



Network analysis on Universities' Twitter data with Community Detection and Topic Modeling

Yung-Ching Hsu
Master Thesis
MSc Applied Data Science, Utrecht University
Supervisor: dr. Dennis Nguyen
Second examiner: dr. Mirko Schaefer
8 July 2022

ABSTRACT

Social media has gained tremendous popularity along with the development of the Internet and technology. However, despite being aware of the performance advantages of integrating and adopting social media, most organizations and universities are unsure how to operate their accounts to reach their intended audiences. Social networking is one aspect of social media through which the accounts representing individuals and organizations create communities. These communities often form around shared ideas and interests. In higher education, Twitter is one of the main social media platforms adopted by institutions and academics. This study performs a novel twist on two popular techniques for studying online social networks: community detection and topic modeling to identify the communities and their topic of interest within universities' Twitter networks. The communities are discovered using the Louvain algorithm, and the topics are extracted from the tweets with Latent Dirichlet Allocation (LDA). The Twitter networks discovered in this study are collected from five accounts of the Faculty of Science at the universities in the Netherlands and encompass more than 600 accounts and 200 thousand tweets. The result shows that research-related topics are the most emphasized by the accounts in the communities. Besides, the study also presents the differences among the five universities and the accounts that are more involved in the major topics.

TABLE OF CONTENT

ABSTRACT.....	1
TABLE OF CONTENT.....	2
1. INTRODUCTION.....	4
2. LITERATURE REVIEW.....	5
2.1. Context: Twitter and Science Communities.....	5
2.2. Network Analysis	6
2.2.1. Community Detection and the Application on Twitter	6
2.2.2. Louvain Community Detection Algorithm	7
2.2.3. Follower Networks as Communities.....	7
2.3. Topic Modeling	8
2.3.1. Latent Dirichlet Allocation (LDA) on Twitter.....	8
3. DATA.....	9
3.1. Data Collection	9
3.2. Preprocessing	10
3.3. Vectorization	11
4. METHOD.....	11
4.1. Community Detection.....	12
4.2. Topic Modeling	12
4.2.1. Model Training.....	12
4.2.2. Topic Identification and Visualization	13
4.2.3. Topic Distribution	13
4.3. Model evaluation.....	13
4.3.1. Community detection	13
4.3.2. Topic Modeling.....	14
5. RESULT.....	14
5.1. Network Analysis	14

5.2. Topic Modeling	16
6. <i>CONCLUSION AND DISCUSSION</i>	20
<i>REFERENCE</i>	23
<i>APPENDIX</i>	28
APPENDIX A	28
APPENDIX B	29
APPENDIX C	31

1. INTRODUCTION

Social media platforms are virtually ubiquitous and are part of various daily activities in people's private and professional lives. In higher education, social media serve diverse purposes such as supporting teaching, marketing and sharing research findings (Reuben, 2008; Madhusudhan, 2012; Zachos et al., 2018). Owing to its popularity and the dialogic potential (Linvill et al., 2012), Twitter is one of the main social media platforms adopted by higher education institutions. Unlike corporations that mainly use social media for marketing, social media platforms are applied in academia for broader reasons, which will be discussed in this study.

Social media platforms have been utilized in many fields and attracted researchers to study them from all aspects, such as how organizations in different domains can manage their accounts more efficiently and effectively. Social networking is one important aspect of social media through which users create communities. These communities often form around shared ideas and interests. Identifying the communities and network structures help in understand the underlying relationships of individuals, which can be beneficial for tasks such as information spreading, scientific collaborations, marketing and recommendations (Bedi & Sharma, 2016). Therefore, community detection has been widely used in social network analysis in various domains (Himelboim et al., 2013; Gurini et al., 2014; Surian et al., 2016). Early research mostly concentrated on the structural characteristics of communities and omitted other crucial elements like their topical characteristics. However, the structural and topic properties of communities may interact mutually. For example, shared interests may form communities, while community structures can strengthen common interests. Therefore, some recent studies have applied community detection on data to discover different opinions on specific topics, especially on public and political issues. For example, Surian et al. (2016) collected the tweets and the users related to HPV vaccines and clustered the communities based on their opinions on the topic. Another research conducted by Ruiz et al. (2021) also utilized community detection methods on the tweets related to the topic of childhood vaccines to target the communities that have concerns about the vaccine with different promotion strategies. Although many studies have been proposed to discover the network structures and topic communities on social media, little work has been done in academics. This study enriches the network analysis literature and intends to understand and benefit the utilization and operations of social media accounts in higher education by analyzing universities' Twitter data. In addition, the methods proposed in the current study can be expanded to explore the network on other accounts and network properties.

The present study aims to discover the main community to which the selected accounts belong and their topic of interest instead of detecting the communities within specific topics. The analysis in this study starts with detecting the communities of high popularity accounts (the accounts with most followers) within the follower-following network and then further discovering their topic of interest. This is based on the idea that information spreads faster in a follower network, and the follower number directly indicates how famous a user is (Zhao et al., 2011). The followers' and followings' information are collected through Twitter API and compared to find the mutual friends with the most followers in the selected Twitter accounts' network. The networks will then be visualized by applying the Louvain algorithm to separate the accounts into different communities. A dataset containing both English and Dutch tweets collected from the accounts within the central communities will further be employed with the

Latent Dirichlet Allocation (LDA) method for topic extraction. These will give insight into the communities and their focus topics in the networks.

The remainder of this thesis is organized in the following manner. The literature review, which includes a brief description of Twitter and its use in higher education and science communities, community detection, and topic modeling, is discussed in Chapter 2. The data is presented in Chapter 3, including the data collection and preprocessing steps. The following section delves deeper into the methods utilized for community detection, topic modeling, decisions on the parameter settings, and evaluation. The results of the study are presented in Chapter 5. Finally, Chapter 6 wraps up the research and gives recommendations for future research.

