-health services are attempting to strengthen their public image. Or, businesses of any type could be advertising support for their employees and customers through hashtag association with the broader mental health community. Professional mental health clinics who use Instagram for business are likely still promoting mental-health-related content.

III. Cluster C contained keywords such as "aliens", possibly due to the United States Department of Defense's recent confirmation of videos showing an unidentified high-speed object [12]. Despite the potentially unreliable nature of data concerning "conspiracy theory" content, there were almost no users labelled by the supervised learning algorithm as a bot or as an Instagram Business Profile (Table 3). There has been online discourse about the pandemic being intentionally manufactured, or as a "cover" during the release of "UFO videos" [13]. Alternatively, users may be including pandemic and conspiracy hashtags together simply because the two events are occurring within the same news cycle. The former is more dangerous, as it lends to the potential spread of misinformation in this global health crisis.

Next, the presence and implications of bot-generated content was analysed. Bot accounts were observed to be heavily associated with #comedy, #memes, #funnymemes, #cute, #fun, and #masks (Fig. 9). There is a high correlation of 0.5 between eigenvector centrality and the percentage of fake users, signifying that these bot accounts are located centrally within the co-hashtag network and therefore have influence (Figure 12). Zarei et al., the originators of the COVID-19 Instagram dataset, suspect that bots and memes could potentially spread misinformation [7]. Social bots may create public panic, propagating rumors over verified information, as was seen during the Ebola and Zika virus pandemics [14]. The #masks is highly-associated with bot accounts within the dataset, but the exact impact of this bot-generated content on the pandemic is unknown.

Lastly, sentiment analysis was performed to determine the degree of positive or negative attitude toward topics within communities from Figure 5. Two metrics were analyzed: caption content and comment content. It was found that Instagram posts within Cluster C ("conspiracy" community) had higher levels of negative comments than those in Clusters B (mental health community) and Cluster A (communities discussing mainstream topics such as fitness or beauty). Cluster B seemed to have a positive "mood", since the comments left by other users were found to be highly positive (Figure 11). In contrast, Cluster C seemed to have a comparatively more negative impact on mood. Hashtags previously found to be common among bots (e.g. #comedy and #memes) were not associated with spreading negative content (Figure 12). Lastly, there was a weak positive correlation ($r = +0.3$) between clustering coefficients and negative comments, which indicates the presence of "echo-chambers" of sentiment (Fig. 12).

Strengths and Limitations

There were some limitations identified in this study. A small training dataset was used for the bot-detection program. Efforts were made to reduce overfitting, but a larger dataset would lend more credibility to the results. As well, the BERT model was trained on an English dataset, so the sentiment analysis only analyzed English texts. It should be noted that the most negatively-labelled communities within Figure 5 were non-English texts. The significance of this is undetermined, as the original Sentiment-140 Twitter data that fine-tuned the BERT model consisted solely of English-language Tweets. Lastly, the Sentiment-140 training-dataset originates from Twitter, lending to a possible mismatch when analysing captions and comments from Instagram.

This study has many strengths. The clustering coefficient algorithm for cohashtag networks flags outliers in the dataset and provides a useful, single-value metric to identify the "nicheness" of conversations. State-of-the-art techniques were applied to this study, such as the BERT neural network. The resulting analysis pipeline is extremely flexible and allows for a wide variety of future applications. Ultimately, this study is one of the first-ever attempts to analyse public response to the COVID-19 pandemic on Instagram. This novel pipeline represents a pioneering attempt in a field currently only analysing data from Twitter.

Future Directions

Our study serves as a starting point to support research in:

1. Further characterization of pandemic behaviours: Examine the breakdown of business users types (e.g. Product/Service, Nonprofit Organization, Entrepreneur etc.). Further categorize Instagram Creator Accounts (public figures, content producers, artists, influencers).

2. Other social media platforms and topics of interest: A hashtag-acquiring program can be attached to the current pipeline to perform SNA on other topics of interest. Researchers can also follow the SNA pipeline to analyse datasets from other platforms (e.g. Reddit).
3. Mainstream trends, mental health and conspiracy: Research should be conducted to evaluate the influence of communities A, B, and C. Specifically, we recommend exploring the potential spread of misinformation and identifying areas where policymakers and governments can target resources to support mental health of communities.
4. A large-scale behavioural analysis of bots and their effects on COVID-19: Since the bot-detection training set was relatively small, further research is needed to analyze the influence of bot-generated content. A separate investigation is required to determine whether bots play a role in creating negative sentiments, and whether these reach into the spheres of health and misinformation. This can better inform policies regulating social bots, and potentially counter inaccurate or unverified health claims on the internet [15].

