SOTICS 2022

The Twelfth International Conference on Social Media Technologies, Communication, and Informatics

ISBN: 978-1-68558-002-5

October 16 - 20, 2022

Lisbon, Portugal

SOTICS 2022 Editors

Nitin Agarwal, University of Arkansas at Little Rock, USA

SOTICS 2022

Forward

The Twelfth International Conference on Social Media Technologies, Communication, and Informatics (SOTICS 2022), held between October 16th and October 20th, 2022, continued a series of events on social eco-informatics, bridging different social and informatics concepts by considering digital domains, social metrics, social applications, services, and challenges. Academic and industrial contributions were expected on algorithms, mechanisms, models, services dealing with challenges in social eco-systems.

The systems comprising human and information features form a complex mix of social sciences and informatics concepts embraced by the so-called social eco-systems. These are interdisciplinary approaches on social phenomena supported by advanced informatics solutions. It is quit intriguing that the impact on society is little studied despite a few experiments. Recently, also Google was labeled as a company that does not contribute to brain development by instantly showing the response for a query. This contrasts with the fact that it has been proven that not showing the definitive answer directly facilitates a learning process better. Also, studies show that e-book reading takes more times than reading a printed one. Digital libraries and deep web offer a vast spectrum of information. Large scale digital library and access-free digital libraries, as well as social networks and tools constitute challenges in terms of accessibility, trust, privacy, and user satisfaction. The current questions concern the trade-off, where our actions must focus, and how to increase the accessibility to eSocial resources.

We take here the opportunity to warmly thank all the members of the SOTICS 2022 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to SOTICS 2022. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the SOTICS 2022 organizing committee for their help in handling the logistics of this event.

We hope that SOTICS 2022 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of social media technologies, communication and informatics.

SOTICS 2022 Chairs

SOTICS 2022 Steering Committee

Elina Michopoulou, University of Derby, UK

SOTICS 2022 Publicity Chairs

Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
Jose Luis García, Universitat Politecnica de Valencia, Spain

SOTICS 2022
Committee

SOTICS 2022 Steering Committee

Elina Michopoulou, University of Derby, UK

SOTICS 2022 Publicity Chairs

Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
Jose Luis García, Universitat Politecnica de Valencia, Spain

SOTICS 2022 Technical Program Committee

Lavanya Addepalli, Universitat Politecnica de Valencia, Spain
Md Shoaib Ahmed, Brain Station 23, Dhaka, Bangladesh
Millicent Akotam Agangiba, University of Mines and Technology, Tarkwa, Ghana
William Akotam Agangiba, University of Mines and Technology, Tarkwa, Ghana
Federico Martín Alconada Verzini, Universidad Nacional de La Plata, Argentina
Samer Al-Khateeb, Creighton University, US
Musfique Anwar, Jahangirnagar University, Bangladesh
Hassan Atifi, Troyes University of Technology (UTT), France
Faical Azouaou, ESTIN - Ecole Supérieure en Sciences et Technologies de l'Informatique et du Numérique, Amizour, Algeria
Qanita Bani Baker, Jordan University of Science and Technology, Jordan
Asif Ali Banka, IUST Kashmir, India
Grigorios N. Beligiannis, University of Patras, Greece
Hernane Borges de Barros Pereira,Centro Universitário Senai Cimatec, Brazil
Christos Bouras, University of Patras, Greece
James Braman, Community College of Baltimore County, USA
Miguel Carvalho, INESC-ID Lisboa, Portugal / Coordinating Center for Communications and Information and Innovation Technologies, Regional Government of the Azores
K C Chan, University of Southern Queensland, Australia
Luisa Fernanda Chaparro Sierra, Tecnologico de Monterrey, Mexico
Dickson K.W. Chiu, University of Hong Kong, Hong Kong
Joshua Chukwuere, North-West University (NWU), South Africa
Subhasis Dasgupta, San Diego Supercomputer Center | University of California San Diego, USA
Dimitri Demergis, Rowan University, USA
Vasily Desnitsky, SPIIRAS, Russia
Nicolás Díaz Ferreyra, University of Duisburg-Essen, Germany
Arianna D'Ulizia,National research council of Italy - IRPPS Research, Italy
Ritam Dutta, Surendra Institute of Engineering & Management | Maulana Abul Kalam Azad University of Technology, West Bengal, India
Abderrahim El Amine, University of Technology of Troyes, France
Luis Enrique Sánchez Crespo, Universidad de Castilla-La Mancha, Spain
Larbi Esmahi, Athabasca University, Canada
Raji Ghawi, Technical University of Munich, Germany

Carlo Giglio, University of Calabria, Italy
Apostolos Gkamas, University Ecclesiastical Academy of Vella of Ioannina, Greece
Barbara Guidi, University of Pisa, Italy
Ekta Gujral, University of California - Riverside / Walmart Inc., USA
Gunjan Gupta, Lightsphere AI Inc, USA
Mahmoud Hammad, Jordan University of Science and Technology, Jordan
Lingzi Hong, University of North Texas, USA
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Anush Poghosyan Hopes, University of Bath, UK
Hana Horak, University of Zagreb, Croatia
Pedram Hosseini, George Washington University, USA
Muhammad Nihal Hussain, University of Arkansas at Little Rock, USA
Sergio Ilarri, University of Zaragoza, Spain
Makoto Itoh, University of Tsukuba, Japan
Igor Jakovljevic, CERN - Open Search Foundation / Graz University of Technology, Austria
Maria João Simões, University of Beira Interior / CICS.NOVA / LabCom.IFP, Portugal
Hanmin Jung, Korea Institute of Science and Technology Information, Korea
Evgeny Kagan, Ariel University, Israel
Attila Kertesz, University of Szeged, Hungary
Tiffany Kim, HRL Laboratories LLC, USA
Yannis Korkontzelos, Edge Hill University, UK
Satoshi Kurihara, Keio University, Japan
Konstantin Kuzmin, Rensselaer Polytechnic Institute (RPI), USA
Johannes Langguth, Simula Research Laboratory, Norway
Jinfeng Li, University of Southampton, UK
Dongxin Liu, University of Illinois, Urbana-Champaign, USA
Yidu Lu, Twitch, USA
Munir Majdalawieh, Zayed University, United Arab Emirates
Estela Marine-Roig, University of Lleida, Catalonia, Spain
Philippe Mathieu, CRIStAL Lab | University of Lille, France
Susan McKeever, Technological University Dublin, Ireland
Kai Meisner, University of the Armed Forces in Munich, Germany
Abdelkrim Meziane, CERIST, Algeria
Konstantsin Miatliuk, Bialystok University of Technology, Poland
Elina Michopoulou, University of Derby, UK
Salvatore Monteleone, CY Cergy Paris Université, France
Jenny Morales Brito, Universidad Autónoma de Chile, Chile
Marcel Naef, University of Zurich, Switzerland
Andrea Nanetti, School of Art, Design, and Media | Nanyang Technological University, Singapore
Cuong Nguyen, Investors' Business Daily, USA
Anastasija Nikiforova, University of Latvia, Latvia
Debora Nozza, University of Milano - Bicocca, Italy
Antonio Opromolla, Link Campus University, Rome
María Óskarsdóttir, Reykjavík University, Iceland
Rachid Ouaret, INP de Toulouse -ENSIACET, France
Luigi Patrono, University of Salento, Lecce, Italy
Cindarella Petz, Technical University of Munich /Bavarian School of Public Policy, Germany
Scott Piao, Lancaster University, UK

Nadja Piedade de Antonio, Universidade Federal do Estado do Rio de Janeiro (UNIRIO), Brazil
Maria Pilgun, Institute of Linguistics - Russian Academy of Sciences, Moscow, Russia
Agostino Poggi, Università degli Studi di Parma, Italy
Vassilis Poulopoulos, University of Peloponnese, Greece
Elaheh Pourabbas, National Research Council of Italy, Italy
Bhavtosh Rath, Target Corporation, USA
Henry Rosales-Méndez, University of Chile, Chile
Debashish Roy, Ryerson University, Toronto, Canada
Cristian Rusu, Pontificia Universidad Católica de Valparaíso, Chile
Mirco Schönfeld, University of Bayreuth, Germany
Ali Shahrabi, Glasgow Caledonian University, Scotland, UK
Soroosh Shalileh, National Research University Higher School of Economics, Moscow, Russia
Vivek Shandilya, Jacksonville University,USA
Anurag Singh, National Institute of Technology, Delhi, India
Evaggelos Spyrou, University of Thessaly | NCSR Demokritos, Greece
Wienke Strathern, Technical University of Munich, Germany
Raquel Trillo-Lado, University of Zaragoza, Spain
Lorna Uden, Staffordshire University, UK
Stefanos Vrochidis, ITI-CERTH, Greece
Gang Wang, Hefei University of Technology, China
Huadong Xia, Microstrategy Corporation, USA
Chenwei Zhang, The University of Hong Kong, Hong Kong

Copyright Information

Table of Contents

A Quantitative Social Network Analysis of Politicians' Tweets to Explore Political Communication

Heidi Schuhbauer, Sebastian Schötteler, Johannes Niu, Bernhard Schiffer, David Wolfarth
Computer Science Department
Nuremberg Institute of Technology
Nuremberg, Germany
heidi.schuhbauer@th-nuernberg.de; sebastian.schoetteler@th-nuernberg.de

Abstract— **This paper illustrates the practical application of cluster analysis, social network analysis and sentiment analysis in a case study. These techniques provide insights into the public communication patterns between German Members of Parliament (MPs) on Twitter around the time of the 2021 federal election. The question of this work was to determine whether a potential shift in communication towards the inaugurated "Ampel" coalition, made up of the parties SPD, Greens and FDP, can be derived from Twitter interactions. In distinct scenarios, mention, retweet, and reply interactions are first considered together and then separately. In these scenarios, the Girvan-Newman Algorithm detects clusters of MPs dependent on the interactions observed. Then, the average inbreeding homophily and other network metrics of the pre- and post-election area are compared. An additional scenario focuses on intra- and inter-party sentiments conveyed within tweet texts. In a fourth scenario, MPs are grouped according to their party affiliation, the average inbreeding homophily values of parties and potential coalitions. The communication clusters of those MPs differ mostly before and after the election. The average sentiment of the parties towards each other changed positively, although no significant tendency could be derived regarding later coalition formations.**

Keywords-Cluster Analysis; Microblog; Network Metrics; Sentiment Analysis; Social Network Analysis.

I. INTRODUCTION

For the political communication between parties, politicians and their constituents, social media platforms play an important role. By communicating through platforms, such as Facebook, Twitter, and Instagram, political actors reach wide audiences within a short period. On these platforms, politicians publicly communicate with each other.

Among those services, Twitter promotes the dialogue between politicians and between politicians and their constituents via mention and reply interactions, which allow users to engage in direct communication. Consequently, social media have become a central component of political communication.

Relations between individual MPs can be examined in more detail using social network analysis. Interactions can be derived from public tweets referring to other people, i.e., retweets of or replies to another user's tweet, or mentions of a user. Analysis of interaction networks explore these relations, as well as their textual contents, which can be examined through sentiment analysis. This article applies such methods to explore changes to the communication of German MPs from selected political parties around the 2021 federal election.

Section 2 of this paper presents related works, formulates the research gap and specifies the hypotheses. Section 3 introduces the methodology used to aggregate and analyze the data. Section 4 presents the results of each perspective and discusses them. Section 5 illustrates the limitations of this research, as well as starting points for possible future work.

II. RESEARCH GAP

Virk [1] compares different Social Network Services (SNS) as a type of social media and explores the special role of Twitter in public communication. The author examines the communication patterns between Twitter users and applies the tie strength theory postulated by Granovetter [3] to conclude that interactions on Twitter – unlike other SNS – focus on content rather than user relationships, and thus can reach wider audiences.

Lassen and Brown [2] examine the Twitter use of members of congress in the United States of America. They state that SNS enable politicians to communicate more directly and personally with peers and supporters by eliminating limits on message visibility, allowing content to be redistributed beyond one's own followers. The application of social network analysis to political networks shows the fragmentation and clustering of politicians, parties, or political systems.

Boireau [4] identifies communities among Belgian MPs along party and linguistic lines. For this purpose, the Girvan-Newman Algorithm (GNA) was applied on a network generated from the MPs' connections to followers, and retweet interactions to find hidden communities and homogeneous clusters by calculating their homophily indices, which express the degree of similarity of members within a cluster.

Caetano et al. [5] analyze social networks between Twitter users during the 2016 American presidential election by analyzing tweets about the candidates. Users were clustered based on their sentiment towards a candidate with their mentioning behavior and hashtag use. By obtaining homophily indices of these clusters, the authors could identify users with high degrees of relative similarity.

Sentiment analysis attempts to quantify attitudes conveyed in a text. Giachanou and Crestani [6] discuss common procedures for sentiment analysis, as well as their respective limitations, e.g., the detection of irony or emotions. The work

ISBN: 978-1-68558-002-5

explicitly focuses on methods suitable to retrieve sentiments from tweets.

Until now, literature does not describe possible changes in Twitter communication behavior between MPs before and after an election. An exploration of the change in tone by analyzing the sentiment of tweets before and after an event has also not yet been described. Interesting aspects of political communication behavior on social media are expected results of this analysis.

Consequently, this article examines how Twitter interactions (mentions, retweets, replies) between MPs of possible coalition partners (CDU, CSU, SPD, Greens, FDP) changed before and after the 2021 German federal election. It furthermore explores potential differences in intra- and inter-party communication and attempts to show whether the political shift towards the inaugurated "Ampel" coalition could be derived from the observed changes.

The following hypotheses form the basis for the communication behavior analysis: The article hypothesizes that different interactions between MPs can be observed during the pre- and post-election period (H1) and that the resulting interaction networks for each period show a difference in intra- and inter-party communication (H2). The article further assumes that "Ampel" MPs' mutual sentiment changed positively (H3). By analyzing the sentiment between parties, as well as the average homogeneity within parties and party groups, political tendencies towards an "Ampel" coalition can be observed (H4).

Thus, this article attempts to describe the change in communication between MPs by analyzing their Twitter interactions before and after the federal election 2021. It aims to understand whether changing interaction intensities between MPs of potential coalition partners yield conclusions about the emerging "Ampel" coalition. This would be of relevance for future research into the interdependencies of political communication on Social Network Services, such as Twitter.

III. Methods

Mention, retweet and reply interactions between MPs from the SPD, Greens, FDP, CDU, and CSU were collected to explore changes in communication on Twitter. One MP using another MP's handle denotes a mention interaction. Retweets refer to the redistribution of another user's tweet and can contain commentary by the retweeter. A reply is defined as a comment posted under another MP's tweet. The resulting social networks of MPs connected by their interactions is analyzed in four separate scenarios.

A. Network Scenarios

Scenario 1 considers all interaction types, while in scenario 2, a) mention, b) retweet, and c) reply interactions were examined separately. For each scenario, MPs were grouped using automated cluster detection and examined for modularity and homophily.

In scenarios 3 and 4, MPs were grouped based on their party affiliation. In scenario 3, interactions were examined for the tweet author's sentiment towards the addressed MP using sentiment analysis. The sentiment for every interaction was evaluated based on the tweet's text. To determine changes to the inter-party relations, each party's average sentiment toward all other parties was then calculated and compared between the pre- and post-election networks. Scenario 4 examined the average homophily within each party and party group. Party groups were based on politically and numerically possible coalition compositions ("Ampel", "Jamaica") and for the Union parties.

B. Data Aggregation

Publicly available Twitter data can be divided into three categories: (1) User information, such as the username, the Twitter handle (identified by @), or account description; (2) following and liking behavior of a user, and the user's followers; (3) the user's tweet timeline, in which all self-published or retweeted tweets appear, as well as the user's replies to others' tweets.