2. LITERATURE REVIEW

2.1. Context: Twitter and Science Communities

Twitter, a microblogging platform launched in 2006, allows users to engage in particular conversations and communicate with other users by posting brief real-time messages, known as ‘tweets,’ in 280 characters or fewer. In each tweet, users can mention other users by adding ‘@’ in front of a username and joining in specific discussions using hashtags (i.e., typing a ‘#’ in front of words). By subscribing to a hashtag, users receive notifications when new tweets involving the hashtag are posted. In addition, other users can like, share, reply, or retweet a tweet after it has been posted, allowing them to exchange and distribute information immediately, participate in public discussions, or draw the attention of target users. With these features, it increases a tweet’s visibility. In 2021, Twitter had around 217 million daily active users worldwide (Twitter, 2021), and around 500 million tweets were sent daily (Sayce, 2019). Twitter’s popularity as a data source and the public application programming interface (API), which allows free access to vast amounts of tweets, have attracted researchers in different domains to conduct studies with Twitter data (Hunt, 2021; Paul et al., 2021; Viegas & Xavier, 2021; Singh, 2022).

Organizations and professionals in various fields widely use Twitter to communicate with their target audiences. In higher education, Twitter is one of the popular social media platforms applied by students, educators, staff, and the public to promote educational activities, distribute information and news, and respond to user inquiries (Almurayh & Alahmadi, 2022). Instructors adopt social media as a teaching technique for various reasons, including promoting student involvement, organizing for teaching, connecting to outside resources, increasing student attention to content, building communities of practice, and discovering resources (Gruzd et al., 2021). Universities use social media, especially Twitter and Facebook, as student recruitment tools (Barnes & Mattson, 2009). As the market of higher educational institutions becomes competitive (Mazzarol & Soutar, 2012), universities have drawn attention to the significance of reputation and branding. Rutter, Roper, and Lettice (2016) provide evidence that universities utilizing branding activity on social media can positively affect student recruiting performance.

In addition, Twitter has received wide acceptance among both academic and non-academic researchers. Scholars communicate to share information, deploy new theories, learn models, gain research ideas, distribute study results, solve experimental or theoretical difficulties, and get critiques and feedback (Jabr, 2011). The use of social media by scholars can “enhance the impact

and reach of scholarship” and “foster the development of more equitable, effective, efficient, and transparent scholarly and educational processes” (Veletsianos & Kimmons, 2012). Therefore, the online presence of researchers and instructors is encouraged by universities (Mewburn & Thomson, 2013). By using social media, notably Twitter, scientists can communicate and share their research findings with both specialized and general audiences, evaluate and discuss scientific work, and collaborate with other scientists (Daneshjou et al., 2021). According to studies, disseminating research through social media such as Twitter is highly effective in increasing citations and broadening its reach in a range of sectors (Wekerle et al., 2018; Zimba & Gasparyan, 2021; Mazurek et al., 2022), which also benefits the researchers’ universities as citations is a factor of university evaluation (Geuna & Martin, 2003; Mazurek et al., 2022).

Higher education institutions are forming and participating in the communities around science and education on Twitter for various purposes. To better understand how universities position themselves in online discourses, it is essential to look into the varied communities they involve and the topics that cluster within communities.

2.2. Network Analysis

2.2.1. Community Detection and the Application on Twitter

Social network structure has been broadly studied to detect communities. By researching the network structures and the communities, it yields insights into how the information spread and the opinions within various community (Amor et al., 2016; Surian et al., 2016). A community is considered a group of users who engage with each other more regularly and are more similar to each other than those outside the group (Pei et al., 2015). The research on community detection is helpful in a range of real-world applications, such as online marketing, policy-making, and recommendation systems.

This work applies the Louvain algorithm to detect the communities within certain accounts’ following networks. Khan and Niazi (2017) have divided the community detection techniques into four main categories: traditional community detection techniques, modularity optimization-based community detection techniques, overlapping community detection techniques, and dynamic community detection algorithms. Modularity optimization-based techniques are widely used for community detection. They are developed to partition the groups by optimizing the modularity, a measurement of the density of connections within and outside the communities. Positive modularity values indicate the potential of community structure, whereas negative ones indicate the opposite. Therefore, one can seek community structure by identifying the network divisions with positive and high modularity values. Due to the low computation speed of previous techniques, Newman and Girvan (2003) proposed using modularity as a fitness function to connect communities based on modularity gain. Blondel’s (2008) Louvain algorithm, a heuristic greedy algorithm, is one of the most prominent methods in the existing literature that assigns the communities based on modularity gain.

Many studies have applied community detection for network analysis on social media, but little research on community detection has been conducted in higher education institutions. This work aims to discover the network structure of the Twitter accounts managed by the Faculty of Science at five universities in the Netherlands and the topic interests within the main

communities to understand their position in the network and help benefit universities' adoption of Twitter.

2.2.2. Louvain Community Detection Algorithm

Blondel (2008) proposed the Louvain algorithm to detect communities in large networks with lower computation costs. Two phases are iterated repeatedly in Louvain's implementation to maximize the modularity. The algorithm first assumes that there is an N-node weighted network. During the first phase, the algorithm assigns each node in the network to a different community; therefore, there are N communities in the initial partition. Next, it assesses the gain modularity by removing each node i from its community and placing it in the neighboring node j 's community. The gain modularity is calculated by:

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m} \right)^2 \right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right], (1)$$

where \sum_{in} is the summed weights of the edges in the community, \sum_{tot} is the summed weights of the edges that link to the nodes in the community, k_i is the summed weights of the edges link to node i , $k_{i,in}$ is the summed weights of the edges from i to nodes in the community and m is the summed weights of all the edges in the network. The node i is then assigned to the community with positive and maximum modularity gain or remains in the same community if the gain is negative. The process iterates for every node until no further advancement can be made to improve the modularity, and the first phase is finished. The second phase starts once the first phase is complete. During the second phase, the nodes that are assigned to the same community during the first phase are grouped to form a new network. Then, the weights of edges from each node in one community to another are added to determine the sum weights of the edges between the two communities. Two phases are iterating until there is no more change and the maximum modularity is reached. With its low computation cost and high quality of performance (Hric et al., 2014), the Louvain algorithm is adopted in this work to partition the accounts in the network into different communities based on maximum modularity.