Conclusion

This study leveraged big data to characterize public behaviour amidst the COVID-19 pandemic, adding to the existing body of conclusions that can be drawn from social media data. Through Social Network Analysis, supervised learning, bot detection, and sentiment analysis, key components of the Instagram dataset were analyzed. Instagram hashtags were used to map out social interactions, represented by nodes in a visual network showing the frequency and importance of each within the dataset. Three main clusters were identified ("Mainstream Trends", "Mental Health", and "Conspiracy Theories") and notable findings were the link between mental health and Instagram Business Accounts, as well as a high number of genuine users engaging in conspiracy spheres. We conclude that the spread of information on social media requires monitoring to determine strategies for countering inaccurate health claims. Investigations of this sort can be performed with sentiment classification. This is important considering the proportion of bot-generated content related to masks; and the presence of a distinct "conspiracy" cluster. Moreover, our results show value in exploring the intersectionality of mental health and COVID-19. Examining both niche and mainstream trends within mental health can offer insight into population needs. Lastly, further study is needed to identify for what reasons Business Profiles are associating their content with mental health, in response to the coronavirus health crisis.

Acknowledgements

The authors of this paper would like to acknowledge the STEMFellowship in providing an opportunity for undergraduate students to further develop their data analysis skills. This paper would also not have been possible without the dataset collected by Koosha Zarei and his team at the Institut Polytechnique de Paris.

References

1. Global digital population as of October 2020. Statista. October 2020. https://www.statista.com/statistics/617136/digital-population-worldwide/.

2. Who director-general's opening remarks at the media briefing on covid-19 - 11 march 2020, World Health Organization. Mar 2020. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

3. Maria Nicola, Zaid Alsafi, Catrin Sohrabi, Ahmed Kerwan, Ahmed Al-Jabir, Christos Iosifidis, Maliha Agha, and Riaz Agha. The socio-economic implications of the coronavirus and covid-19 pandemic: A review. *International Journal of Surgery (London, England)*, Apr 2020.

4. Paul J. J. Welfens. Macroeconomic and health care aspects of the coronavirus epidemic: Eu, us and global perspectives. *International Economics and Economic Policy*, page 1–68, May 2020.

5. Most popular social networks worldwide as of January 2021, ranked by number of active users. Statista. January 2021. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

6. Siddique Latif, Muhammad Usman, Sanaullah Manzoor, Waleed Iqbal, Junaid Qadir, Gareth Tyson, Ignacio Castro, Adeel Razi, Maged Kamel Boulos, and Jon Crowcroft. Leveraging data science to combat covid-19: A comprehensive review. *IEEE Transactions on Artificial Intelligence*, 1(1):85-103, August 2020.

7. Koosha Zarei, Reza Farahbakhsh, Noel Crespi, and Gareth Tyson. A first Instagram dataset on COVID-19. *arXiv* preprint, April 2020.
8. Dawn Iacobucci, Rebecca Mcbride, and Deidre Popovich. Eigenvector centrality: Illustrations supporting the utility of extracting more than one eigenvector to obtain additional insights into networks and interdependent structures. *Journal of Social Structure*, 18:1–21, October 2017.

9. Naoki Masuda, Michiko Sakaki, Takahiro Ezaki, and Takamitsu Watanabe. Clustering coefficients for correlation networks. *Frontiers in Neuroinformatics*, 12:7, 2018.

10. Andres Laan, Gabriel Madirolas, and Gonzalo G. de Polavieja. Rescuing collective wisdom when the average group opinion is wrong. *Frontiers in Robotics and AI*, 4:56, 2017.

11. Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Re´mi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. *Association for Computational Linguistics Anthology*, 38–45, October 2020.

12. Statement by the Department of Defense on the Release of Historical Navy Videos. United States Department of Defense. 2020. https://www.defense.gov/Newsroom/Releases/Release/Article/2165713/statement-by-the-department-of-defense-on-the-release-of-historical-navy-videos/.

13. Adam M. Enders, Joseph E. Uscinski. The coronavirus conspiracy boom, April 2020.
14. Jon-Patrick Allem and Emilio Ferrara. Could social bots pose a threat to public health? *American Journal of Public Health*, 108(8):1005–1006, August 2018.

15. Don E. Schultz. Extending the extended self in the digital world. *Journal of Marketing Theory and Practice*, 22(2):143–145, December 2014.

Appendix

For references to the code used, please visit https://github.com/brenda-shen/2020-Big-Data-Challenge and https://github.com/brenda-shen/2020-Big-Data-Challenge/tree/master/Instagram-dataset-analysis/Appendix.