As a basis for this study, publicly available tweets from MPs of the 19th (2017-2021) and 20th (2021-2025) legislative sessions were collected for the period from July 26, 2021, 0:00 a.m. to November 26, 2021, 12:00 p.m. The end date was chosen to serve as cut-off due to the official presentation of the coalition agreement between the SPD, Greens, and FDP on November 24, 2021. To collect reactions to this announcement, two more days were added. The period between the closing of polls on September 26, 2021, 6 p.m. and the end date covers 60 days and is considered as the post-election period. An equally long time before the closing of polls was considered for the pre-election period.

Twitter accounts were selected from all MPs with a public Twitter timeline who are members of the parties SPD, CDU, CSU, Greens, and FDP. Members of the parties "The Left" and AfD were not included in this analysis, as neither party was relevant for coalition negotiations after the election. The timelines of all selected accounts were then scraped from Twitter's website.

Data Collection. Scraping of timelines was done using the Python package Scweet [7]. Scweet uses the Chrome plugin Selenium [8] to access the desired Twitter page, to extract the information of the tweet from the page and save it to a CSV-file.

Data Processing. A custom Java application was developed to generate uniformly formatted and sanitized datasets. The data originally scraped from Twitter included the timelines of all MPs, i.e., all their tweets, retweets of and replies to other tweets within the time frame. The information generated for each of these messages included the time of publication, the author's username and handle, the textual contents of the tweet, as well as information on whether it was posted as a retweet of, or reply to another tweet. If other users were mentioned within the tweet, they could be identified through their handle.

Additionally, the application enriches the data with information on party affiliation and membership of the 19th or 20th legislative period. It produced output data in the GEXF-format [9], which is limited by specified procedures. First, all tweets that did not represent a connection between two MPs were removed. The dataset was then divided into a pre- and a post-election partition. For this purpose, all tweets that were created before the time of the closing of polls on September 26, 2021, 6:00 p.m. were assigned to a first partition. The elements from the timeline after this date were assigned to a second partition. Additionally, the output is restricted to specific interaction types. This allowed the creation of one pre-election and one post-election dataset for each of the scenarios defined.

Data Description. The data set collected from Twitter consisted of 26,888 German language tweets from 736 Twitter accounts. 15,770 of these tweets were posted before and 11,118 after election day. 1,030 MPs were elected for the 19th and 20th legislative periods. 71.5% of them maintained a Twitter account. Once filtered, the dataset consisted of 622 accounts and 9,582 tweets. After removing all tweets that did not connect two MPs, 5,766 tweets from 466 MPs remained in the pre-election dataset and 3,816 from 476 MPs in the post-election dataset. Figure 1 shows the percentage distribution of all tweets among the parties before and after the election.

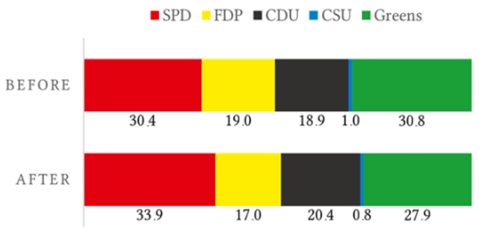

Figure 1. Percentage distribution of MPs' tweets by party

The pre- and post-election data contain nodes and edges depending on the interaction types selected during the data processing step. Scenarios 1 and 4 thus contained all MPs, while scenario 2 contained three separate data sets, differentiated by interaction types. Scenario 3 handled only those interaction types whose tweet text field were not empty. The aggregated data and source code can be accessed at [13].

C. Cluster Detection

Cluster detection extracts groups of individuals from a network based on similarity of one or more attributes. This work used connectivity-based clustering, which identifies clusters based on the connections between nodes in the network, as well as the weights of connections. For this purpose, the Girvan-Newman Algorithm [10] was used. This algorithm assumes that members of a cluster have more connections to other members of the same cluster, and fewer connections to other nodes in the remaining network. By iteratively removing connections whose Edge Betweenness Centrality (EBC) is the highest, clusters are separated from each other. The EBC is defined as "the number of [the] shortest paths between pairs of vertices that run along it" [10]. In each step, the edge with the highest EBC is removed from the network and its modularity is calculated. The modularity of a network denotes how well clusters are separated from each other. The iteration continues until every connection between nodes has been eliminated. The intermediate step with the highest modularity is the result of the algorithm.

To guarantee that an MP's allocation to a cluster is based on their interactions and not their party affiliation, a χ^2 test is performed on the network. The test's p-value denotes the probability p of MPs' party affiliation determining the results of the cluster detection.

D. Sentiment Analysis

The textual contexts of MPs' tweets were examined to analyze the sentiment for which the Python package TextBlob [11] was used. The package uses a lexicon-based approach to compute the sentiment. For the analysis of German language texts, the plugin TextBlobDE [12] was used. A predefined dictionary of words associated with positive or negative emotions is used to weigh a text's sentiment. An individual score is assigned to each word in the examined text. The overall sentiment is defined by the average sentiment across all words in the text. The algorithm generates a polarity score from –1.0 to +1.0 for each tweet, which classified the tweet as either positive, neutral, or negative. Each tweet in the data set is then enriched with the polarity value, as well as the polarity class as additional attributes.

E. Homophily

The homophily index *H* measures a cluster's relative homogeneity. To determine *H* for a cluster i, the connections of all nodes of the cluster are examined. Caetano et al. [5] calculate $H_i = \frac{s_i}{si+d_i}$ where s_i denotes homogeneous links, i.e., those that connect a node of class *i* to other nodes of the same class, while d_i denotes heterogeneous connections, i.e., those that connect a node of class *i* to nodes of another class. By normalizing H_i over the whole network, *H* can be compared across different clusters. This inbreeding homophily index *IH* is determined by $IH_i = \frac{H_i - w_i}{1-w_i}$, where w_i denotes the relation of nodes between cluster *i* and the total number of nodes in the network. Clusters whose IH_i is greater than 0 are considered homogeneous. The average of *IH* across all clusters in a network is used to compare the clusters detected in the pre- and post-election networks.

F. Evaluation

The procedure resulted in a set of network pairs, each consisting of a pre- and a post-election network. The two networks created for scenario 1 contained all MPs that have interacted via mentions, retweets, or replies within the respective timeframe. The number of connections between two nodes weighted the edges.

Scenario 2 generated one network pair for each of the three interaction types. Thus, one pre- and one post-election network each were generated which included all those MPs that a) mentioned each other, b) replied to one another, and c) retweeted each other. Edges represent the connections. They are weighted by the interaction count. These scenarios were examined separately. For each network automated cluster detection was applied. The *H* and *IH* indices were calculated to determine the homogeneity of each cluster. Additionally, the number of nodes and edges in the network, the number of clusters identified by the GNA, as well as their networks' average homophily and inbreeding homophily indices and the maximum modularity were determined. Statistical significance was ensured using the χ^2 test. The results of these analyses were then compared for the pre- and post-election network pair. To illustrate the results of the automated cluster detection, each pre- and post-network pair is visualized as a cluster graph.

In scenario 3, each party's average sentiment towards all other parties was examined. For this purpose, MPs were clustered according to their party affiliations.

Scenario 4 looked at the inbreeding homophily of each party, as well as the coalition options before and after the election. The *IH*-values for the coalitions where also checked for statistical significance using the χ^2 test and its p-value.

IV. Results

A. *Scenario 1: Multiple Interactions*

In scenario 1, automated cluster detection included all interaction types. An overview of the collected metrics can be found in Table I.

TABLE I. NETWORK AND CLUSTER METRICS CONSIDERING ALL INTERACTIONS

Metric	Value (pre)	Value (post)	Difference
Number of nodes	466	476	10
Number of edges	5766	3816	-1950
Number of clusters	256	188	-68
Maximum modularity	0.026	0.356	0.330
Average *IH*	0.0212	0.0571	0.0359
p-value from χ^2-Test	< 0.001	< 0.001	

The number of MPs (nodes) tweeting after the election did not vary much from that before the election. However, the number of connections (edges) was reduced by 33%, which suggests that tweeting activity was distributed more equally among MPs after the election. The GNA identified 256 clusters of the pre-election network with 466 MPs, and very low modularity, homophily and inbreeding homophily indices. After the election, 476 MPs could be assigned to 188 clusters. The maximum cluster size was reduced by 54.5% to 97. The modularity increased by 1369%, from 0.026 to 0.356, and homophily and inbreeding homophily also increased significantly. Figure 2 shows a visualization of these clusters. Node colors represent each MP's party affiliation. The size of a node depicts the sum of all incoming and outgoing edges, i.e., the node's degree. Edges were omitted from these figures for improved visibility.

Pre-election, the visualization shows a distinctive, large cluster which unites MPs across all parties. Outside of this cluster many MPs are scattered into tiny groups or unassigned to any notable cluster. Post-election, four large clusters separated along party affiliation can be identified. A heterogeneous group of MPs was not assigned to any notable cluster.

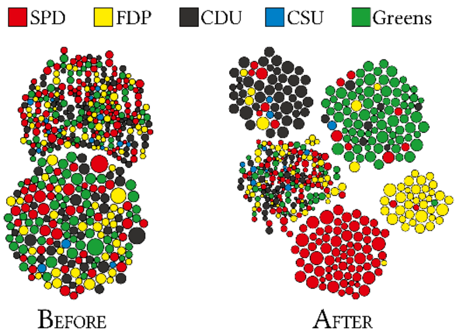

Figure 2. Clusters found by GNA before and after election considering all interactions

The pre-election results of scenario 1 show that MPs were likely allocated to the dominant cluster based on their general activity on Twitter. Nodes with higher degrees were allocated to the dominant cluster. Post-election, distinct clusters are clearly separable, which consist mainly of MPs of either the SPD, CDU, Greens or FDP. The number of nodes that could not be allocated to any major cluster decreased. This indicates that post-election, MPs predominantly communicated within their own parties, while they communicated much more openly before the election. The overall count of interactions decreased significantly.

B. *Scenario 2: Single Interactions*

When interaction types are considered separately, these findings can be analyzed in more detail.

Mentions. In this particular scenario, clusters were determined based on mentions only. Table II shows the collected metrics.

TABLE II. METRICS OF NETWORK AND CLUSTERS DERIVED FROM MENTIONS

Metric	Value (pre)	Value (post)	Difference
Number of nodes	433	428	-5
Number of edges	3247	1758	-1489
Number of clusters	95	38	-57
Maximum modularity	0.237	0.441	0.204
Average *IH*	0.1158	0.4550	0.3292
p-value from χ^2-Test	< 0.001	< 0.001	

Almost as many (433 vs 428) MPs mentioned one another in the pre- and post-election period. Interactions decreased by 54%, and the number of detected clusters decreased by 40%. After the election, 38 clusters with a modularity of 0.441 could be identified, compared to 95 clusters with a modularity of 0.237 before the election. Average IH across all clusters in both networks increased by more than 300%. Figure 3 visualizes the detected clusters.

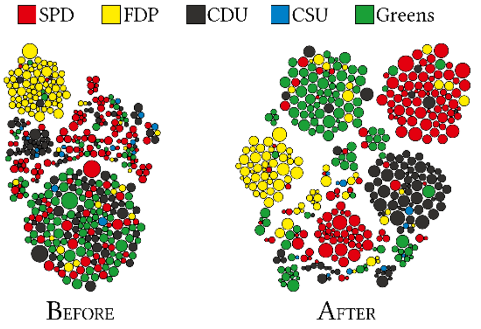

Figure 3. Clusters found by GNA before and after election considering only mentions

Pre-election, three distinct clusters can be identified, one portraying a large cluster mainly dominated by Greens but including MPs across all parties, one dominated by FDP MPs, and a smaller one dominated by CDU MPs. The large, heterogeneous cluster dominated by Green MPs could be caused by many mentions of the Greens' chancellor candidate, Annalena Baerbock.

Distinct clusters are detected in the post-election network separated along party lines. Two SPD clusters are found, as well as several smaller but still homogeneous clusters. The number of mentions increased. A subsequent analysis revealed that the distinct party clusters might be caused by MPs congratulating their party peers.

Retweets Cluster analysis detected several well-separated clusters with relatively high homogeneity before and after the election. A possible explanation is that MPs attempted to promote tweets of party peers. The clusters in the post-election network were smaller. Retweets play a smaller role in the communication among MPs.

Replies. Solely considering reply interactions, one large and many small clusters were found in the pre-election network. The main cluster contains many nodes with a high in– and out-degree. In the post-election network, more nodes are identified but fewer connections between them are found. Two main clusters were identified, notably consisting mainly of SPD and Green party members. One cluster of CDU and FDP MPs indicates active conversations between these two parties, potentially on the FDP's willingness to enter coalition negotiations with the SPD and Greens shortly after the election which supports hypothesis H1.

C. Scenario 3: Sentiment Analysis

Each interaction's textual content was analyzed to retrieve the parties' mutual sentiment. The average sentiment of interactions from MPs of one party towards MPs of the other parties was calculated. The results are shown in Table III. Notably, polarity does not score very highly overall, except for the sentiment from MPs of the CSU towards MPs from the CDU. FDP MPs communicated neutrally in general. The SPD scores positively towards the "Ampel" parties. On average, Green party MPs showed positive polarities only towards other MPs of their own party.

TABLE III. AVERAGE SENTIMENT BETWEEN PARTIES BEFORE THE ELECTION

Target / Source	SPD	FDP	CDU	CSU	Greens
SPD	0.25001	0.21293	0.00002	-0.11499	0.35683
FDP	0.05095	-0.01008	0.06981	0.09734	0.01032
CDU	0.00070	0.02997	0.13179	-0.12469	0.00483
CSU	0.04297	-0.01875	0.70728	0.10625	0.11405
Greens	-0.16582	0.03257	-0.16458	0.00053	0.35588

Table IV shows the average sentiment between parties after the election. The post-election sentiments between parties notably tend towards an overall positive sentiment. The SPD received overall positive interactions, especially from the CDU. The SPD communicated relatively neutrally, both internally, as well as towards their subsequent coalition partners. The polarity of the interactions among MPs of the Greens and interactions from MPs of the CSU towards CDU MPs did not change significantly from their pre-election scores. The overall sentiment across all parties after the election was on average more positive than before the election. The FDP especially shows notable increases in positive sentiments towards the SPD and the Greens, considering that the FDP moved towards the "Ampel". This strongly hints at successful coalition negotiations which ended with the signing of the coalition contract.

TABLE IV. AVERAGE SENTIMENT BETWEEN PARTIES AFTER THE ELECTION

Target / Source	SPD	FDP	CDU	CSU	Greens
SPD	0.00166	0.25408	0.31106	0.84063	0.06433
FDP	0.33102	0.54495	-0.11953	0.00391	0.27281
CDU	0.79865	-0.00598	0.09291	-0.08487	0.67012
CSU	-0.16250	0.24688	0.59688	0.12500	0.39146
Greens	0.43225	0.09978	0.62791	-0.06024	0.38109

D. Scenario 4: Party and group dependent clustering

In this scenario, MPs were clustered along party affiliation. Additionally, the two potential government coalitions, "Ampel" (SPD, Greens, FDP) and Jamaica (CDU, CSU, Greens, FDP), as well as the Union (CDU, CSU), were clustered. To compare the homogeneity within each cluster, the average *IH* before and after the election was calculated and compared. Table V displays the average *IH* values of each party, as well as the coalition and union clusters for the pre- and post-election networks.

The biggest differences are within the SPD and CDU. Their relative homophily increased. CSU and FDP decreased in *IH*. SPD received the biggest increase in homogeneity. This could be explained by their win of the election, and the positive feedback MPs received from their peers, as well as the election of SPD MPs Olaf Scholz as chancellor and Bärbel Bas as president of the parliament. The biggest positive change among grouped MPs took place in the "Ampel" coalition, but *IH* increased for the Jamaica and Union clusters as well. However, a significant statistical independence of these findings is not reliably provable, as the χ^2-test results in relative high p-values for the pre- and post-election homophily.