2.2.3. Follower Networks as Communities

On Twitter, an account can represent not only an individual but also an organization or a group of people with the same objectives. In this research, communities are detected through accounts' follower-following network based on its feature of homophily that accounts represent individuals or organizations tend to follow others with similar interests, background and viewpoints. Homophily can be observed in human social networks, where people with similar characteristics, such as age, ethnicity, work, educational background, social status, etc., gather. It affects people's social worlds in receiving related information, developing similar attitudes, and experiencing similar interactions (McPherson et al., 2001). When depicted as networks, nodes representing similar people or organizations are clustered together and more closely connected by edges (Newman, 2002). This phenomenon is also discovered in the follower-following network on Twitter. Kang and Lerman (2012) demonstrate that topically similar individuals are more likely to be connected through the following relationship than users who are not. To understand the underlying pattern and usage of Twitter by members of U.S. congress, Peng et al. (2016) found that members of Congress tend to follow or engage with colleagues who share similar political viewpoints, native state, chamber, and public concerns. Individual users and the accounts that represent organizations usually follow or are followed by others like them,

primarily when similarity is based on interests or viewpoints; the accounts tend to be more firmly connected to those in common and disconnected to those with different interests or opposing viewpoints. Du and Gregory (2017) also studied following networks on Twitter, and the result shows that new connections are triple or even more likely to be created within the same communities while existing edges linked to different communities are more likely to disconnect. Also, through a follower network, information flows faster and widely by passing through fewer nodes (Zhao et al., 2011). Based on the feature of homophily, by clustering the followers based on whom they follow and who is following them, this study aims to find the common interests shared in the same community. From an application standpoint, identifying the communities in the network can help the accounts create relevant content and target the accounts that may potentially increase the mutual engagement and information flow.

2.3. Topic Modeling

Social scientists have utilized topic modeling to automatically extract topics from large textual dataset and demonstrated that topic modeling can identify novel topics from texts without the influence of possibly skewed perspectives (Hopkins & King, 2010; Quinn et al., 2010, Jelveh et al., 2018). Despite the fact that topic modeling provides many benefits, it has several limitations. A significant drawback is the loss of interpretability. It is challenging to interpret topics generated by complicated algorithms since their outputs are produced based on mathematical properties, whereas interpretation depends on the objectives of the analysis, the researcher's perspectives, and domain expertise (Hagen, 2018). Besides, various decisions need to be made during the process of topic modeling and each choice may affect the result. Since the outputs are data-driven, the accuracy of the topics generated by the models are questionable.

2.3.1. Latent Dirichlet Allocation (LDA) on Twitter

Latent Dirichlet Allocation (LDA) is one of the most popular techniques in topic modeling, and extensive studies have been conducted using LDA in text mining to understand the sentiment and topics in extensive text collections. It is an unsupervised generative probabilistic method for modeling discrete data collections, such as corpora, first introduced by Blei, Ng, and Jordan in 2003. LDA uses word probabilities to represent topics. By looking at the words with the highest probabilities within a topic, people can understand what the topic is. LDA considers that each document is represented as a probability distribution across latent topics, and each topic is also represented as a probabilistic distribution over words that all documents and word distributions of topics share a common Dirichlet prior (Blei et al., 2003). To compare the performance of different topic modeling techniques, Qomariyah et al. (2019) applied both Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) on more than ten thousand Tweets posted by Surabaya citizens and evaluated their performances with the topic coherence. The results show that LDA performs better than LSA because LDA considers the relationship between Tweets in the corpus, whereas LSA does not.

Latent Dirichlet Allocation (LDA) for topic modeling on Twitter data has been widely used in previous research for several purposes. For example, Sanadras et al. (2020) collected all the tweets with the hashtag #CrisisUNAL from 2011 to 2015 and analyzed them with LDA to understand the conversations about the financial crisis at the National University of Colombia on Twitter. Another research conducted by Coelho and Figueira (2021) applied LDA to 18,000 Tweets collected from 12 top higher education institutions listed in the 2019 Center for World

University Rankings (CWUR) to understand the trend of different topics evolving with time. The results found that these institutions' strategies and topics have changed after the explosion of cases during the COVID-19 pandemic. There are many other publications on the application of LDA in various domains (McCallum et al., 2005; Linstead et al., 2008; Eidelman et al., 2012; Chen et al., 2015). However, the application of LDA in Tweets is not limited to topic extraction for target issues. It is commonly used with other techniques for a broader purpose. For example, Surian et al. (2016) examined Tweets related to HPV vaccines with topic modeling and community detection to find out where specific opinions are concentrated within communities. On a similar issue, Lyu et al. (2021) combined topic modeling with sentiment analysis on Twitter data to understand the opinions, emotions, and concerns of people worldwide on the COVID-19 Vaccine. In this study, LDA is applied with community detection to discover the topic interest within the detected community.

Research reveals that organizations are unclear of how to manage social media accounts to generate favorable outcomes, even though they are aware of the performance benefits of the adoption and integration (Hanna et al., 2011). The higher education industry is no exception (Rutter et al., 2016), with confused social media marketing and inconsistent techniques that eventually limit the possibility of building relationships with prospective audiences. Numerous papers applied network analysis and topic modeling on Twitter data in various fields (Zhao, 2013; Tremayne, 2014; Grandjean, 2016); however, fewer studies have been conducted on social media accounts managed by the Faculties in higher education institutions. This study builds on previous works and aims to explore further the social network structure and the topics of interest from the Twitter account of the Faculty of Science within five universities in the Netherlands by answering the following research questions:

RQ1: What communities do the selected accounts belong to?

RQ2: What are the topics focused by the accounts in the selected accounts' community?

RQ3: What are the differences in the communities and their interest topics among the selected accounts?

By answering these research questions, the study gives insights on the Twitter network structures and the topic of interest in higher education. In practice, the study presents the analysis on what topics of content universities can create to reach their target audiences and who should they reach to spread knowledge and information to more people through the influence of the leading accounts.

3. DATA

3.1. Data Collection

The dataset is collected from five universities in the Netherlands that have Twitter account specifically for Faculty of Science, which are Utrecht University (UU), University of Amsterdam (UVA), Vrije Universiteit Amsterdam (VU), Leiden University (LU) and Technische Universiteit Delft (TU Delft). Therefore, these accounts are chosen for their comparability. The accounts of @UUBeta, @uva_science, @VU_Science, @LeidenScienceEN, and @tnwtudelft are owned by the Faculty of Science at UU, UVA, VU, LU, and TU Delft, respectively. In order

to discover their Twitter network, the popular accounts (the accounts with the most followers) from the followers and following lists of each account are collected. The followers' and followings' id and their number of followers are collected to find the network's leading accounts. After discovering the communities, the tweets posted by the accounts in the communities are further collected to understand their topic of interest. The Tweepy¹ and advertools² libraries are used to connect to the Twitter API. Twitter API v1.1's "GET friends/ids," "GET followers/ids" and "GET users/lookup" are used to get the follower and following accounts' information, and "GET statuses/user_timeline" is used to get the tweets' information. The metadata retrieved for the tweets includes the author's information and the tweet ID, the date when the tweet was posted, the language used in the tweet, and the full text of the tweets. The most recent tweets (including the retweet tweets) posted by the accounts that belong to the same community as the five selected accounts in the network are collected until 30 June 2022. Due to the return limitation of "GET statuses/user_timeline," only the most 3200 tweets are returned. For accounts that have posted more than 3200 tweets, the Twitter API returns slightly more than 3200 but does not exceed 3250 tweets; for those with less than 3200 tweets, the API returns all the tweets. The dataset is imbalanced because some accounts have more tweets while others have fewer. In this case, if an account is more focused on a particular topic, the outcome of topic modeling may be affected when more tweets from that account are gathered, and vice versa. However, the tweets from all the accounts within the main communities are used in order to have a more comprehensive understanding of the topics in each community. Therefore, some processes, which will be explained in the methodology, are taken to solve this problem. The total tweet number collected for each selected account is shown in Table 1. Ethical data handling is ensured through anonymization (except for organizations that can be considered as public figures in a sense) and aggregation of the data.

Twitter Account	Number of tweets collected in:	
	English	Dutch
UUBeta (UU)	52879	2195
UvA_Science (UVA)	17598	19573
VU_Science (VU)	20876	10617
LeidenScienceEN (LU)	41617	17842
TNWTU Delft (TU Delft)	2580	16245

Table 1: The number of tweets collected from each account.

3.2. Preprocessing

Before applying LDA on the tweets to discover the topics, the text needs first to be preprocessed. SpaCy³ is a commonly used library for natural language processing. It supports several languages and is easily operated with its embedded Linguistic features, such as tokenization, lemmatization, and part-of-speech tagging that are used to preprocess text in this study. SpaCy returns a token-type object that stores the specified information, such as the lemmatized and tokenized word, and the part-of-speech tag (POS-tag), for each word in the text. These objects

¹ Information about Tweepy API can be found on: <https://docs.tweepy.org/en/stable/api.html>

² Information about Adverttools API can be found on: <https://adverttools.readthedocs.io/en/master/>

³ Information about SpaCy can be found on: <https://spacy.io/models>

are then used for later analyses. In this study, the text of each tweet is preprocessed with cleansing, lowercasing, lemmatizing, punctuation and stopwords removal, and tokenizing before training with LDA. The cleansing step first removes all the emojis, mention (@username), hashtag (# and the words after it), the characters 'RT' indicate a retweeted tweet, links, numbers, and the unit, and signs (+, >, <, =, |, *, ^, \$). Besides, only tweets posted in English or Dutch are included in the analysis. Then, punctuation and stopwords removal, lemmatizing, lowercasing, and tokenizing are processed with spaCy's `en_core_web_sm` for English tweets and `nl_core_news_sm` for Dutch tweets. Punctuation and stopword removal steps eliminate all the punctuation and terms that frequently appear in the text but do not provide information for the analysis, for instance, 'I,' 'we,' 'you,' and 'the.' Except for the stopwords listed by spaCy⁴, some words that frequently appear in the texts but with no meaning or cannot be distinguished among topics, such as 'dank,' 'van,' 'de,' 'een,' 'nee,' 'bron,' 'morge,' 'lol,' and the units for time, like 'minute,' 'hour,' 'day' and 'year' are manually added into the list when preprocessing English tweets. The full stop word list and the code for the analysis can be found on GitHub⁵. In addition, only the nouns are extracted from the preprocessed tweet texts to get better human interpretability by omitting the terms that do not contribute to the interpretation of topics, albeit the sentiment is lost. However, losing the sentiment does not affect the result in this study since the goal of this study is to find the topic of interest instead of the opinion in the communities. By removing all the verbs and adjectives, which may appear in several topics simultaneously, the results are more explainable. Next, lowercasing and lemmatizing steps turn all the characters into lowercase letters and all the words to their original form; for example, 'am,' 'was,' 'been,' and 'being' will all be converted to 'be.' After these steps, each word is tokenized and saved for further use. For text collected from accounts' descriptions, the preprocessing steps are the same as those for preprocessing tweets' text.

3.3. Vectorization

The tokenized texts are further vectorized to create data that computers can interpret. During the vectorization process, `gensim.corpora` dictionary is used to transform the text into a meaningful series of numbers. Two parameters, `MIN_DF` and `MAX_DF` are set to indicate the minimum and maximum frequency of a word. The values of these two parameters are usually determined varied based on different dataset and research purposes. In this study, `MIN_DF=5` and `MAX_DF 0.7` are specified to only include words appear in more than five and less than 70% of the tweets. This is to generate more focused topics by omitting the terms that appear too much or too little. Then the collection of words is converted to its bag-of-words (BOW) representations, which is required for model training, with `doc2bow` function.

4. METHOD

This study aims to discover the communities within the social network of the five selected Twitter accounts and explore the topic interest of their communities. The goal is achieved by applying the community detection technique to the popular accounts extracted from the selected accounts' follower-following network to separate them into different communities based on the modularity gain and then extracting the topics from the tweets posted by the accounts in the same community as the five chosen accounts.

⁴ The stop words list can be found on spaCy developers' GitHub page: https://github.com/explosion/spaCy/blob/master/spacy/lang/en/stop_words.py
⁵ The code for the analysis in this study can be found on: https://github.com/Hannayc/Thesis_2022/blob/main/Thesis%202022.ipynb

4.1. Community Detection

First, followers' and followings' ids from each of the selected accounts are collected through Twitter API and compared with each other to find the mutual friends (the accounts that follow and also followed by the selected accounts). The follower numbers of these mutual friends are further extracted through Twitter API to get the five accounts with the most followers, which are considered the most popular accounts in the network. All mutual friends will be included if the number of friends found is less than five. The exact process is then applied to these five accounts to find the other five most popular accounts in their network. After repeating the process three times, the accounts with highest popularity within three distances from the selected accounts are found. These accounts are represented with nodes, and their mutual following relationships are presented with undirected edges. In order to recognize the communities in the network, the Louvain algorithm is applied to separate the nodes into different groups by maximizing the modularity. Finally, the network for each selected account is plotted into graph using the python libraries Networkx and Matplotlib.