TABLE V. RELATIVE IH IN PARTIES AND PARTY GROUPS

	Before	**After**	**Difference**
CDU	0.4749	0.5596	0.0848
CSU	0.0721	0.0516	-0.0205
SPD	0.5272	0.6392	0.1120
Greens	0.5682	0.5729	0.0047
FDP	0.5397	0.4618	-0.0779
“Ampel” Coalition	0.4519	0.6272	0.1754
Jamaica Coalition	0.5209	0.5307	0.0098
Union Group	0.4632	0.5422	0.0791
p-value from χ^2-Test	0.057764	0.106983	

V. CONCLUSION

This paper illustrates the application of techniques from social network analysis, sentiment analysis and cluster analysis in combination to analyze communication on social media especially on micro blogs.

H1 is proven, as differences are found for mention and reply interactions. The networks for each interaction type yield differences in both intra- and inter-party interactions, which is shown by the results of the GNA. These findings are statistically significant due to the low p-values. H2 can therefore be considered as true. The p-value of the χ^2 test indicates a low likelihood that party affiliation influences the assigned cluster.

H3 cannot be answered clearly. MPs’ mutual sentiment changed positively. The FDP’s positive change towards the coalition partners SPD and Greens can be considered as a sign of a generally improved attitude towards these parties. However, the notable overall increase in positivity across most parties could indicate that the findings of the FDP are not unique. The generally positive attitude between parties after the election can be caused by MPs congratulating one another. A lack of German language sentiment analysis models for short text fragments limits this research. Improved models utilize machine learning techniques and so can comprehend sentiments on a broader level and can also recognize nuances.

Statements about H4 are not reliable. However, while positive tendencies towards an “Ampel” coalition can be shown from both the sentiment analysis and the inter-party and intra-coalition homogeneity, neither can be proven as statistically significant.

Definitely results are: Different interactions between MPs can be observed during the pre- and post-election periods and the resulting interaction networks for each period show a difference in intra- and inter-party communication. However, this paper handles the political communication only via Twitter. Results are partially transferable to other countries.

Future work may include “The Left” and AfD in these considerations to produce more information. Expanding the evaluated timeframes or continuous monitoring would produce more data. Analyzing follower and friend networks and MPs’ liking behavior in combination with the findings of this article would yield insights into differences in parties’ mutual relationships around elections.

REFERENCES

[1] A. Virk, “Twitter: The Strength of Weak Ties”, University of Auckland Business Review, Vol. 13, No. 1. University of Auckland, Auckland, AUK, NZL, pp. 19-21, Jan 2011.

[2] D. S. Lassen and A. R. Brown, “Twitter: The Electoral Connection?”, Social Science Computer Review, Vol. 29, No. 4. SAGE Publications, Thousand Oaks, CA, USA, pp. 419-436, Nov 2011.
DOI: https://doi.org/10.1177%2F0894439310382749

[3] M. S. Granovetter, “The Strength of Weak Ties”, American Journal of Sociology, Vol. 78, No. 6. The University of Chicago Press, Chicago, IL, USA, pp. 1360-1380, May 1973.

[4] M. Boireau, “Uncovering Online Political Communities of Belgian MPs through Social Network Clustering Analysis”, Proceedings of the 2015 2nd International Conference on Electronic Governance and Open Society: Challenges in Eurasia (EGOSE '15). Association for Computing Machinery, New York, NY, USA, pp. 150-163, Nov 2015. DOI: https://doi.org/10.1145/2846012.2846049.

[5] J. A. Caetano, H. S. Lima, M. F. Santos, and H. T. Marques-Neto, “Using sentiment analysis to define Twitter political users’ classes and their homophily during the 2016 American presidential election”, Journal of Internet Services and Applications, Vol. 9, Article 18, Sep 2018. Springer Open, DOI: https://doi.org/10.1186/s13174-018-0089-0.

[6] A. Giachanou and F. Crestani, “Like It or Not: A Survey of Twitter Sentiment Analysis Methods” ACM Computing Surveys, Vol 49, No. 28. Association for Computing Machinery, New York, NY, USA, pp. 1-41, Jun 2017. DOI: https://doi.org/10.1145/2938640.

[7] Y. A. Jeddi, Scweet (Version 1.6), [Online]. Available from: from https://github.com/Altimis/Scweet, Dec 2021.

[8] The Selenium Project, Selenium (Version 4.1.0), [Online]. Available from: https://github.com/seleniumhq/selenium, Dec 2021.

[9] GEXF Working Group, GEXF File Format (Version 1.2), The Gephi Community Project, 2009. [Online]. Available from: http://gexf.net, Dec 2021.

[10] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks”, Proceedings of the National Academy, Vol. 99, No. 12. National Academy of Sciences of the United States of America, Washington DC, USA, pp. 7821-7826, Jun 2002.
DOI: https://doi.org/10.1073/pnas.122653799.

[11] S. Loria, TextBlob (Release v0.16.0) Documentation. TextBlob: Simplified Text Processing. [Online]. Available from: https://textblob.readthedocs.io/en/dev/index.html#, Dec 2021.

[12] M. Killer, textblob-de. (Version 0.4.3), German language support for TextBlob by Steven Loria. [Online]. Available from: https://github.com/markuskiller/textblob-de, Dec 2019.

[13] Datawork, [Online]. Available from: https://git.informatik.fh-nuernberg.de/wolfarthda82341/sna-germanys-members-of-parliament-on-twitter, Oct 2022.

Sentiment Analysis of Twitter Posts on COVID-19 Cultural Dimensions: Collectivist vs. Individualist

Daniel Dobler*, Leo Donisch†, Melanie Koeppel‡, Patricia Brockmann§
Computer Science Department Nuremberg Institute of Technology
Nuremberg, Germany
Email: *doblerda75546@th-nuernberg.de, †donischle75565@th-nuernberg.de,
‡koeppelme76459@th-nuernberg.de, §patricia.brockmann@th-nuernberg.de

***Abstract*—Social distancing requirements during the COVID-19 pandemic have led to an increase in the importance of social media to maintain communication channels. This paper describes an initial investigation to English Twitter posts about the COVID-19 epidemic. The goal is to determine whether differences in opinion between users from different cultural backgrounds can be discerned. As a first prototype, a classification of tweets according to collectivist and individualistic cultures is attempted. Training data is used to generate feature vectors to train a neural network. Sentiment analysis is employed to classify the posts as positive, negative or neutral. Potential consequences for education and possible adaptive measures for collectivist and individualistic cultures are suggested.**

***Index Terms*—social media; sentiment analysis; cultural; collectivist; individual; education.**

I. Introduction

Physical and social distancing requirements during the COVID-19 pandemic have made it more difficult for people to physically spend time with friends, family members and colleagues. The need to discuss experiences and exchange ideas with others remains a fundamental human need. To fill this void caused by restrictions of in-person meetings, the importance of social media channels has increased [1].

Attitudes toward contact restrictions imposed to combat the spread of COVID-19 have varied considerably among citizens of different countries. Levels of resilience in dealing with stress situations caused by lock-downs have also varied considerably around the world, especially among young people. Cultural dimensions, as described by Hofstede [2], may play a role in explaining some of these different responses. A high level of power distance may positively affect respect for positions of authority and thus increase acceptance of temporary restrictions. Conversely, a culture which highly values individualism may experience lower compliance with health regulations. In collectivist cultures, which value the group higher than the individual, people may willingly adhere to health measures, in order protect weaker members of the society. Depending on the cultural dimensions of a country, educational measures could be specially adapted to help students cope with pandemic measures.

The goal of this work is to build a proof-of-concept prototype to investigate whether it is possible to differentiate between tweets on Covid-19 from different cultures by applying sentiment analysis. Sentiment Analysis is defined as the computational analysis of opinion, analysis and subjectivity in text [3]. Using natural language processing techniques, text can be classified according positive or negative polarity. To perform this investigation, a large collection of Twitter posts were cleaned, pre-processed and their sentiments analyzed. An evaluation was made to determine whether different types of cultures express more positive, negative or neutral opinions on the COVID-19 virus. For this first prototype, a focus is placed on collectivist and individualist cultures.

The research questions examined in this study are:

- R1: Can sentiment analysis deliver meaningful insights into the opinions of the COVID-19 pandemic expressed in Twitter posts?
- R2: Do cultural dimensions associated with users from collectivist vs. individualist cultures affect their expressed opinions?

First, an overview of the related literature is surveyed in Section II. The methods employed in this work are described in Section III. In Section IV, initial results of the prototype model for sentiment analysis are presented. Finally, conclusions and plans for future work are discussed in Section V.

II. Related Work

A. Cultural Dimensions

Hofstede [2] was one of the first investigators to apply multivariate statistical methods to analyze data from a large, international survey of thousands of information technology professionals. The differences observed in cultural perspectives among respondents from different countries were scored according to six dimensions:

1) Power distance: How a society views inequalities between individuals
2) Collectivism vs. individualism: Preference for loosely or tightly-knit social frameworks
3) Masculine vs. feminine: Achievement and assertiveness vs. cooperation and caring
4) Uncertainty avoidance: Degree to which unknown or ambiguous situations are viewed as threatening
5) Long-term vs. short-term orientation: Thrift and planning for the future vs. challenges of the present
6) Indulgence: Immediate gratification vs. restraint.

In addition to these six cultural dimensions, Hall [4] differentiates between high and low context cultures. In low context cultures, explicitly written and spoken words are the primary

ISBN: 978-1-68558-002-5

source of meaning. Thus, communication in low context cultures can often seem quite verbose. Western countries, such as Germany, tend to be classified as low context cultures and also place a high value on individualism. In high context cultures, personal relationships between people, such as their level of familiarity or differences in societal status, can play an intrinsic role in communication. Unspoken communication, such as facial expressions, gestures and pauses can sometimes convey more meaning than the actual written or spoken words. East Asian countries, such as Japan, are classified as high context cultures and also tend to value collectivism.

Kim et al. [5] found major differences in the usage of social media by university students in Korea and the U.S. Korean students were more motivated to use social media to obtain social support from existing relationships, while American students were more interested in seeking entertainment. Korean students also had a smaller number of contacts in their networks than American students. This appears to coincide with Hofstede's [2] findings, which categorize Korea as a collectivist culture and the U.S. as an individualistic culture.

B. Sentiment Analysis of Social Media

A number of researchers have applied the technique of sentiment analysis to social media. Chakraborty et al. [6] conducted a widespread literature review of over 200 papers on the subject of social networks and the use of sentiment analysis in social media. They review different techniques of sentiment analysis and point out important challenges which should be addressed, such as rumor detection and community shaming. Strathern et al. [7] explored the use of sentiment analysis to detect so-called ”firestorms” on Twitter. These firestorms can be triggered by negative online dynamics, which result in uncontrollable escalation which result in real harm to people. Tsao et al. [8] performed a literature review of 81 studies on online social media and COVID-19. They identified five main public health themes: surveying public attitudes, identifying infodemics, assessing mental health, detecting or predicting COVID-19 cases, analyzing government responses to the pandemic and evaluating the quality of health information in prevention education videos. Their main criticism is the scarcity of studies documenting real-time surveillance with data from social media. Aggregated data from Facebook was found to show that COVID-19 is more likely to spread between regions with stronger social network connections [9].

Sentiment analysis conducted during a nationwide lockdown in one single country, India, showed that although a number of negative sentiments were expressed, such as fear, disgust, and sadness, the overwhelming sentiments were positive, especially trust [10]. A different study from India utilized sentiment analysis on tweets. They found that popularity has an effect on the accuracy of information disseminated over social media. The most popular retweets skewed highly negative and did not contain any significant information [11]. Another study compared topic modeling for English and Portuguese tweets related to COVID-19. They found that the top ten topics for both languages were mostly similar [12].

Kruspe et al. [13] analyzed Twitter messages collected during the first few months of the pandemic in Europe. They performed a sentiment analysis using multilingual sentence embeddings and separated the results by country of origin. They found that lockdown measures correlated with a deterioration of sentiment in almost all of the countries surveyed. A sentiment analysis Twitter messages from different countries was performed by Imran et al. [14]. They divided up countries into geographic regions. The U.S. and Canada were grouped together as North American countries. India and Pakistan were grouped together as South Asian countries. Sweden and Norway were grouped together as Nordic countries. They found a high level of correlation between countries in the North American group. The South Asian group also showed a high level of correlation within the group. Sweden and Norway, however, showed opposite trends in polarity. This result is quite surprising and inspires further inquiry.

One cause which may be explain highly different sentiments between geographically close countries may lie in different values for certain cultural dimensions. Sentiment analysis geared toward clusters of countries which share similar values on specific cultural dimensions has not yet been handled in the literature. This research gap is addressed in this paper.

III. Methods

The goal of this study was to investigate possible differences in the sentiment of groups with similar values of cultural dimensions. For this first proof-of-concept experiment, individualist vs. collectivist cultures were examined.

A. Data Source

This work was conducted on a publicly available data set which contains tweets about the COVID-19 pandemic. This data set is freely available on the open data platform on Kaggle and includes 44,955 tweets from March 12th through March 16th, 2020 [15]. Each data record includes information about the user, their screen name (encoded to preserve user privacy) and the date the tweet was posted. In addition to the text content of each tweet, information about the location where it was posted and the sentiment of the text was included.

The sentiment of each tweet (very positive, positive, neutral, negative, very negative) was determined manually by the author of the data set and serves as the label value which the algorithm used in this experiment attempts to predict. Manual labeling can be quite difficult, even for a human. Furthermore, individual, subjective opinions may also bias this labeling process. To get a more objective evaluation, three of the authors of this work manually labeled 100 tweets. In 41 % of these tweets, the three given labels were unambiguous. Through a majority decision, a label could be assigned to 96 % of the tweets. Four examples could not be labelled, since they received one vote each of positive, neutral and negative. For 47 % of the tweets, the appropriate value was determined. For 53 % of the tweets, the manual evaluation would have given a different label to the tweet than the author of our source.

B. Data Partitions

The data set first needs to be partitioned according to the country of origin. These countries will then be assigned to groups with similar values on cultural dimensions. Twitter users input their geographic locations as strings. As a consequence, a location can sometimes contain a country, federal state, city or any other character data. The only location data which can be easily mapped to a certain value for one cultural dimension is the country. Therefore, only the data samples which included a country as part of the location data were used in this first prototype experiment. To refine samples with an easy to convert country, all of the data samples with country values that consist of only one word were selected. In the next step, a function from Kaggle [16] was combined with the Python library pycountry [17] to recognize countries. The snippet was adjusted so that it could identify countries by the common name, the official name and ISO (International Organization for Standardization) alpha-3 code. The recognition of ISO alpha-2 codes was disabled, because they can often match with both a federal state or with a country.

Once all samples for the test data set were prepared, the location value had to be associated with the correct cultural dimension group: individualist or collectivist. To achieve this for all of the 152 extracted countries, a dictionary was established and used to map countries to one of the two groups. The strength of the score for the dimension of individualism vs. collectivism (IDV) in each country was compared to the values of Hofstede [2]. If the value for a country was greater or equal than 50, the sample was assigned the tag “I”, for an individualistic culture. If the value was below 50, it was assigned the tag “C”, for a collectivist culture.

C. Pre-processing Pipeline

In order to evaluate text automatically, it first must be cleaned of unnecessary information and then transformed into a format which can be analyzed. Fig. 1 shows the pre-processing steps conducted before the data analysis, based on the recommendations of [18].