4.2. Topic Modeling

4.2.1. Model Training

In this study, the LDA is applied to extract the most dominant topics used by the accounts in different communities through analyzing the entire corpus of their recent tweets. The tweets collected are trained and analyzed separately by their language and community. The path to the LDAMallet program on the local disk, the dictionary, the BOW representations (corpus) created during vectorization, and some parameters, such as the topic number, `optimize_interval`, and iteration number, are input to train the model with LDAMallet from `gensim.models.wrappers`. Generally, a higher number of iterations leads to a better convergence, and a lower `optimize_interval` gives a better model fit; however, optimizing these values can be computationally expensive (Binkley et al., 2014). Given that the dataset is relatively small, I train the model with the following number of iterations {1000, 2000, 3000} and `optimize_interval` {5,10} and compared the results. The final LDA models are input with the following hyperparameters: 2000 iterations and 10 for `optimize_interval`.

One of the difficulties in topic modeling is determining the number of topics. The model's final performance depends on a solid separation between various clusters. Therefore, the coherence score, assessed by computing the degree of semantic similarity between high-scoring terms in the topic, is used to compare different topic numbers to decide the optimal number of topics for the corpus. This study uses the CV metric to calculate the coherence value. It works by first segmenting the data into word pairs and calculating the probabilities, a confirmation measure is then calculated to reflect how strong a word set supports the other, and finally, each confirmation measure is summed into an overall coherence score (Syed et al., 2017). Usually, the score increases when the number of topics increases. However, the increase gets smaller as the topic number gets high. In practice, too few topics could result in vast entities that combine various themes that should be separated, whereas too many topics might lead to similar entities that cannot be identified meaningfully. Therefore, this study considers the elbow of the curve when determining the optimal number of topics. The idea behind the approach is to identify a threshold beyond which an additional increase in the topic number is not worthwhile with the declining rate of the increase of coherence score. After training the models with the numbers considering

the coherence values, the numbers are adjusted again based on human interpretability. The final chosen topic numbers are presented in Table 2.

Twitter Account	English tweets	Dutch tweets
UUBeta (UU)	4	3
uva_science (UVA)	10	4
VU_Science(VU)	7	3
LeidenScienceEN (LU)	4	5
tnwtudelft (TU Delft)	2	4

Table 2. The chosen topic number used for each model.

4.2.2. Topic Identification and Visualization

After model training with the data, the ten words with the highest probability and the top eighty words within each topic are extracted using `lda.show()` and plotted in word clouds, respectively. The word cloud presents popular words in different sizes based on their frequency. The higher the probability of a word, the more distinctive it is for that particular topic. Each topic is then manually labeled with a subject by considering the terms that represent it. The subjects are examined by comparing the labeling results from two people to see whether the interpretations are appropriate and validated by checking a random number of tweets in the topics.

4.2.3. Topic Distribution

In this study, topic distribution is employed to identify the most popular topic within various topics in the community. The topic probability distribution for each tweet is first calculated. A probability of 1 indicates the maximal probability of the topic occurring in the tweet, and a 0 indicates the opposite. Each tweet consists of several topics, each of which is principally made up of several main terms. Next, the summed probability of each topic per account is calculated by grouping the probability based on accounts' names. However, the number of tweets collected per account is imbalanced, which can lead to a biased result, as mentioned in Chapter 3.1. Therefore, the summed probability per topic is divided by the number of tweets collected from each account. By doing this, the probability distribution for each account becomes comparable. Finally, the probability for each topic is summed up to find the most focused topic in the community.

4.3. Model evaluation

4.3.1. Community detection

Several studies have used statistical techniques to evaluate the effectiveness of community detection algorithms (Orman et al., 2012; Yin et al., 2015; Fortunato & Hric, 2016). The most common measures include Precision, Recall, F-Measure, and Modularity. Precision is the ratio of the number of correctly identified communities to the total number of detected communities. Recall is a similar method that measures the proportion of the number of correctly identified communities to known communities. The values for both measurements range from 0 to 1, where 1 is the best and 0 is the worst. However, a maximum value of Precision will result from treating each node as a separate community, and a single community composed of all nodes will get the highest Recall value (Linhares et al., 2020). Therefore, both methods are unsuitable for

evaluating community detection performance. F-Measure, also ranges from 0 to 1, strikes a balance between Precision and Recall. It is calculated by taking the harmonic mean of the Precision and Recall measurements. The value is near one when the detected communities match the known communities. However, known communities are necessary for the evaluation, whereas Modularity, which is explained in Chapter 2.2.1, can be calculated for any network (Linhares et al., 2020). A network with high modularity value indicates more distinct and less interconnected groups. This work evaluates the partition of the Louvain algorithm by computing the Modularity of each network.

4.3.2. Topic Modeling

The result of topic modeling can be evaluated using several methods, including human judgment and quantitative approaches. The procedure of some human judgment approaches, for example, manually checking the words within each topic, can be time-consuming, and the human interpretability varies between persons depending on the use and domain knowledge.

Quantitative approaches, such as perplexity and coherence value, on the contrary, are more automated and standardized. Although perplexity has been used in many cases, Chang et al. (2009) discovered the negative correlation between perplexity and human interpretability, indicating that the higher the perplexity score, the lower the human interpretability in the topics. As a result, coherence is established to capture the context between words. This study combined both CV coherence value and human judgment to assess how similar the words within a topic are to terms within other topics produced by the model.

5. RESULT

The methods and research process have been described in the previous sections; this section will present the analysis results of the study. In Chapter 5.1, the evaluation of Louvain algorithm and the Twitter network of each selected account will be displayed. The findings of topic modeling for each discovered community mentioned in Chapter 5.1 will then be shown in Chapter 5.2.

5.1. Network Analysis

Louvain algorithm is applied to detect the communities in the five selected accounts' Twitter follower-following networks. The algorithm identified between 13 to 16 communities with sizes from 6 to 22 accounts in each network. The networks of the five Twitter accounts of the Faculty of Science in UU, UVA, VU, LU, and TU Delft are shown in Figure 1, and the total numbers of nodes, edges and communities in the networks are shown in Table 3. The chosen five accounts are highlighted with the black dotted line circle in the center of the networks (see Figure 1). Each node in the network represents a Twitter account, and each edge represents the mutual following relationship between each pair of accounts. In the networks, nodes with the same colors belong to the same communities. Through the networks, the most popular accounts in each community and the various communities formed by these accounts are discovered. The quality of the partition of the communities in each network is measured by computing their modularity (see Table 4). The network of uva_science has the highest modularity with 0.8083, whereas tnwtudelft's network has the lowest modularity with 0.7211. However, all networks show high modularity, which indicates good partitions of the communities.