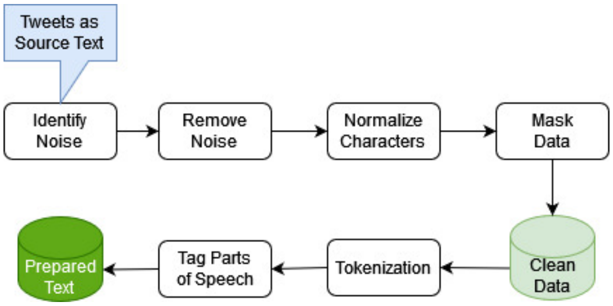

Fig. 1. Pre-Processing Pipeline

1) Data cleaning: One of the first steps for efficient noise removal is to correctly identify the noise in the given context. To achieve this, the length of each tweet was first calculated. A maximum number of 280 characters are allowed per tweet. If the length of a tweet exceeded this maximum, then it was flagged for closer analysis. Another measurement employed was the impurity score, which indicates the share of suspicious characters in a text. This enables recognition of noisy tweets and to measure improvements of data cleaning. For each tweet in the data set, its length and impurity score was calculated. If one particular tweet had a length longer than 280 characters and a high impurity score, this tweet was flagged for more detailed analysis. Tweets which didn’t exceed the maximum length but had an high impurity score were also flagged for further analysis. This approach did not always produce perfect results; not all of the noise could be effectively identified.

TABLE I
STOP WORDS

Before Removal	Polarity	After Removal	Polarity
The lockdown is good	pos	lockdown good	pos
The lockdown works	pos	lockdown works	pos
I did not like the lockdown	neg	like lockdown	pos
This lockdown is no good	neg	lockdown good	pos

Two successive cleaning approaches were implemented. First, artifacts of the extraction method are identified and removed. Next, a more specific cleaning method incorporates the insights of the analysis. These steps try to minimize tweet-specific patterns, such as URLs and user handles. These methods were implemented using functions, which were derived by Albrecht [18]. To summarize, these steps are necessary to remove unwanted patterns and simultaneously minimize word variants, in order to enable to learn a more precise language model [18] [19]. With all these measures, the tweets became cleaner and smaller, as measured by the impurity score.

An additional source of noise are so-called ”stop words” [20]. Stop words are parts of speech, such as prepositions, conjunctions or determinants, such as “and”, “or”, and “a”. Simply blindly removing these words from a text corpus is not ideal. If stop words such as “not” are removed, the sentiment of a tweet completely changes. This can cause problems, because the sentiment label is the opposite of what the tweet implied, as shown in Table I. In order to maintain the original sentiment, stop words which reverse the sentiment should not be removed from the tweets. They need to be removed from the default stop word list in the library SpaCy [20].

2) Tokenization: In order to use text in a machine learning algorithm, text must be correctly segmented into analyzable elements. This step is called tokenization and the results are referred to as n-grams. In the first step, each tweet gets segmented into one-word-tokens, called unigrams. Linguistic attributes are attached to each token, to achieve more precise n-grams in a later step. The process of segmenting text into smaller elements and attaching the linguistic attributes was done using a python library called SpaCy [21].

The linguistic attributes contain a boolean value, which indicates whether a given unigram is a stop word. When utilizing the stop word removal method described in Subsection III-C1, only non-sentiment changing words are removed. A further linguistic attribute is the Part-Of-Speech Attribute, which can be used to extract grammatical insights of a tokenized tweet [18]. Here, it is used to construct meaningful n-grams. Building n-grams is important if context information plays a key role in the analysis.

The library Textacy contains a useful function which allows for POS-tag pattern search, similar to regular expressions [22]. With this pattern-search function, a phrase which starts with an adjective and ends with a noun, so-called adjective-noun-phrases, can be extracted. Capturing sentiment adjectives plays a key role. The adjective-noun-phrases and adjective-adverb-phrases were extracted according to methods described in [18].

One major drawback is that not every tweet is structured grammatically correctly and therefore may not yield good results. Thus, using only bigrams is not ideal. For this reason, the default unigrams were extended with the adjective-noun-phrases found to capture context information. These tokens could then be passed into the desired vectorization method, which will be described in the next subsection.

D. Feature Engineering

In order to use machine learning algorithms with text, words need to be transformed into a numerical representation. One approach often used is called the "Bag of Words" method. In this approach, every token represents a phrase or word learned in the vocabulary. For each tweet, the frequency of how often each of these words occur is calculated. Thus, the first thing this method does is to learn the vocabulary list of the data set and then to transform each tweet into word frequencies. If a word in a new tweet is not already in the vocabulary list, it is ignored. In this way, the majority of the words captured must be contained in the training data set [18]. However, this approach is prone to over-weighting frequently used words and under-weighting less frequently used words. To counter this trend, the Term Frequency - Inverse Document Frequency (TF-IDF) value is calculated. The TF-IDF calculation boosts words which are less frequently used and slightly lowers the weighting of frequently used words [23].

Two additional parameters were also used: maximum and minimum document frequency. These parameters restrict the algorithm from learning words which occur too often or too seldom. A maximal document frequency of 80% is used in this work. This means that tokens which appear in more than 80% of the tweets were not be used. Tokens which appear in less than five tweets were also omitted. By limiting the size of the vocabulary, the dimension of the vector is reduced, thus making learning more efficient. The disadvantage is that some information does get lost, because the discarded terms could potentially carry important meaning [18]. In this case, both learning time and the metrics used both improved.

E. Neural Network

As an initial experiment, a pre-trained model which uses the Bidirectional Encoder Representations from Transformers (BERT) [24] from the TensorFlow repository was used. This is a non-case sensitive, smaller version of BERT, which was pre-trained for English on Wikipedia and BooksCorpus. A notebook for the data set investigated was published on Kaggle [15] for free use. The model distinguishes between five classes of the sentiment of a text. It uses a special tokenizer for BERT and one layer of the pre-trained BERT model, positioned ahead the other three layers of the neural network. In total, the model has 109,533,701 trainable parameters, which requires high computing times. In an experiment with an average Windows laptop, only 15 samples could be used due to the processing capacity limitations. Because of the small number of samples, only a 40% accuracy rate could be achieved. This accuracy rate is lower than flipping a coin and was thus judged to be too low for further pursuit.

As an alternative, a simpler model of SciKit-Learn (sklearn) with a Multi-Layer Perceptron Classifier (MLPClassifier) offers was implemented. Using the module "neural network" from sklearn, it was easy to adapt the model to fit the research goal and to adjust its parameters. In order to analyze the sentiment of Twitter posts, it would have been possible to use a number of different machine learning algorithms, such as a random forest algorithm. Neural networks were selected for this implementation, because they can easily be easily trained to recognize complex patterns [25]. During the training phase, the network can be fine-tuned. For example, the selection of methods for weight optimization or value of the learning rate provide a high amount of flexibility and fine tuning potential. The architecture of the neural network, the number of hidden layers and the number of neurons in each hidden layer can also have an impact on the results [25].

For this prototype model, hyper-parameters were optimized to help improve accuracy. First, a subset of the data were used to try out different combinations of the random key for the initialization of the weights and different hidden-layer sizes. The best results were achieved with hidden-layer-sizes of 100, 35, 11 and 7. The Train-Test-Split random-seed was evaluated by testing different random-seeds.

IV. Results

Because this work describes a proof-of-concept prototype, results here are described in a step-by-step fashion. The first prototype designed was a simple model with just one hidden-layer, consisting of 100 neurons and an output-layer which contained five sentiments: very positive, positive, neutral, negative, very negative. The results were of this first prototype were extremely bad, because it was very difficult to distinguish between "positive" and "extremely positive" sentiments.

For this prototype it was sufficient to differentiate between positive and negative sentiments. Thus, the number of labels was reduced: Positive and very positive labels were combined to a single label, positive. Negative examples were combined in a similar way and the neutral examples were first temporarily omitted. With these simplifications, the model achieved an accuracy rate of 83%. After further improvements to the pre-processing, neutral labels were once again reintroduced.

Table II show the results of the final model configuration. The worst F1-Score was achieved when attempting to classify posts with a neutral sentiment.

TABLE II
PERFORMANCE METRICS

Metric	Score	Metric	Score
Accuracy	73.5%	Precision	71.0%
Recall	72.2%	F1	71.6%
Sensitivity	72.2%	Specificity	74.6%

Serrano et al. recommend measuring Roc-Auc curves to improve the performance of classification models [26]. The Roc-Auc-Curve is defined as the Area Under the Receiver Operator Curve. The value for the Roc-Auc-Curve in this result was 87%.

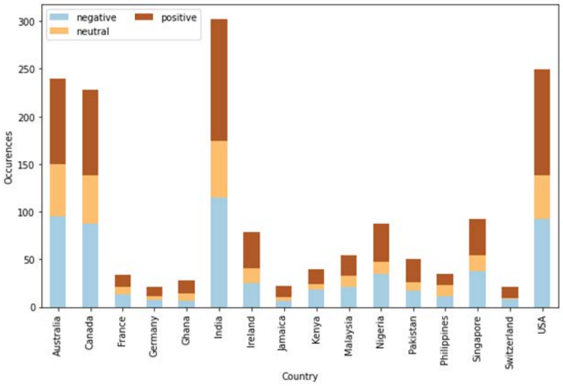

Fig. 2. Results of Sentiment Analysis by Country

Fig. 2 shows the preliminary results of the sentiment analysis for some of the individual countries investigated. Ghana has a score for individualism of 15 and would be classified as a collective country. The high percentage of positive (50%) to negative (25%) comments seems to support the hypothesis that collective countries show more positive sentiments. Singapore, with its low individualism score of 20 would be considered a collectivist society. Its percentages of positive (41%) and negative (39%)comments were almost identical. Switzerland, a country known neutrality and a high score of 68 for individualism, showed almost exclusively positive (52%) and negative (43%) sentiments, with very few neutral (5%) sentiments. Although Kenya, with an individualism score of 25 would be classified as a collective country, the percentage of negative comments (51%) is much higher than the percentage of negative comments for the U.S.A. (36%), a country with one of Hofstede's highest individualism scores (91%) [2].

The results comparing the sentiments recognized for individualist and collectivist countries is shown in Fig. 3. Against expectations, there are no statistically significant differences between the sentiment of collectivist and individualist cultures. Therefore the hypothesis that individualist cultures have a more positive sentiment about the coronavirus cannot be confirmed. Although the prototype model was able to achieve an accuracy of 73.5%, the validity of the location data and the labels remains uncertain. Another alternative explanation may lie in the binary cut-off value of 50 for Hofstede scores, which was used when assigning countries to either the individualist or the collectivist group. Some countries, such as India, with an score of 48 on individualism, were automatically be assigned to the collectivist group. An assignment based on mean or median values could possibly deliver more exact results.

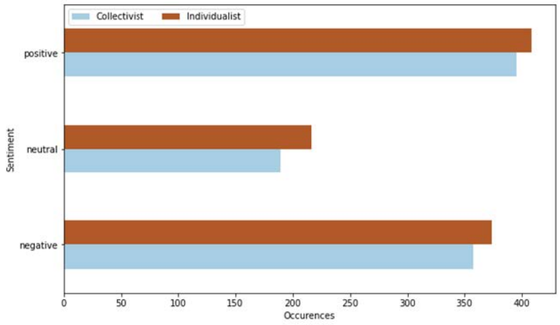

Fig. 3. Results Grouped by Cultural Dimension

A. Limitations

One drawback which may limit the validity of the results of this proof-of-concept prototype is that only a small subset of the data was used for this initial investigation. This subset is not necessarily representative. Furthermore, this data set was collected at the beginning of the pandemic, over a relatively short time period. Sentiments probably changed in different countries over the course of the pandemic. It would definitely be a good idea to investigate how sentiments in different countries changed over time.

The problem that the Twitter data is very noisy necessitated heavy pre-processing of the tweets, which may have introduced additional bias and thus skewed results. Uysal and Gunal [27] showed that the choice of pre-processing methods can have a significant effect on classification accuracy.

A further source of bias could have been introduced when assigning tweets to specific countries. Countries with more than one word, such as the "United Arab Emirates", are underrepresented in the test data set. Other countries were not included in the analysis, because Hofstede [2] did not provide a rating for 34 of the nations in the data set. It was also not possible to establish whether the location corresponds to the place where the tweet was posted or to the actual cultural background of the user who posted it. One method to improve the quality of the location data for the analysis would be to implement recognition of latitude an longitude coordinates.

A further potential problem is that the labeling process was done manually by the author of data set [15]. Errors could already have been introduced during this manual labeling. Differentiating between positive and negative sentiments is difficult, even for a human. The sentiment of a statement can be misinterpreted or experienced diversely, since it is a very subjective value. One suggestion would be to explore the use of crowd-sourcing to aid in the labelling of tweets.

V. CONCLUSION

An initial, proof-of-concept prototype to analyze tweets related to COVID-19 was described. The research questions posed at the beginning can now be answered:

- R1: Can a simple sentiment analysis model deliver meaningful insights into the opinions of the COVID-19 pandemic expressed in Twitter posts?
- R1: TRUE, the initial prototype described here was able to correctly classify the sentiments of 73.5% of tweets.
- R2: Do cultural dimensions of users from collectivist vs. individualist cultures affect their expressed opinions?
- R2: FALSE: No significant differences were found in the sentiments from collectivist vs. individualist countries.

Future work on the implementation of the complete model will include applying sentiment analysis to look at other Hofstede-Dimensions, such as power-distance, masculinity vs. femininity, uncertainty avoidance, long-term vs. short-term orientation and indulgence vs. restraint [2].

Methods to obtain better location data from each tweet will be explored. By using longitude and latitude data, it would be possible to get further information about the federal state or even the county, to analyze differences within a country. Potentially, countries grouped by other characteristics could be compared, such as the G7-States to the rest of the world. A closer look at the demographics of the users who post the tweets could be warranted. Do age, gender or other demographics affect sentiment on the COVID-19 virus?

Further work will include use of larger data sets gathered over longer time periods. The feasibility of cloud-sourcing platforms to improve the quality of data labeling will be explored. The usage of a pre-trained model such as BERT [24] could be useful to improve the recognition of different sentiments. A comparison of different methods, eXtreme Gradient Boosting (XGBoost), Random Forest or linear regression algorithms should be explored. A final idea for future work would analyze whether cultural dimensions affect the sentiment of students during the pandemic.

This work is part of a larger research project to develop hybrid courses to teach global software engineering to geographically separated groups of students. Once the full model has been implemented, investigations will focus on how teaching methods can best be adapted to students who come from different cultural backgrounds during and after the pandemic.

ACKNOWLEDGMENT

This work was supported by a grant from the German Academic Exchange (DAAD) program for International Virtual Academic Collaboration (IVAC) and with support from the Ritsumeikan University in Japan and the Nuremberg Institute of Technology in Germany.

REFERENCES

[1] W. Hussain, “Role of social media in covid-19 pandemic,” *The International Journal of Frontier Sciences*, vol. 4, no. 2, pp. 59–60, 2020.

[2] G. Hofstede, G. J. Hofstede, and M. Michael, *Cultures and Organizations: Software of the Mind*. McGraw-Hill, 2010.

[3] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” *Foundations and Trends® in information retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[4] E. T. Hall, *Beyond culture*. Anchor Books, 1989.

[5] Y. Kim, D. Sohn, and S. M. Choi, “Cultural difference in motivations for using social network sites: A comparative study of american and korean college students,” *Computers in human behavior*, vol. 27, no. 1, pp. 365–372, 2011.

[6] K. Chakraborty, S. Bhattacharyya, and R. Bag, “A survey of sentiment analysis from social media data,” *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 450–464, 2020.

[7] W. Strathern, M. Schoenfeld, R. Ghawi, and J. Pfeffer, “Against the others! detecting moral outrage in social media networks,” in *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2020, pp. 322–326.

[8] S.-F. Tsao, H. Chen, T. Tisseverasinghe, Y. Yang, L. Li, and Z. A. Butt, “What social media told us in the time of covid-19: a scoping review,” *The Lancet Digital Health*, vol. 3, no. 3, pp. e175–e194, 2021.

[9] T. Kuchler, D. Russel, and J. Stroebel, “Jue insight: The geographic spread of covid-19 correlates with the structure of social networks as measured by facebook,” *Journal of Urban Economics*, vol. 127, p. 103314, 2022.

[10] G. Barkur and G. B. K. Vibha, “Sentiment analysis of nationwide lockdown due to covid 19 outbreak: Evidence from india,” *Asian journal of psychiatry*, vol. 51, p. 102089, 2020.

[11] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, “Sentiment analysis of covid-19 tweets by deep learning classifiers—a study to show how popularity is affecting accuracy in social media,” *Applied Soft Computing*, vol. 97, p. 106754, 2020.

[12] K. Garcia and L. Berton, “Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa,” *Applied soft computing*, vol. 101, p. 107057, 2021.

[13] A. Kruspe, M. Häberle, I. Kuhn, and X. X. Zhu, “Cross-language sentiment analysis of european twitter messages duringthe covid-19 pandemic,” *arXiv preprint arXiv:2008.12172*, 2020.

[14] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, “Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets,” *Ieee Access*, vol. 8, pp. 181 074–181 090, 2020.

[15] A. Miglani, “Coronavirus tweets nlp - text classification,” Sep 2020, accessed on 18.07.2022. [Online]. Available: https://www.kaggle.com/datatattle/covid-19-nlp-text-classification

[16] Monga, “Names of countries,” Aug 2018, accessed on 18.07.2022. [Online]. Available: https://www.kaggle.com/datatattle/covid-19-nlp-text-classification

[17] Flyingcircusio, “Flyingcircusio/pycountry: A python library to access iso country, subdivision, language, currency and script definitions and their translations.” accessed on 18.07.2022. [Online]. Available: https://github.com/flyingcircusio/pycountry

[18] J. Albrecht, S. Ramachandran, and C. Winkler, *Blueprints for Text Analytics Using Python*. O’Reilly Media, 2020.

[19] D. Forsyth, *Applied Machine Learning*. Springer, 2019.

[20] J. Nothman, H. Qin, and R. Yurchak, “Stop word lists in free open-source software packages,” in *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, 2018, pp. 7–12.

[21] Y. Vasiliev, *Natural Language Processing with Python and SpaCy: A Practical Introduction*. No Starch Press, 2020.

[22] B. DeWilde, “textacy documentation,” 2021, accessed on 18.07.2022. [Online]. Available: https://textacy.readthedocs.io/en/

[23] S. Qaiser and R. Ali, “Text mining: use of tf-idf to examine the relevance of words to documents,” *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25–29, 2018.

[24] S. Ravichandiran, *Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT*. Packt Publishing Ltd, 2021.

[25] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. ” O’Reilly Media, Inc.”, 2019.

[26] A. J. Serrano, E. Soria, J. D. Martin, R. Magdalena, and J. Gomez, “Feature selection using roc curves on classification problems,” in *The 2010 international joint conference on neural networks (IJCNN)*. IEEE, 2010, pp. 1–6.

[27] A. K. Uysal and S. Gunal, “The impact of preprocessing on text classification,” *Information processing & management*, vol. 50, no. 1, pp. 104–112, 2014.

#MyIBDHistory on Twitter

Predicting IBD Type from Personal Tweets

Maya Stemmer, Gilad Ravid, Yisrael Parmet
Department of Industrial Engineering and Management
Ben-Gurion University of the Negev
P.O.B. 653, Beer Sheva, Israel
e-mails: mayast@post.bgu.ac.il, rgilad@bgu.ac.il, iparmet@bgu.ac.il

***Abstract*—Inflammatory Bowel Disease (IBD) is a chronic inflammation condition of the digestive system that is usually classified into one of two diseases: Crohn's Disease (CD) or Ulcerative Colitis (UC). If neither one is diagnosed with certainty, the patient is diagnosed with IBD Unclassified (IBD-U). IBD patients form communities on Twitter and exchange thoughts regarding their diseases. In 2018, IBD patients shared their disease history on Twitter and signed their tweets with the hashtag *#MyIBDHistory*. In their tweets, they mentioned their age at diagnosis, the medications they have tried over the years, whether they underwent any surgeries, and more. In this research, we analyzed patients' tweets containing the hashtag *#MyIBDHistory* and built a classifier that predicts the IBD type (CD or UC) of a patient. We transformed the disease history described in the tweets into tabular classification features and assessed their importance. We identified key features that helped distinguish CD from UC and used the classifier to predict the disease type of IBD-U patients. Our results were correlated with IBD related research, as the two most prominent features that tilted the classification towards CD were having a fistula and suffering from nutrient deficiency.**

Keywords-Twitter; IBD; data analysis; logistic regression.

I. Introduction

Inflammatory Bowel Disease (IBD) is a chronic inflammation condition of the digestive system characterized by flares and remission states. The two primary diseases identified with IBD, Crohn's Disease (CD) and Ulcerative Colitis (UC), are usually diagnosed in young patients (in the age range of 15-30 years) [1,2]. Distinguishing between CD and UC is not trivial as their symptoms and effects may overlap. When the disease features are inconclusive and do not enable a certain diagnosis of CD or UC, patients are diagnosed with IBD Unclassified (IBD-U) [3,4]. The incidences of IBD are rapidly increasing, and it has evolved into a global disease [5].

There are no medications or surgical procedures that can cure IBD. Treatment options can only help with symptoms, affecting each patient differently. They involve prescription drugs and lifestyle-related solutions, such as diets and therapies. Symptoms include abdominal pain, diarrhea, and fatigue; severe cases may result in hospitalization or surgical interventions [6,7]. As chronic bowel diseases, both CD and UC require day-to-day care for drug consumption and special nutrition.

Patients describe IBD as an embarrassing disease that causes immediate disruption of daily activities. They experience difficulties adjusting to the changes it entails and consider themselves different from their peers. Since IBD is identified with frequent bowel movements, people do not hasten to share their disease with others [8,9]. IBD patients attribute part of the embarrassment to a lack of public awareness. Outsiders cannot see that a person's stomach hurts or that their bowels are scarred. The disease is invisible, and others might doubt that it exists [10,11].

The embarrassment caused by IBD and the need to confide in people who undergo similar experiences help explain the creation of IBD-related communities on Twitter. IBD patients are the most common type of users who talk about IBD on Twitter [12]. They use Twitter for sharing their own experiences and for seeking social support. They exchange thoughts about symptoms and medications and recommend treatments to one another [13]. By sharing their life experiences with the disease on Twitter, patients fight disease invisibility and raise public awareness about IBD [14].

The hashtag *#MyIBDHistory* was first initiated in 2018 by a Twitter account promoting IBD-related discussion called @bottomlineibd. The account's manager is the IBD patient and advocate Rachel Sawyer, founder of The Bottom Line IBD community. Sawyer challenged her fellow IBD patients to write their own IBD medical history in a single tweet and sign it with the hashtag *#MyIBDHistory*.

This research aimed to analyze patients' tweets containing the hashtag *#MyIBDHistory* and to determine the disease type of an IBD patient based on their symptoms and treatments. We constructed a list of classification features and used LASSO logistic regression to predict whether a patient suffered from CD or UC. We identified key features and our results correlated with IBD-related research. To adhere to ethical norms and maintain user privacy, we only publish aggregated results that do not reveal the specific users. Sawyer herself gave her informed consent to be mentioned in this study.

The rest of the paper is organized as follows: in Section II we explore related research regarding health and IBD on Twitter, in Section III we describe in detail the methods used in this research, in Section IV we review the results of our research, in Section 0 we discuss the implications of the results, and in Section VI we conclude and suggest future research.

II. Related Work

Many studies have used machine learning models [15,16] and neural networks [17] for predicting whether an individual suffers from a specific illness. Nonetheless, standard logistic regression continues to hold a prominent role in performing such tasks [18]. It yields as good a performance as more complicated models while being easy to interpret [19,20]. Using a small dataset, logistic regression even shows more stable results compared to machine learning models and neural networks [21,22]. Lately, there has been a growing interest in the use of penalized regression for IBD diagnosis [23].

During the past years, text mining and social network analysis have been used to detect personal health mentions on Twitter [24,25], identify depression [26], or track the spread of the covid-19 pandemic [27]. Regarding IBD, two previous studies [28,29] have automatically identified patients with IBD among the IBD community on Twitter, but did not differentiate between CD and UC. The authors in [30] implemented a deep-learning method for extracting medical entities from social media and showed state-of-the-art results on other diseases. Nonetheless, their method failed to distinguish CD from other IBD complications because of their overlapping symptoms.

In this study, we addressed a set of users who openly declared their IBD on Twitter and tried to distinguish between CD patients and UC patients. We used a penalized logistic regression model to predict whether a patient with IBD suffered from CD or UC based on the information they shared in their tweets.

III. Methods

In this section, we describe the data we gathered for this study and the method we used to analyze the data and predict the patients' disease type.

A. Data Collection and Preparation

On September 29th, 2021, we used Twitter academic API to collect all tweets containing the hashtag *#MyIBDHistory*. We performed a full-archive search that was not limited to a specific timeframe. We excluded retweets and limited our search to tweets written in English. Two hundred six tweets, written by 140 different users, were collected. The earliest tweet containing the hashtag was published on July 26, 2018. The hashtag was intensively used until mid-August 2018, and sporadically used later. The latest tweet containing the hashtag was published on July 23, 2019, as a reminder of last year's discourse. Our study was based solely on those publicly available tweets and did not perform any clinical intervention.

The collected tweets were manually processed by the authors of this paper who are experts in statistics and social network analysis. One hundred twenty-five users were IBD patients telling their IBD story, while others were engaged spectators who did not contribute a story of their own. Patients mentioned their age at diagnosis, the medications they have tried over the years, whether they underwent any surgeries, and more. Some patients described their history in minute detail, insisting on fitting everything into several tweets; others wrote in general and focused on milestones. We were interested in transforming the heterogeneous data written by patients into fixed categorical features, so we could analyze the data using statistical algorithms.

We carefully read all patients' tweets and processed them into a tabular framework containing categorical features. We did not decide on the features in advance; we derived them from the information the patients shared in their tweets. With every new tweet we read, we added features to our framework or updated the existing ones based on the information in contained. The only feature we added, though not mentioned in the tweets, was gender.

We deduced each patient's gender by manually looking into their Twitter profile and investigating their full name and profile picture. Notice that this process can be done automatically, as Pérez-Pérez et al. showed [28]. We also investigated their user description (bio) since many users explicitly mentioned how they should be addressed (e.g., she/her) or used informative phrases (like father/husband) in their bio. The combination of full name, profile picture, and bio was enough to determine the gender of 118 patients: eighty females and thirty-eight males.

We were unable to determine the gender of seven responders. Two of them twitted from social enterprise accounts that did not reveal personal details regarding their authors. The other five accounts were no longer available on Twitter, and we did not want to specify their gender only based on their screen name. We marked the gender of these seven users as Unknown.

The first feature we derived from the tweets was the type of the disease – whether the patient was diagnosed with CD, UC, or IBD-U. Thirty-three patients did not mention their disease type in their tweets, and we searched their Twitter profiles for the information. We were able to determine the disease type of all thirty-three patients based on their previous tweets and their Twitter bio. Seventy-six patients had CD, forty-three had UC, and six had IBD-U.

The second feature we derived from the tweets was the patient's age at diagnosis. Only sixty-seven patients mentioned their age at diagnosis, and we left the feature blank for all other patients. Since the logistic regression classification model ignores all records with missing values, we had to forfeit the entire feature or drop half the records in our dataset. Since IBD patterns in childhood differ from adult-onset disease, and distinguishing CD from UC in children differs from the equivalent task in adults [31,32], we were unwilling to give up the age feature.

Based on previous literature [1,2], we considered three age groups that are meaningful to the outburst of IBD: under 15 years old (y/o), between 15 y/o and 35 y/o, and over 35 y/o. We transformed the continuous age feature into one categorical feature, indicating whether the patient belonged to one of the three age groups. Thirteen patients were under 15 y/o when diagnosed with IBD, forty-five patients were between 15 y/o and 35 y/o, and nine were over 35 y/o. The fifty-eight patients from whom we were not able to derive their age did not belong to any of the age groups.

Based on the diverse types of drugs known to treat IBD [1,33], we created six binary features to describe the patients' medical treatments: anti-inflammatory medications (meds), steroids, antibiotics, biologic meds, immune system suppressors, and other meds. We considered each med feature positive if the patients explicitly mentioned they had tried at least one medication from that specific drug class. A negative value meant that the patients either have tried the drug class but failed to mention it or explicitly mentioned they have not. Fifty-eight tried anti-inflammatory meds, seventy-seven tried steroids, fourteen tried antibiotics, seventy-six tried biologic meds, seventy-six tried immune system suppressors, and thirteen tried other types of medication.

We constructed another two binary features: whether the patient was initially diagnosed with a different disease or a different type of IBD and whether the patient had a fistula. Twenty-two patients wrote they were misdiagnosed, and seventeen wrote they suffered from a fistula. We considered the features positive for the relevant patients and negative for the rest.

We considered three categorical features: whether the patient underwent any weight changes, whether they changed their diet as a mandatory or preventive action, and whether it took a long time to confirm their diagnosis. Thirteen patients emphasized losing weight or being extremely underweight, while only one mentioned gaining weight with medications. Eleven patients experienced forced diet changes, resorting to a liquid diet or even tube feeding, and four patients mentioned their diet as a way of controlling their disease. Thirty-five patients said they had suffered for a long time before eventually being diagnosed, and three patients mentioned that their diagnosis was part of emergency surgery. We considered each of the three categorical features unavailable whenever a patient did not explicitly mention one of its values.

We extracted one ordinal variable from the text: whether the patient was ever hospitalized or even had surgery. Eighty-five patients underwent at least one surgery, fifteen patients were hospitalized but have not had surgery, and six explicitly wrote they have never been hospitalized. We considered those who did not regard hospitalization in their tweets as those who said they were never hospitalized.

We transformed each categorical/ordinal feature into a set of binary features based on the number of categories it contained. TABLE I. summarizes the features we gathered from the tweets by showing each feature and the presence of its values in our dataset. The right column of the table explains how we eventually used the features in our classification model.

B. Predicting Disease Type

We wished to predict the type of IBD based on the symptoms the patients had and the treatments they received. We tried several learning algorithms that showed consistent results and decided to focus this paper on logistic regression because of its simplicity and interpretability.

TABLE I. CLASSIFICAION FEATURES - DESCRIPTION AND VALUES

Feature Name	Description and Values	Type and Model Use
Disease type	CD: 76, UC: 43, IBD-U: 6	Binary, dependent: CD or UC IBD-U as new data
User	Unique Twitter screen name	String, unique identifier For internal use only
Gender	Females: 80, males: 38, unknowns: 7	Two binary features
Age group	Under 15: 13, 15-35: 45, over 35: 9, unknowns: 58	Three binary features
Meds: anti-inflammatory	Yes: 58, no: 67	Binary
Meds: steroids	Yes: 77, no: 48	Binary
Meds: antibiotics	Yes: 14, no: 111	Binary
Meds: biologics	Yes: 76, no: 49	Binary
Meds: immune suppressors	Yes: 76, no: 49	Binary
Meds: others	Yes: 13, no: 112	Binary
Wrong diagnosis	Yes: 21, no: 104	Binary
Fistula	Yes: 17, no: 108	Binary
Weight	Lost: 13, gained: 1, neither: 111	Two binary features
Diet	Mandatory: 11, lifestyle: 4, neither: 110	Two binary features
Pre-diagnosis (Prior to diagnosis)	Prolonged suffering: 35, Emergency surgery: 3, neither: 87	Two binary features
Hospital	Surgery: 85, hospitalized: 15, neither: 25	Two binary features

We considered the disease type as the dependent variable and the rest of the features as independent variables and used logistic regression to predict whether the patient suffered from CD (1) or UC (0). We excluded the six patients suffering from IBD-U from our dataset and considered them as unlabeled new observations. We used the seventy-six patients suffering from CD and the forty-three patients suffering from UC to train and validate our model.

We used the scilkit-learn (sklearn) package in python [34] to split our dataset into training (80%) and test (20%) sets and to build a LASSO logistic regression model. We decided to use L1 regularization since we had twenty-one explanatory variables and a relatively small dataset. The LASSO logistic regression would help eliminate unnecessary independent variables [35].