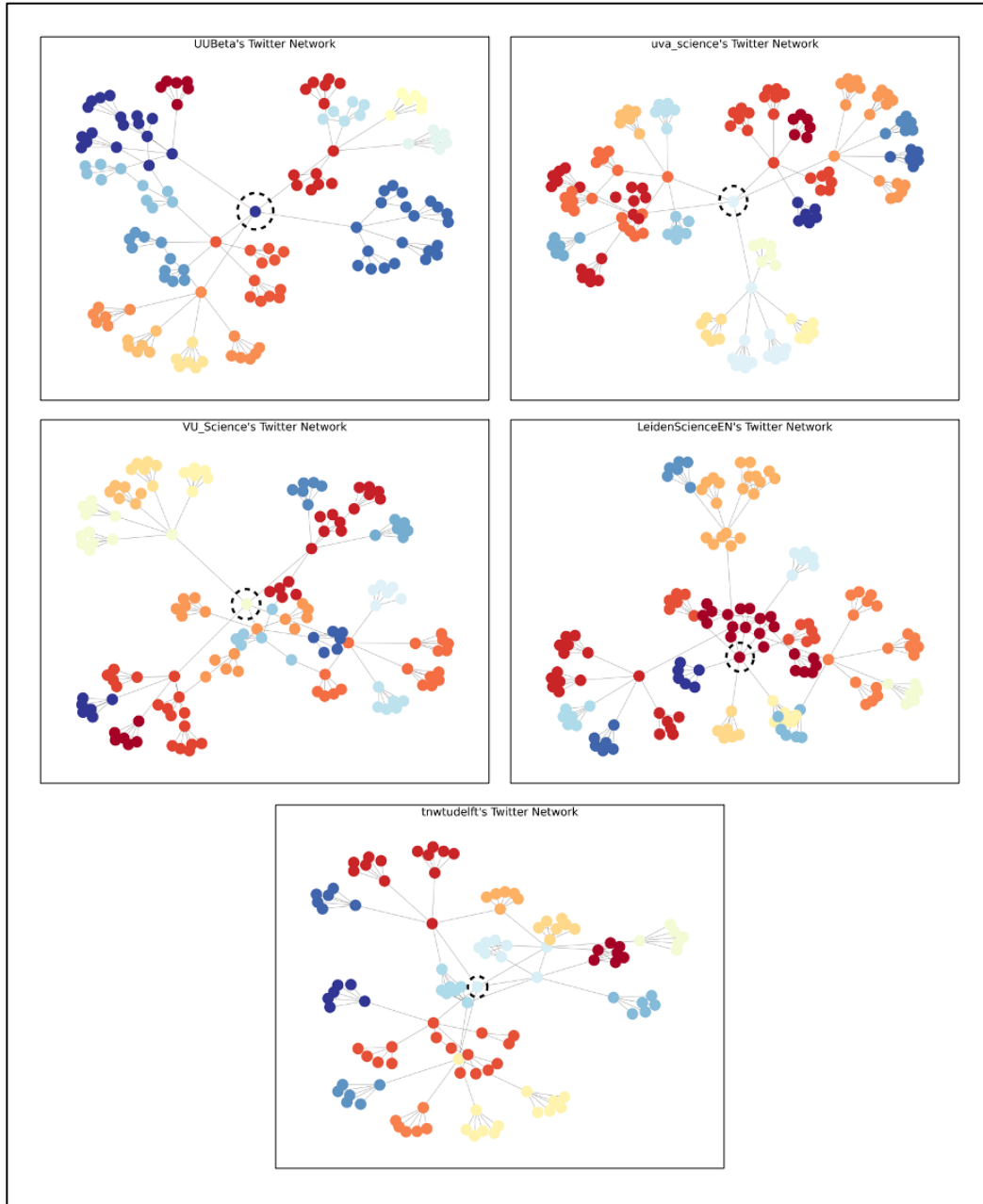


Figure 1: Twitter's networks of Faculty of Science at UU, UVA, VU, LU and TU Delft.

Twitter Account	Number of nodes in the network	Number of edges in the network	Number of communities in the network	Number of nodes in the account's community
UUBeta (UU)	134	148	13	18
uva_science (UVA)	155	155	16	14
VU_Science (VU)	148	154	16	13
LeidenScienceEN (LU)	145	155	14	22
tnwtudelft (TU Delft)	114	145	14	9

Table 3. The number of nodes, edges and communities in the networks and the user numbers in the selected accounts' communities.

Network	UUBeta (UU)	uva_science (UVA)	VU_Science (VU)	LeidenScienceEN (LU)	tnwtudelft (TU Delft)
Modularity	0.7369	0.8083	0.7699	0.7582	0.7211

Table 4. The Modularity calculated for each network (the values are rounded to the fourth decimal place).

Among all the communities, the users in the five selected accounts' communities in the networks are extracted for the analysis. This is to understand the topic interests in the community and how these selected accounts position themselves in the network. In order to understand the interests and fields of the accounts in the same community, accounts' descriptions on their Twitter profiles are collected, preprocessed and trained with one topic in LDA. The word clouds with the most frequent words in the accounts' descriptions are displayed in Appendix A. The result shows that the accounts place the most emphasis on research and the scientific disciplines are the majority in these groups.

5.2. Topic Modeling

The tweets posted by the users in the five accounts' communities are collected and trained with LDA to find the topic of interest within each community. Based on the coherence scores show in Appendix B, the numbers corresponding to the elbow of the curves are used to train the models. Although these numbers are retrieved by using an iterative method to compute coherence values for different topic numbers, it is possible that the selected numbers are not the best. Therefore, the final input topic numbers are decided based on both coherence scores and human judgment to get a more explainable result. The CV coherence scores computed for the final chosen models are shown in Table 5. The values for the models trained on Dutch tweets are sufficient, whereas those for the models trained on English tweets are comparatively low. However, a high coherence value does not guarantee high human interpretability; the method of human judgment can also lead to biased results based on different domain knowledge, and personal perspectives.