We used five-fold cross-validation on our training data to evaluate different regularization parameter values ($c \in \{0.1, 0.5, 1, 10\}$) and select the best one ($c = 1$). Then we trained a LASSO logistic regression model with the best regularization parameter on the entire training set. We applied the obtained classifier to the test set to evaluate its performance and estimated feature importance by investigating the regression coefficients.

Finally, we trained our model on all 119 records of CD and UC patients and used the obtained classifier to classify the IBD-U patients. This means, we trained our model on the entire dataset of labeled data to predict the classification of new observations.

IV. RESULTS

TABLE II. shows five classification metrics evaluating the performance of our regression model. The table shows the evaluation of the model on the test set (when trained on the training set) and the evaluation of the model when trained on the entire dataset (for predicting the class of IBD-U patients). TABLE III. shows the confusion matrices of both cases. We can see that while our model successfully identified the CD patients, it had difficulty identifying the UC patients.

TABLE II. REGRESSION MODEL EVALUATION RESULTS

Evaluation Measure	**Evaluation Data**	
	Test Set	***Entire Dataset***
Accuracy	0.75	0.7563
Precision	0.7273	0.7527
Recall	1.0	0.9211
F1	0.8421	0.8284
Area Under the Receiver Operating Characteristic Curve (AUC ROC)	0.625	0.6931

TABLE III. CONFUSION MATRICES FOR THE REGRESSION MODEL

Evaluation Data		**Predictions**	
		Predicted UC	***Predicted CD***
Test Set	***True UC***	2	6
	True CD	0	16
Entire Dataset	***Ture UC***	20	23
	True CD	6	70

Figure 1 demonstrates the importance of the features in our model by showing the regression coefficient of each feature. We can see that the strongest feature was fistula, tilting the classification in favor of the CD class. Indeed, in our dataset, none of the patients who mentioned suffering from a fistula had UC: sixteen of them had CD, and one had IBD-U. The second strongest feature was a mandatory diet change, again favoring the CD class. Out of the eleven patients who mentioned changing their diets due to nutritional deficiencies, ten had CD, and one had IBD-U. Then, there was a group of three features favoring a CD prediction with notable coefficients: Pre-diagnosis: prolonged suffering, Meds: biologics, and Hospital: surgery.

Nine features turned out to be unimportant and were omitted from the regression model: Gender: female, Age group: under 15, Age group: 15-35, Meds: antibiotics, Meds: immune suppressors, Meds: others, Weight: gain, Diet: lifestyle, and Pre-diagnosis: emergency surgery. In Figure 1, we can see that their coefficients were shrunk to zero by the LASSO algorithm. Another insignificant feature was Hospital: hospitalization, whose coefficient was close to zero (0.04). Gender: male had the second smallest absolute coefficient of 0.125.

TABLE IV. presents the features and predictions for the six IBD-U patients. Though five of them were classified as CD patients, we can see that for only three of them, the classification probability was greater than 0.6, and for only one of them, the classification was done with great confidence (probability greater than 0.99). The probabilities of the other three predictions were close to 0.5 (between 0.4 and 0.6), meaning that the classification between CD and UC was inconclusive. The results in TABLE IV. are highlighted with a gray scale based on the strength of the classification.

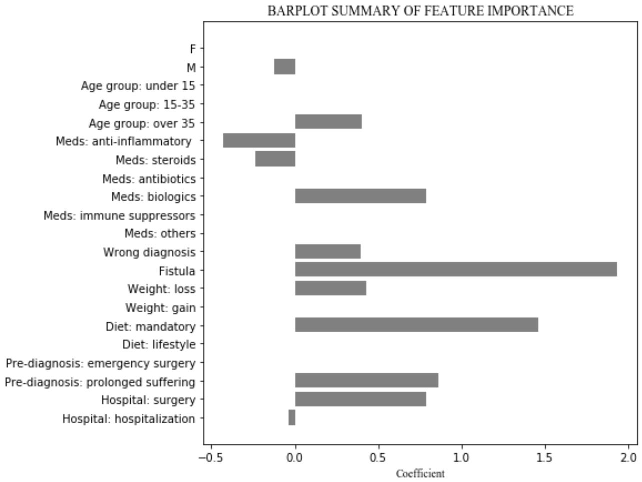

Figure 1. Barplot of feature importance based on regression coefficients.

TABLE IV. FEATURES AND PREDICTIONS FOR IBD-U PATIENTS

Feature/ Prediction	**Patient**					
	IBD1	***IBD2***	***IBD3***	***IBD4***	***IBD5***	***IBD6***
Gender: female	0	1	1	0	1	1
Gender: male	1	0	0	1	0	0
Age group: under 15	0	0	1	0	0	0
Age group: 15-35	0	0	0	1	0	0
Age group: over 35	1	0	0	0	1	0
Meds: anti-inflammatory	1	0	0	0	1	0
Meds: steroids	1	0	1	1	1	0
Meds: antibiotics	0	0	1	0	0	0
Meds: biologics	1	0	1	1	1	1
Meds: immune suppressors	1	0	1	0	0	0
Meds: others	0	0	0	0	0	0
Wrong diagnosis	0	0	0	0	0	0
Fistula	0	0	1	0	0	0
Weight: loss	0	0	1	0	0	0
Weight: gain	0	0	0	0	0	0
Diet: mandatory	0	0	1	0	0	0
Diet: lifestyle	0	0	0	0	0	0
Pre-diagnosis: emergency surgery	0	1	0	0	0	0
Pre-diagnosis: prolonged suffering	1	0	0	0	0	0
Hospital: surgery	0	1	1	1	0	0
Hospital: hospitalization	1	0	0	0	0	0
Probability	0.622	0.567	0.994	0.603	0.428	0.589
Class	1	1	1	1	0	1

The results from the feature importance analysis and the prediction for IBD-U patients correlate and demonstrate the challenge in predicting the IBD type. As can be seen in Figure 1, substantially more features favored a CD classification and the few favoring a UC classification had relatively small coefficients. The results in TABLE IV. show the influence of the features on the prediction: When the model observed one of the strong features, it confidently returned a CD prediction. When none of the strong features were present, the model returned an equivocal prediction based on weaker features.

Patient IBD3 who had a fistula and a mandatory diet change, the two strongest classification features, was unambiguously classified as a CD patient. Patients IBD1 and IBD4 both used biologic medications and showed prolonged suffering and surgery, respectively. Hence, they were both classified as CD patients, but not with the same confidence. Each of the other three patients showed only one of the medium level features, if any. Their predictions were, therefore, indecisive with probabilities close to 0.5. Patient IBD5 who used both anti-inflammatory medications and steroids, the two strongest features favoring UC, was the only one classified as a UC patient.

V. Discussion

In this section, we discuss the study's principal findings, describe its strengths and limitations, and suggest future work.

A. *Principal Findings*

In this study, we collected and analyzed tweets containing the hashtag *#MyIBDHistory*, where IBD patients described their disease history in just one tweet. We transformed the natural language text of the tweets into a tabular database with binary features that indicated the symptoms and the treatments the patients experienced. Then, we trained a LASSO logistic regression model that predicts the type of the disease, CD or UC, based on these binary features. We analyzed the importance of our classification features and used the classifier to predict the disease type of patients with IBD-U.

We did not compare the performance of our classifier with classification results from previous studies [28-30] since our prediction task differed from theirs. To the best of our knowledge, this is the first study to use social media data for differentiating between CD and UC.

The feature importance analysis and the prediction of disease type for IBD-U patients showed the complexity of distinguishing between these two diseases. In some cases, the CD's distinctive characteristics helped identify it. In other cases, the prediction probabilities were approximately 0.5, indicating the ambiguity of the classification.

The use of LASSO logistic regression helped to eliminate unnecessary independent variables that did not contribute to the classification. A regular logistic regression would give a small coefficient, but not zero. If we had used regular logistic regression, we had to perform a procedure of model selection such as forward selection or backward elimination to get exactly zero coefficients.

The two key features that helped distinguish CD from UC were having a fistula and resorting to mandatory diet changes due to nutrient deficiency. These findings align with IBD-related literature since suffering from fistulas or malnutrition is common with CD, but seldom occurs with UC [36,37]. The IBD-U patient who was classified as a CD patient with great confidence also suffered from fistula and malnutrition.

The female indicator was ignored entirely by the classifier and the male indicator had the second smallest absolute coefficient. Overall, our data contained more females than males, even though there are more male Twitter users than female Twitter users [38]. Moreover, the prevalence of CD is similar within the two genders and the prevalence of UC is geographically dependent [39]. Both facts can explain why the gender features did not contribute to the classification.

B. *Impact on Health-Related Research*

Twitter is becoming an online space for health-related conversations where patients share personal experiences on a global scale [25]. The platform is available for patients at any time, allowing them to get support from others sharing their disease. It constitutes a huge database of personal health information that can enrich traditional medical data.

Twitter research enables to collect data from substantial amounts of patients simultaneously and to perform both personal and aggregative analyses. Hence, such research may derive not only personalized insights but also global comprehensions regarding the disease.

This study demonstrates how findings from Twitter research on IBD patients correlate with existing medical knowledge regarding the disease. Predicting the type of IBD will help physicians when determining the right treatments for patients. Insights from such study can serve as a decision support system for physicians facing a challenging diagnosis. Therefore, further mining Twitter for health-related data may complement and enhance healthcare research.

C. *Limitations*

We focused our research on Twitter and manually processed tweets containing the hashtag *#MyIBDHistory*. Therefore, our patients' dataset was relatively small and contained only 125 patients. Enriching the dataset, by identifying more patients on Twitter or expanding the search to other social media, could significantly improve the classifier's performance and make the classification more precise.

Our limited data were also imbalanced: we had 76 CD patients and only 43 UC patients. Nonetheless, we used a 0.5 classification threshold such that any probability greater than 0.5 indicated a CD prediction. This could explain the bias of our model towards the CD class.

The limited dataset and its imbalance were inherent in the information available on Twitter. We did not filter the data other than excluding retweets and focusing on tweets written in English. All tweets containing the hashtag *#MyIBDHistory* that met this description were used in this study.

Finally, though both having a fistula and undergoing surgery had meaningful coefficients, the two features are correlated since surgery is usually necessary for treating fistulas [37]. Our surgery feature does not differentiate between bowel surgeries and fistulotomies.

VI. Conclusion and Future Work

In the era of personalized medicine and patient-centered care, it is important to derive insights that reflect the patients' perspective, as manifested in social media. Collecting and analyzing patients' data on Twitter shows that CD and UC are not easily distinguishable and highlights two key features that help identify CD from UC. It also points out insignificant features for separating the two. The findings provide an additional foundation for existing medical knowledge regarding IBD.

In a previous study [29], aiming to identify patients with IBD on Twitter, we trained and evaluated a classifier that distinguishes patients with IBD from other users who tweet about the disease. In future research, we intend to enrich our patients' dataset by applying the classifier and identifying more patients with IBD. Then, we can mine their Twitter timelines for the key features found in this study and enable the analysis of big data.

Future research should compare the results of such scalable analysis with those presented in this study and evaluate the contribution of collecting patients' data automatically. On the one hand, one can achieve a much larger dataset of patients. On the other hand, the dataset may contain erroneous identification of patients due to the imperfection of the classifier. It would be interesting to investigate how this trade-off affects the quality of the results.

This research suggests that there is room for collaboration between physicians and engineers regarding understanding chronic diseases. The personal information shared by chronically ill patients on Twitter can be used to understand better the disease and how it affects patients' lives. The presented methods, which were applied to IBD, can also help to explore other medical conditions. Although such analysis should not strive to replace physicians or draw conclusions of clinical nature, it may provide complementary recommendations for healthy lifestyles based on the wisdom of the crowd.

Acknowledgment

This study was supported by a grant of the ERA-Net Cofund HDHL-INTIMIC (INtesTInal MICrobiomics) under the umbrella of Joint Programming Initiative "A healthy diet for a healthy life".

References

[1] Crohn’s & Colitis Foundation of America, "The facts about inflammatory bowel diseases," Inflammatory bowel diseases, vol. 2, p. 1, 2014. [PDF].

[2] I. Trivedi and L. Keefer, "The emerging adult with inflammatory bowel disease: challenges and recommendations for the adult gastroenterologist," Gastroenterology Research and Practice, 2015.

[3] B. S. Kirschner, "Inflammatory Bowel Disease Unclassified (IBD-U)/Indeterminate Colitis," Textbook of Pediatric Gastroenterology, Hepatology and Nutrition, pp. 393-399, Springer, Cham, 2022.

[4] D. A. Winter et al., "Pediatric IBD-unclassified is less common than previously reported; results of an 8-year audit of the EUROKIDS registry," Inflammatory bowel diseases, vol. 21, no. 9, pp. 2145-2153, 2015.

[5] G. G. Kaplan, "The global burden of IBD: from 2015 to 2025," Nature reviews Gastroenterology & hepatology, vol. 12, no. 12, pp. 720-727, 2015.

[6] B. Norton, R. Thomas, K. G. Lomax, and S. Dudley-Brown, "Patient perspectives on the impact of Crohn’s disease: results from group interviews," Patient preference and adherence, vol. 6, pp. 509-520, 2012.

[7] D. T. Rubin et al., "The impact of ulcerative colitis on patients’ lives compared to other chronic diseases: a patient survey," Digestive diseases and sciences, vol. 55, no. 4, pp. 1044-1052, 2010.

[8] J. Devlen et al., "The burden of inflammatory bowel disease: a patient-reported qualitative analysis and development of a conceptual model," Inflammatory bowel diseases, vol. 20, no. 3, pp. 545-552, 2014.

[9] N. J. Hall, G. P. Rubin, A. Dougall, A. Hungin, and J. Neely., "The fight for ‘health-related normality’: a qualitative study of the experiences of individuals living with established inflammatory bowel disease (IBD)," Journal of health psychology, vol. 10, no. 3, pp. 443-455, 2005.

[10] D. O. Frohlich, "The Social Construction of Inflammatory Bowel Disease Using Social Media Technologies," Health Communication, vol. 31, no. 11, pp. 1412-1420, 2016.

[11] G. G. Macdonald et al., "Patient perspectives on the challenges and responsibilities of living with chronic inflammatory diseases: qualitative study," Journal of Participatory Medicine, vol. 10, no. 4, e10815, 2018.

[12] A. Rowe, S. Rowe, A. Silverman, and M. L. Borum, " P024 Crohn’s disease messaging on twitter: who’s talking?," Gastroenterology, vol. 154, no. 1, pp. S13-S14, 2018.

[13] P. O'Neill, B. Shandro, and A. Poullis, "Patient perspectives on social-media-delivered telemedicine for inflammatory bowel disease," Future Healthcare Journal, vol. 7, no. 3, p. 241, 2020.

[14] D. O. Frohlich and A. N. Zmyslinski-Seelig, "How Uncover Ostomy challenges ostomy stigma, and encourages others to do the same," New media & society, vol. 18, no. 2, pp. 220-238, 2016.

[15] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," PloS one, vol. 12, no. 4, e0174944, 2017.

[16] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient's health parameters," Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 8, no. 12, pp. 59-65, 2016.

[17] N. H. Ismail, N. Liu, M. Du, Z. He, and X. Hu. "A deep learning approach for identifying cancer survivors living with post-traumatic stress disorder on Twitter," BMC Medical Informatics and Decision Making, vol. 20, no. 4, pp. 1-11, 2020.

[18] M. E. Shipe, S. A. Deppen, F. Farjah, and E. L. Grogan, "Developing prediction models for clinical use using logistic regression: an overview," Journal of thoracic disease, vol. 11, no. Suppl 4, pp. S574-S584, 2019.

[19] S. Nusinovici et al., "Logistic regression was as good as machine learning for predicting major chronic diseases," Journal of clinical epidemiology, vol. 122, pp. 56-69, 2020.

[20] E. Christodoulou et al., "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," Journal of clinical epidemiology, vol. 110, pp. 12-22, 2019.

[21] T. van der Ploeg, P. C. Austin, and E. W. Steyerberg, "Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints," BMC medical research methodology, vol. 14, no. 1, pp 1-13, 2014.

[22] S. Nalluri S, R. Vijaya Saraswathi, S. Ramasubbareddy, K. Govinda, and E. Swetha, "Chronic heart disease prediction using data mining techniques," Data engineering and communication technology, pp. 903-912, Springer, 2020.

[23] B. Stankovic et al., "Machine Learning Modeling from Omics Data as Prospective Tool for Improvement of Inflammatory Bowel Disease Diagnosis and Clinical Classifications," Genes, vol 12, no. 9, p. 1438, 2021.

[24] L. Luo, Y. Wang, and D. Y. Mo, "Identifying COVID-19 Personal Health Mentions from Tweets Using Masked Attention Model," IEEE Access, 2022.

[25] Z. Yin, D. Fabbri, S. T. Rosenbloom, and B. Malin, "A scalable framework to detect personal health mentions on Twitter," Journal of medical Internet research, vol. 17, no. 6, e4305, 2015.

[26] Y. Zhang et al., "Monitoring depression trends on twitter during the COVID-19 pandemic: observational study," JMIR infodemiology, vol. 1, no. 1, e26769, 2021.

[27] M. Lopreite, P. Panzarasa, M. Puliga, and M. Riccaboni, "Early warnings of COVID-19 outbreaks across Europe from social media," Scientific reports, vol. 11, no. 1, pp. 1-7, 2021.

[28] M. Pérez-Pérez, G. Pérez-Rodríguez, F. Fdez-Riverola, and A. Lourenço, "Using twitter to understand the human bowel disease community: exploratory analysis of key topics," Journal of medical Internet research, vol. 21, no. 8, e12610, 2019.

[29] M. Stemmer, Y. Parmet, and G. Ravid, "Identifying Patients With Inflammatory Bowel Disease on Twitter and Learning From Their Personal Experience: Retrospective Cohort Study," Journal of medical Internet research, vol. 24, no. 8, e29186, 2022.

[30] S. Scepanovic, E. Martin-Lopez, D. Quercia, and K. Baykaner. "Extracting medical entities from social media," In Proceedings of the ACM Conference on Health, Inference, and Learning, pp. 170-181, 2020.

[31] A. Bousvaros et al., "Differentiating ulcerative colitis from Crohn disease in children and young adults: report of a working group of the North American Society for Pediatric Gastroenterology, Hepatology, and Nutrition and the Crohn's and Colitis Foundation of America," Journal of pediatric gastroenterology and nutrition, vol. 44, no. 5, pp. 653-674, 2007.

[32] A. S. Day, O. Ledder, S. T. Leach, and D. A. Lemberg, "Crohn’s and colitis in children and adolescents," World journal of gastroenterology: WJG, vol. 18, no. 41, p. 5862, 2012.

[33] A. B. Pithadia and S. Jain, "Treatment of inflammatory bowel disease (IBD)," Pharmacological Reports, vol. 63, no. 3, pp. 629-642, 2011.

[34] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," the Journal of machine Learning research, vol. 12, pp. 2825-2830, 2011.

[35] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267-288, 1996.

[36] J. J. Ashton, J. Gavin, and R. M. Beattie, "Exclusive enteral nutrition in Crohn's disease: Evidence and practicalities," Clinical nutrition, vol. 38, no. 1, pp. 80-89, 2019.

[37] J. Cosnes, C. Gower–Rousseau, P. Seksik, A. Cortot, "Epidemiology and natural history of inflammatory bowel diseases," Gastroenterology, vol. 140, no. 6, pp. 1785-1794, 2011.

[38] D. Noyes, "Distribution of Twitter users worldwide as of January 2021, by gender," 2021.

[39] T. Greuter, C. Manser, V. Pittet, S. R. Vavricka, and L. Biedermann, "Gender Differences in Inflammatory Bowel Disease," Digestion, vol. 101, no. 1, pp. 98-104, 2020.

A Multidimensional Analysis of YouTube Communities in the Indo-Pacific Region

Ugochukwu Onyepunuka, Thomas Marcoux, Mainuddin Shaik, Mayor Inna Gurung, Nitin Agarwal
COSMOS Research Center
University of Arkansas at Little Rock
Little Rock, AR, USA
email: {uponyepunuka, txmarcoux, mxshaik, mgurung, nxagarwal}@ualr.edu

***Abstract*—Although YouTube is the second most used social media platform in the world today, there is a need for more systematic research on video-based social network platforms, as few studies provide insights into the dynamics of their online discourse. Data, such as comments and other user engagement statistics, give insight into what information content creators are spreading and what moves communities. The analytical framework utilized in this research presents a multidimensional view into the geopolitical discourse on YouTube, with a focus on the Indo-Pacific region. We identified major YouTube channels that were discourse movers on polarizing topics, such as the treatment of Uyghur muslims, and COVID-19. We provided context on information consumption behaviors of these communities, as well as the engagement trends within these channels. This revealed segregated communities that often engaged in toxic behavior when clashing with opposite communities, as well as some inorganic engagement trends, allowing us to identify communities that may use automated means to push their narrative.**

Keywords-indo-pacific; China; Uyghur; YouTube; misinformation; social media; information operations; cyber influence campaigns.

I. INTRODUCTION

In this study, we approach how YouTube videos can and have been used as vehicles of misinformation or political propaganda. We observe content popular in the Indo-Pacific region and explore three main topics: general discourse about the Chinese government, the Uyghur crisis, and COVID-19.

Compared to the significant body of work that studies the online discourse on social media platforms like Twitter, there is little research on discourse within video-based social networking platforms, such as YouTube, which is one of the most popular platforms in the Asia-Pacific region. Carrying out systematic research on video-sharing platforms like YouTube provides immense opportunities to gain situational awareness that could be pivotal in strategic policy making, especially in trade and defense. Data, such as video titles, comments, views, and likes can give insight into what information content creators and media houses are spreading along with the audience's engagement with those contents. Engagement statistics, such as likes and views, could also serve as potential sources to gain implicit knowledge about community interests on certain topics. Analyzing trending and influential content on social media platforms requires a rigorous, systematic approach and could yield many benefits, helping stakeholders to detect existing trends and content generators, track upcoming trends and accompanying narratives, and how discourse evolves.

Through this research, we aim to present a methodological pathway on how to identify suspicious and inorganic content in cyber-influence operations. We use the word "suspicious" when a message is being amplified through inorganic means: through commenter bots, inorganic growth in a channel's subscribers, or a video's views - driving traffic to the content.

The rest of this study is structured as follows. First, we will discuss the work done by other researchers in comparable research in Section 2, describe our methodology in Section 3, including data collection, processing, and topic modeling methodology. Finally, in Section 4, we will discuss our findings before offering closing thoughts.

II. LITERATURE REVIEW

Third party web traffic reports [1] tell us that YouTube is the second most popular website, and accounts for 20.4% of all search traffic. According to official YouTube sources [2], 1 billion hours of videos are watched each day. Another study [3] found that 60% of YouTube videos are watched at least 10 times on the day they are posted. The researchers also highlighted that, if a video does not attract viewership in the first few days after upload, it will likely not gain traction later on. YouTube provides an overwhelming amount of video data, with over 500 hours of content uploaded every minute on average. That number was 300 in 2013 [4]. In previous publications [5, 6], we identified YouTube as a potential vehicle of misinformation and proposed the use of YouTube metadata for understanding and visualizing these phenomena by observing data trends. Previous research [7] has studied engagement patterns of YouTube videos and highlighted the related videos engagement trends, later designated as the "rabbit hole effect", where users will be recommended increasingly relevant videos. In some cases of polarizing content, this effect has been shown to be a contributing factor to user radicalization [8]. Recent research on the same subject leverages advanced Natural Language Processing (NLP) techniques on text entities, such as video comments [9] but we could find little work available on the video content itself.

It is important to provide some social and political context for this work. One major key idea of this study is China's desire to develop a 21st-century Maritime Silk Road coined as the belt and road initiative. One of the important hubs for the project Xinjiang, is home to Uyghurs Muslims [10]. The People's Republic of China has opened a number of what it calls "vocational education" camps, claiming to de-radicalize extremists individuals in the area. However, the West has called it cultural genocide and political brainwashing. The government of China is taking a range of countermeasures to mitigate those claims by utilizing the video-sharing platform YouTube. In these videos, content creators are calling the West hypocrites who are spreading fake news, and producing content mimicking trusted news channels like CNN [11]. While this type of content increases misinformation and gives raise to propaganda, it can have a severe impact on society. For instance, during COVID-19, 87 % of the users encountered relevant misinformation that suggested consuming ethanol or bleach as preventative measures for COVID-19 [12]. Similarly, politicizing YouTube content gave rise to anti-Asian hatred with the use of words like Chinese virus to refer to the pandemic [13]. Other research [14] suggests that this type of content is heavily concentrated among a small group that has high prior racial resentment. In contrast, YouTube content has been utilized to create public awareness as well. Studies [15] suggest that many content creators used the platform as a mean to convey information about diseases, interviewing subject matter experts, giving evidence and argument.

III. METHODOLOGY

This study uses a three-step methodology to assess YouTube content's suspiciousness. First, we collected YouTube video metadata and comments, separated them according to the main narrative they pushed, and performed engagement and network analysis on each of the subsets identified.

A. Narrative Segregation

We collected data from YouTube using a group of keywords relating to each narrative theme provided by subject matter experts from Arizona State University. The relevant keywords were identified by studying coverage in the Indo-Pacific region with further reviews to improve the inclusiveness. A data collection task was set up for each narrative using the group of keywords as a parameter to pull YouTube data relating to the narrative. To perform an analysis on a narrative, we queried our database using the keywords in the full-text search query. Table II is a snippet of the keywords for each narrative and Table I shows total collection statistics. YouTube data was collected using the official YouTube Data API and following the methodology described by Kready et al. [16].

TABLE I
VIDEO COLLECTION STATISTICS

Videos	25,673
Channels	6,806
Comments	5,538,730

B. Engagement Analysis

To uncover any suspicious activity within the channels found in our research, we examined the engagement trends over certain periods. Our analysis includes time series data based on channels' activity trends, such as, daily view count, daily subscriber count, daily video count, total views, total subscribers, total comments, and total videos; following the methodology described by Kirdemir et al. [17].

Engagement trend analysis was executed on the channels we collected data for, to discover channels exhibiting suspicious behaviors. There are 5 key steps employed by the script in discovering suspicious behaviors in a channel:

1) Rolling window correlation analysis
2) Anomaly detection
3) Rule-based classification
4) Principal component analysis
5) Clustering

The first step is the rolling window correlation analysis that groups the data into a rolling window of 100 days and computes the pairwise correlation between the video production and engagement metrics for each window. It aims to capture channels inauthentic behaviors by analyzing the correlation between engagement metrics, e.g. channels with increasing views, but decreasing subscribers. The output from this step is the start and end date for each window and the value for correlation pairs between the metrics:

- Views and subscribers
- Views and videos
- Views and comments
- Subscribers and videos
- Subscribers and comments
- Videos and comments

The output from the rolling window correlation analysis goes into the anomaly detection step to train a Long Short-Term Memory (LSTM) model on the time series data and identify anomalies in the correlation between engagement metrics. LSTM models are artificial neural networks that, unlike traditional feedforward neural networks, use feedback connections. This lets the model process not only single data points, but also entire sequences of data by using a recurrent network. To capture anomalous periods, a threshold was set to average peak of the data for each channel's correlation pairs to capture all data points that were placed above that threshold.

To compute the anomaly threshold, we grouped the data into a rolling window of 100 days, then computed the pairwise correlation between the video engagement metrics. The output from this was passed into the anomaly detection step, where the data was trained on the LSTM model. It ran through each dataset (1000 data points approximately, where one data point represents 100 days) with a batch size of 32 and a lookback size of 1. The loss from the computation was represented as the anomaly confidence score. To capture the anomalous periods, the threshold was set to the anomaly confidence score for each channel's correlation pair. This resulted in a list containing the anomalous data points for each correlation pair.

TABLE II
NARRATIVE EXTRACTION SAMPLE

S/N	Narrative	Date	Keywords
1	China	2018 - 2021	'Komunis Cina—China pengaruh Indonesia', 'Menguasai Cina—China—Tiongkok— Tionghoa ekonomi Indonesia', ...
2	Uyghur	2018 - 2021	'Uighur—Uyghur Indonesia', 'Penindasan Uighur—Uyghur bebaskan', 'Kejam Uighur—Uyghur', ...

A rule-based classification algorithm was used to determine a suspicion score ranging from 0 (least suspicious) to 1 (most suspicious) for each engagement metric pair. The suspicion scores for all indicators are then aggregated to create a single suspicion score for each observation.

C. Network Analysis

The YouTube co-commenter network represents the connections between commenters on YouTube videos, where the edge weight indicates the number of videos they commented on together. To identify the communities in the network, we used the ForceAtlas2 layout in Gephi, a force-directed layout similar to other algorithms used in network spatialization. This also showed the top node in each community, and the type of content the community engaged with. The modularity measure was used to partition the network into clusters or communities. As the degree centrality measure was used to find out the top nodes, it also shows how many ties/edges a node has. Once the node information was extracted for all nodes in a community, our dataset is queried to extract samples of the content some of the nodes engaged with.

IV. RESULTS

In this section, we discuss the thoughts of our data collection team and the ground truth as they were observed, and compare these with the results obtained through our topic modeling visualization tool. The results show YouTube co-commenters networks for the different Indo-Pacific narratives where edges between commenters indicate that the commenters were active in 10 or more of the same videos. The results highlight the top nodes and different communities in the network. It also shows the type of content that the members of the various communities engaged with or shared. We notice that the communities tend to converge towards news channels, and then segregate based on their preferred narrative.

A. Network Analysis - China Narrative (February – August 2021)

TABLE III
CHINESE NARRATIVE COMMUNITIES

S/N	Top Commenter Name	Community	Degree
1	Nilesh Bhattacharya	0 (Blue)	108
2	thndrngest	0 (Blue)	72
3	Beware of the Leaven of the USA	0 (Blue)	66
4	Discover China 探索中	1 (Purple)	173
5	True North Strong and Free	1 (Purple)	95
6	Colchicum autumn crocus	1 (Purple)	66
7	Pub Comrad	4 (Green)	113
8	Olympic - 2022	4 (Green)	76
9	Last Chang	4 (Green)	72

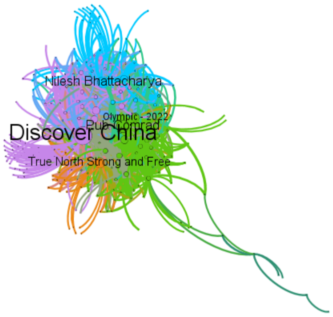

Fig. 1. China narrative (February – August 2021).

Figure 1 illustrates some of the data shown in Table III. It highlights three communities. The majority of the channels from community 0 publish videos relating to China. It is labelled as such due to the presence of high viewership videos, such as "Xi Jinping: China's president and his quest for world power", and "Global brands face backlash in China for rejecting Xinjiang cotton". The comments from the top nodes in this community showed support for China, but showed some robotic or translated patterns, e.g., "CPC is indeed the Chinese people's party. Long may it stay strong", "Long live the people's Republic of China".

Videos in community 1 discussed China centenary celebration. It is labelled as such due to the presence of high viewership videos, such as "Xi Jinping leads celebrations marking centenary of China's ruling Communist Party", and "China's largest military parade marks National Day". The comments from two of the top nodes in this community were pro-China with comments on Chinas' centenary celebration and others. While the comments from the other node were anti-China, and were critical of the decision to hold the celebration while citizens' grievances went unaddressed.