Twitter Account	Coherence values for:	
	Models train on English tweets	Models train on Dutch tweets
UUBeta (UU)	0.3237	0.6186
uva_science (UVA)	0.3866	0.7145
VU_Science(VU)	0.3589	0.6829
LeidenScienceEN (LU)	0.2802	0.6984
tnwtudelft (TU Delft)	0.4159	0.5110

Table 5: The coherence values for different models.

After training the model on the corpus preprocessed from the tweets, the top ten words in each topic are manually given a label to produce recognizable subjects (see Appendix C). Some topics contain distinct and precise words that are logically connected to each other, while some contain words that are less discriminative and irrelevant to the other words in the topic. To demonstrate the topics focused within each community, the distribution for each topic is summed and compared to find the most used topic in English and Dutch tweets (see Appendix C).

The word clouds and the distributions for the top five accounts of the most focused topics in UUBeta's community are showed in Figure 2. Except for the five universities' accounts for Faculty of Science, all the accounts' names in this study are anonymized and represented in characters for privacy reasons. In UUBeta's community, the topics of Social Issue, Sport Media, Student Life and Politics are found in English tweets. Politics related issues are most discussed by the accounts with words, such as 'law,' 'government,' 'party,' 'country,' 'mp' (Members of Parliament). Besides, Student Life topic also captures a lot of tweets from individuals sharing information about universities' programs, news and events. In Dutch tweets, the topics are most focusing on education and research field that words related to educators and research project in different domains are widely mentioned. Besides, a relatively small part of the tweets is related to the discussion and concerns about environmental issues.

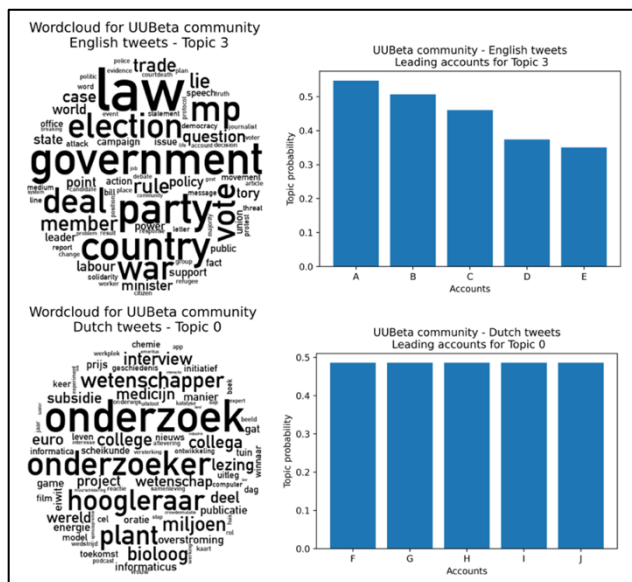


Figure 2: The word clouds and the distributions for the top five accounts of the main topics in UUBeta's community.

In uva_science's community, English tweets are classified into ten topics including, Music Concert, Environmental study, War Criminal, Covid-19, Job, Pandemic Life, Political Campaign, Science, Biking, and one that is more general with no specific topic. Among all topics, the topic of Music Concert appears the most in the tweets. The words, such as 'album,' 'gig,' 'song,' 'ticket,' 'band' and 'music' are frequently used by accounts to share about the music performances they attended. However, the ten subjects are more evenly discussed in the tweets. In Dutch tweets, Social Discussion topic is more commonly focused by users to talk about social problems and policies than the other topics in all tweets. Figure 3 displays the word clouds and distributions for the top five accounts of the most popular topics in the uva_science's community. It is found that the topic distribution for the top five accounts in the most used topics are more evenly distributed and uva_science is the second among the top five accounts that focus in Social Discussion topic.

The word clouds and the distributions for the top five accounts of the most focused topics in LeidenScienceEN's community are showed in Figure 5. In LeidenScienceEN's community, Education is the most highlighted subject among all the topics in both English and Dutch tweets. LeidenScienceEN is also one of the top accounts that highly emphasize in this topic. In English tweets, the discussion of the topic is narrower as the words are more related to doctoral degree, such as 'research,' 'student,' 'phd,' and 'paper.' Compared to English tweets, the subject is discussed in a broader perspective using terms like student, research, professor, university, book, education and child, in Dutch tweets. Other topics, for example, Disease Study, Book/Author, Neuroscience, Event and Information are also shared in the community.

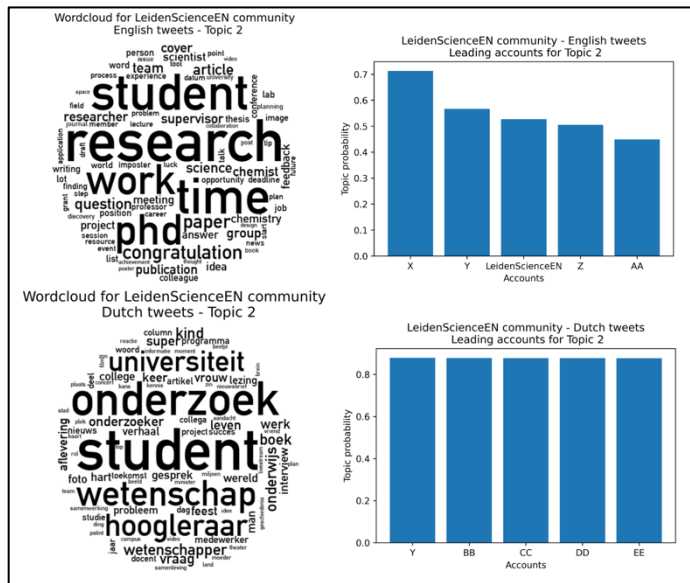


Figure 5: The word clouds and the distributions for the top five accounts of the main topics in LeidenScienceEN's community.

In tnwtudelft's community, there are much fewer accounts and English tweets collected through the processes compare to all the others. This can be the reason that tnwtudelft has only posted tweets in Dutch, therefore, attracts less English-speaking users in the community. However, regardless their speaking language, the accounts within this community are paying more attention on topic related to research, especially on the sponsorship of studies. Words, including 'grant,' 'funding' and 'miljoen' are frequently mentioned with 'research,' 'onderzoek,' 'researcher' and 'project.' Figure 6 displays the word clouds and distributions for the top five accounts of the most emphasized subjects in tnwtudelft's community. Out of all the accounts that tweet in English, tnwtudelft's account focuses the most on research topics.