Top videos in community 3 also captured polarizing videos from other communities, which tended to communicate anti-China sentiment, e.g., "Taiwan: 'China preparing for final military assault", and "Global brands face backlash in China for rejecting Xinjiang cotton". The comments made by the top nodes in this community were vehemently anti-China, with comments such as: "Of the 14 countries bordering China, it has conflict with 13 of them including Russia", and "Without constant lying (on top of intimidation) the Chinese regime can't exist".

TABLE IV
UYGHUR NARRATIVE COMMUNITIES

S/N	Top Commenter Name	Community	Degree
1	Discover China	0 (Green)	145
2	Nilesh Bhattacharya	0 (Green)	102
3	thndrngest	0 (Green)	95
4	Beware of the Leaven of the USA	1 (Purple)	210
5	Arthur Lincoln	1 (Purple)	85
6	Aaron Baldwin	1 (Purple)	63
7	Hüseme Erbolat	2 (Blue)	67
8	John Francisco	2 (Blue)	55
9	lin hai	2 (Blue)	35
10	Yuni Sukawana	3 (Orange)	71
11	Siti Nurjannah	3 (Orange)	71
12	Camay Chayo	3 (Orange)	68

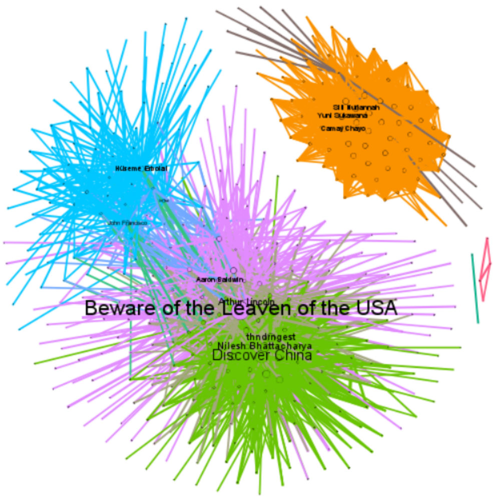

Fig. 2. Uyghur narrative (February – August 2021).

B. Network Analysis - Uyghur Narrative (February – August 2021)

Figure 2 illustrates some of the data shown in Table IV. It highlights four communities. In community 0, we noticed the top videos focused on Chinas' forced labor camps and the oppression of the Uyghurs, e.g., “Uyghurs Who Fled China Now Face Repression in Pakistan”, “What's happening with China's Uighurs? — Start Here”, and “GLOBALink — Chinese scholars debunk ”forced labor” claims with own investigation”. The comments made by top nodes in this community showed support to China and referred to China's oppression of the Uyghur Muslims as accusations. They also showed dislike for the western media.

The top videos in the nodes community 1 engaged with tended to be educational content, with titles such as “How Xinjiang Became Muslim ft. Let's Talk Religion”, and “What's China's 're-education camp' in Xinjiang really about?”. The comments from the top nodes in this community were mixed. One showed support for China and talked about how western media frames China as negative in their news, with another top node being anti-China.

Top videos in community 2 included content from popular politainment channels, such as “Last Week Tonight with John Oliver” and “VICE News”, and also featured engaging titles, such as “China's Vanishing Muslims: Undercover In The Most Dystopian Place In The World”. The comments made by the top nodes in this community were very polarized and argumentative, with both sides of the issues present.

The channels in community 3 share educational content on Islam, with top video in this community pertaining to the teachings of Islam. The title of the videos, translated to English, includes: “Extraordinary Ustadz Abdul Somad answered all the questions of this Malaysian congregation”, and “2 ways to ask for God's help — Palembang”. The comments from the top nodes were positive, and praised the speakers and their material.

C. Network Analysis - COVID-19 Narrative (All)

TABLE V
COVID-19 NARRATIVE COMMUNITIES

S/N	Commenter Name	Community	Degree
1	True North Strong and Free	0 (Orange)	16
2	Lau Billy	0 (Orange)	3
3	Frederic Chen	0 (Orange)	1
4	Nilesh Bhattacharya	1 (Green)	12
5	Jef Chen	1 (Green)	11
6	Beware of the Leaven of the USA	1 (Green)	11
7	Colchicum autumn crocus	2 (Purple)	13
8	thndrngest	2 (Purple)	8
9	LVPN 1	2 (Purple)	6

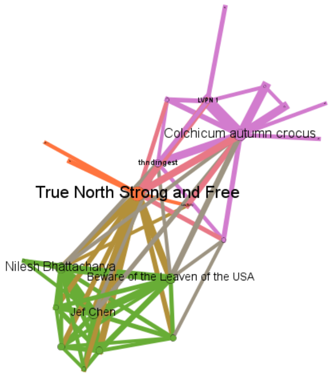

Fig. 3. COVID-19 narrative (all dates).

Figure 3 illustrates some of the data shown in Table V. It highlights three communities. In community 0, we once again see mostly News channels, including “CGTN”, “New China TV”, “球Global Times', “DW News”, and “China Daily 中日”. Most of the top videos in this community include content from the “China Global Television Network” (CGTN) channel and include “Fighting Together: China sends aid to Colombia to combat virus”, and “Ambassador Lin Songtian discusses China-Africa pandemic aid”. Most videos seemed to discuss Chinese international aid in vaccine distribution. The comments made by the top nodes in this community showed high toxicity and attacked the video content, as well as the other commenters. The top nodes showed polarized attitudes in regards to China, as well as inorganic behavior. Notable, one of the top nodes posted emojis exclusively.

The top nodes in community 1 engaged with the videos listed below, with a majority of videos making reference to China, e.g., “The Truth About The COVID Origin and the Lab Leak Theory”, and “China says it has not received COVID-19 aid from U.S. government”. We noticed obviously inorganic behavior, with some of the comments made by one of the top nodes, with repeating comments for every video commented on. The comments from the other two nodes hinted towards the commenters being pro-China: “China is the future”, “China is a good example for the world”, etc.

Videos the top nodes from community 2 engaged with only include two news channels: CGTN, and News China TV. Top videos from these channels discussed COVID-19 aid, e.g., “Medical teams return home with Wuhan in their hearts”, and “Why did Sichuan experts volunteer to help Italy?”. Comments were consistent with this pro-China sentiment.

D. Engagement Trends Analysis - Indo-Pacific Channels

Engagement trend analysis is used to uncover channel-level suspicious activity on YouTube by processing metrics of video production and engagement through a multi-step analytical pipeline including rolling window correlation analysis, anomaly detection, peak detection, rule-based classification, Principal Component Analysis (PCA), and unsupervised clustering. From the list of 5,000 Indo-Pacific channel ids collected, we were able to get daily data on 3,517 channels from Social Blade, an online tool tracking social media statistics and analytics, commonly used for traffic and earnings estimations.

Figure 4 shows the output of the anomaly detection model, which returned anomalous periods with a large feature set. PCA was then used to reduce the dimensions of the dataset and a scatterplot was created using the first two principal components and the suspicion score was used to color the data point, as reported in Table VI, which shows the highest scores. Once the suspicion score was generated for every observation of the 3,517 Indo-Pacific channels, the channel with the highest suspicion score was “Breaking News TV” with a suspicion score of 0.72. The anomalous period for this data point occurred between 2019-03-24 and 2019-10-08. The channel has now been taken down.

The final step was using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to identify clusters (groups of channels with similar engagement trends) from the PCA scatter plot - see Figure 5. Figure 5 matches Figure 4 and helps match communities to trends of suspicious engagement.

V. Conclusion and Future Work

In this study, we presented a multi-dimensional analysis of popular political and societal discourse in the Indo-Pacific region. We focused on YouTube, the second most used social media platform in the world today, and highly popular in the Indo-Pacific region, as a source of information and entertainment. To understand the patterns behind these online discourse dynamics, the interests of these communities, and the content preferences of the discourse movers (channels), we used network analysis and narrative segregation to identify various communities within YouTube networks, and provided context on their information consumption behaviors and engagement metrics. This revealed highly polarized user to user interactions, as well as some inorganic engagement trends, allowing us to identify “suspicious” communities. Through this study, we want to present the pathway on how to analyze video-based platforms, especially YouTube, to obtain situational awareness during any social cyber-influence operation. Future points of improvement for this research include considering bot accounts and their impact, and further automating the process of connecting community detection and clustering to engagement trends, creating a suspicion score at the community level instead of the channel level. This will allow analysts

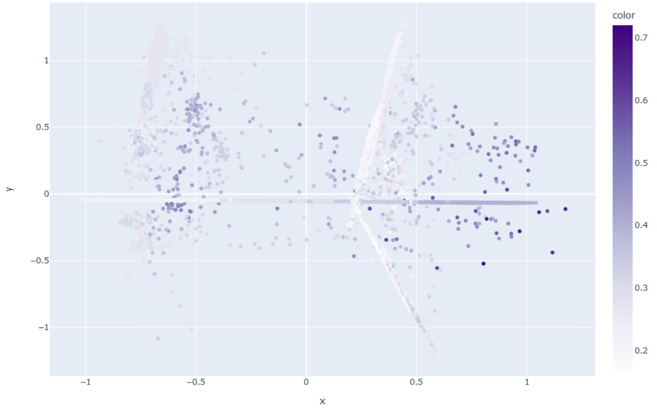

Fig. 4. Principal component analysis scatter plot showing suspicion scores.

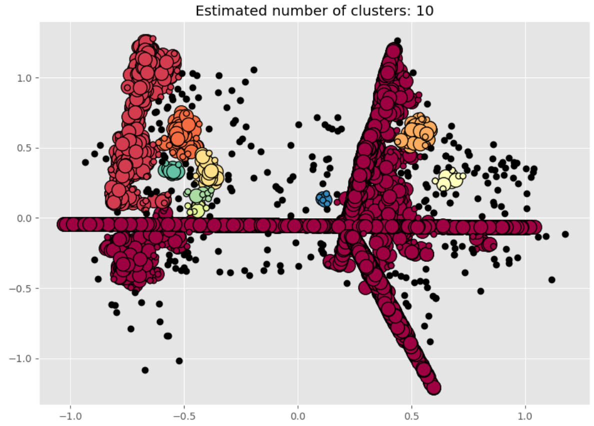

Fig. 5. Scatter plot showing clusters identified from DBSCAN.

TABLE VI
CHANNELS WITH MOST SUSPICIOUS DATA POINT

Channel ID	Channel Name	Suspicious Score	Date Range
UCN_qIhm7BAq9Qxa6j9XMVXA	Breaking News TV	0.72	2019-03-24 to 2019-10-08
UCrGZO3wJ20CWiy36Fdu6vdw	Badminton Talk	0.68	2019-04-24 to 2019-11-08
UCkQSMH1vP1Kcx-SPArUYLLQ	viral_makkodak	0.67	2020-02-07 to 2020-09-04
UCmvtaFkiWOSHhnOwt4Fz68g	SAFA News	0.64	2019-11-26 to 2020-06-22

to automatically identify communities that tend to participate in online information operations, purposefully or otherwise.

ACKNOWLEDGEMENT

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540, N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-20-1-0262, W911NF-16-1-0189), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

REFERENCES

[1] *Youtube.com Traffic Analytics and Market Share — Similarweb*, URL: https://www.similarweb.com/website/youtube.com/#overview (visited on 2022-07-15).

[2] *How YouTube Works - Product Features, Responsibility, & Impact*, URL: https://www.youtube.com/intl/en-GB/howyoutubeworks/ (visited on 2022-07-15).

[3] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, “Analyzing the Video Popularity Characteristics of Large-Scale User Generated Content Systems”, in: *IEEE/ACM Transactions on Networking* 17.5 (2009-10), pp. 1357–1370, ISSN: 1558-2566, DOI: 10.1109/TNET.2008.2011358.

[4] J. Hale, *More Than 500 Hours Of Content Are Now Being Uploaded To YouTube Every Minute*, 2019-05, URL: https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/ (visited on 2022-07-15).

[5] M. N. Hussain, S. Tokdemir, N. Agarwal, and S. Al-khateeb, “Analyzing Disinformation and Crowd Manipulation Tactics on YouTube”, in: *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM ’18, event-place: Barcelona, Spain, IEEE Press, 2018, pp. 1092–1095, ISBN: 978-1-5386-6051-5.

[6] T. Marcoux et al., “Understanding Information Operations Using YouTubeTracker”, in: *IEEE/WIC/ACM International Conference on Web Intelligence - Companion Volume*, WI ’19 Companion, Thessaloniki, Greece: Association for Computing Machinery, 2019, pp. 309–313, ISBN: 9781450369886, DOI: 10.1145/3358695.3360917, URL: https://doi.org/10.1145/3358695.3360917.

[7] X. Cheng, C. Dale, and J. Liu, “Statistics and Social Network of YouTube Videos”, in: *2008 16th Interntional Workshop on Quality of Service*, 2008, pp. 229–238, DOI: 10.1109/IWQOS.2008.32.

[8] L. Tang et al., ““Down the Rabbit Hole” of Vaccine Misinformation on YouTube: Network Exposure Study”, in: *J Med Internet Res* 23.1 (2021-01), e23262, ISSN: 1438-8871, DOI: 10.2196/23262, URL: http://www.ncbi.nlm.nih.gov/pubmed/33399543.

[9] J. C. Medina Serrano, O. Papakyriakopoulos, and S. Hegelich, “NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube”, in: *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online: Association for Computational Linguistics, 2020-07, URL: https://www.aclweb.org/anthology/2020.nlpcovid19-acl.17.

[10] M. Clarke, “The Belt and Road Initiative: Exploring Beijing’s Motivations and Challenges for its New Silk Road”, in: *Strategic Analysis* 42.2 (2018), Publisher: Routledge _eprint: https://doi.org/10.1080/09700161.2018.1439326, pp. 84–102, DOI: 10.1080/09700161.2018.1439326, URL: https://doi.org/10.1080/09700161.2018.1439326.

[11] B. Alpermann, “’In other news’: China’s international media strategy on Xinjiang –CGTN and Xinhua on YouTube”, in: 2020-10.

[12] S. Zhang, W. Pian, F. Ma, Z. Ni, and Y. Liu, “Characterizing the COVID-19 Infodemic on Chinese Social Media: Exploratory Study”, in: *JMIR Public Health Surveill* 7.2 (2021-02), e26090, ISSN: 2369-2960, DOI: 10.2196/26090, URL: http://www.ncbi.nlm.nih.gov/pubmed/33460391.

[13] Y. Yang, C. Noonark, and C. Donghwa, “Do YouTubers Hate Asians? An Analysis of YouTube Users’ Anti-Asian Hatred on Major U.S. News Channels during the COVID-19 Pandemic”, in: *Global Media Journal - German Edition* 11.1 (2021-07), URL: https://www.globalmediajournal.de/index.php/gmj/article/view/198.

[14] A. Y. Chen, B. Nyhan, J. Reifler, R. E. Robertson, and C. Wilson, *Subscriptions and external links help drive resentful users to alternative and extremist YouTube videos*, 2022, DOI: 10.48550/ARXIV.2204.10921, URL: https://arxiv.org/abs/2204.10921.

[15] F. A. Sofian, “YouTubers Creativity in Creating Public Awareness of COVID-19 in Indonesia: A YouTube Content Analysis”, in: *2020 International Conference on Information Management and Technology (ICIMTech)*, 2020, pp. 881–886, DOI: 10.1109/ICIMTech50083.2020.9211149.

[16] J. Kready, S. A. Shimray, M. N. Hussain, and N. Agarwal, “YouTube Data Collection Using Parallel Processing”, in: *2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 2020, pp. 1119–1122, DOI: 10.1109/IPDPSW50202.2020.00185.

[17] B. Kirdemir, O. Adeliyi, and N. Agarwal, “Towards Characterizing Coordinated Inauthentic Behaviors on YouTube”, en, in: *The 2nd Workshop on Reducing Online Misinformation through Credible Information Retrieval (ROMCIR 2022) held with the 44th European Conference on Information Retrieval (ECIR 2022)*, Stavanger, Norway, 2022-04.