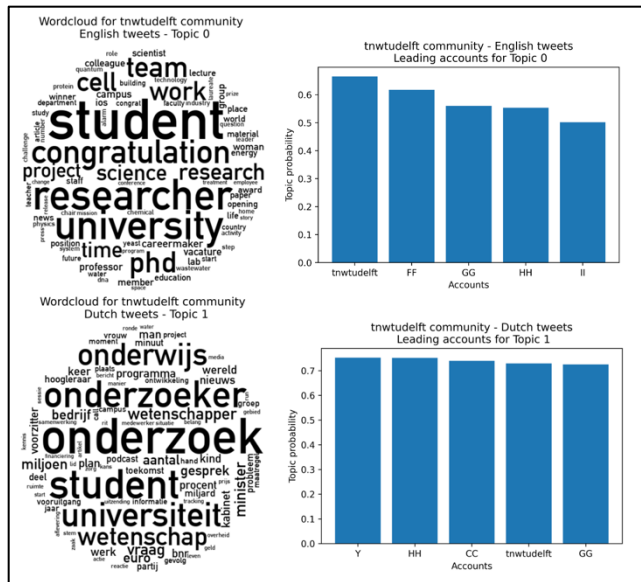


Figure 6: The word clouds and the distributions for the top five accounts of the main topics in twnwtudelft's community.

6. CONCLUSION AND DISCUSSION

This network analysis study has three primary objectives: (a) to discover the main community in each selected Twitter account (@UUBeta, @uva_science, @VU_Science, @LeidenScienceEN, and @twnwtudelft), (b) to identify the topic interests within each discovered main community, and (c) to compare the differences of topic interests among all the selected accounts. To answer RQ1: 'What communities do the selected accounts belong to?', this study collects 696 users from the five selected Twitter accounts and applied Louvain algorithm to partition the accounts into different communities in the five networks, which are shown and discussed in Chapter 5. By extracting the descriptions on the profiles from the 76 accounts in the five chosen accounts' communities, the result indicates that Science is the main domain of most users in the communities. To answer RQ2: 'What are the topics focused by the users in the selected accounts' community?', 202,022 tweets from the accounts within 5 communities in 5 networks are analyzed using LDA to discover their topic of interests. From the most used topic in each community, it is discovered that the users are mainly center on the topics in academia, such as university, education, study program, and especially on research related topics in these communities. To answer RQ3: 'What are the differences in the communities and their interest topics among the selected accounts?', the results of community detection and topic modeling of the five accounts are compared to find the difference in the communities and the topic of interests. In UUBeta's and LeidenScienceEN's communities, there are much more English tweets than Dutch tweets posted by the users, especially for the users in UUBeta's community. This shows that they are connecting more to English-speaking users. However, it is the opposite in twnwtudelft's community. From the result of topic modeling, it is found that education, research and science topics are widely posted in all of the five communities, especially for the community of LeidenScienceEN and twnwtudelft, these are the main topics share by the accounts. Nevertheless, broader topics besides academic subject, such as politics, music and nature are also highly discussed in UUBeta's, uva_science's and VU_Science's community.

For ethical concerns, the data collected in this study is available to the public that anyone can access freely through Twitter API or view it on Twitter platform. However, it is possible to find more personal information from users' ids, names and other attributes. Considering users' privacy, this study does not show any information of the users and hides the users' account id when plotting the networks to keep them anonymous.

This research contributes in understand the communities that the universities position themselves and their purpose of using social media. In practice, this analysis is beneficial for higher education institutions to understand who are the most popular users and what are the communities in their social media network, and their topic of interests. For example, this study discovers that only UUBeta is not within the top five accounts that focus on the most popular subjects in the community among the five target accounts. To draw more attention from the community, UUBeta may design contents that has include the major topics in their tweets. By recognizing the communities and their interest topics in the network, universities are able to create relevant content to reach their target audiences or spread information to more people through the followers of the leading accounts in the network by using the functions of mention and hashtag.

There are several limitations on this study. The first limitation is that the dataset is collected from Twitter, unlike news or literatures, many users are communicating in pseudo-language using abbreviations and informal words, such as 'thx,' 'LOL' and 'LMAO.' In addition, some terms are used as a metaphor or with other words as a phrase; some are used to refer to other situations or objects than their original meaning. Moreover, the analysis for Dutch words performed in this study is translated using Google Translate, some words may not be accurately expressed. Given that the terms are representing diverse definitions and the technical limitation with the language, these could lead to some topic being misinterpreted. The second limitation is that only the English and Dutch tweets are collected and analyzed. Among all the Dutch tweets, many are posted with some English words, which are removed or turned into different words in Dutch during the preprocessing steps, within the text. These may result in losing some topics from the removing or wrongly identified words, and the tweets posted by non-English and non-Dutch speaking users in the communities. Last, the effectiveness of topic modeling heavily depends on data's quality and the determined settings of parameters, especially the number of topics and iterations. Data's quality is depending on the steps of data collection, cleaning, and preprocessing. Dimensionality reduction might negatively impact the data's quality by deleting some information from the dataset prior to analysis during the data preparation stage. For the parameters setting, different topic numbers provided in the model can produce entirely different results. When the number is too small, one topic can turn into numerous topics, and when the number is too large, several topics can merge into one. This may potentially result in a distorted systemic outcome.

Some changes and additions can be made to this study for future research. First, the partition for the discovered communities in this study is based on the following relationship of the accounts that have most followers in the network. Further research may apply community detection techniques considering other characteristics, such as gender, location, interest, etc., or combining various features to separate accounts with similar trait into different groups. Second, this study collects the Twitter data from the Faculty of Science at several universities in the Netherlands.

Future studies may collect data from other Faculties or universities in different countries and compare the difference. For example, comparing the topic interests of Western society and Asian society. This can help in understand the distinct subjects and network structures under different culture background. Third, this study only focuses on the selected accounts' community, further research can expand the study by analyzing all the communities within the network to see what topics are discussed by other accounts. In sum, this research opens up several new avenues for future research on the network of social media in higher education that can be explored to understand how universities are engaging in different communities and give suggestions on how they can reach their target audiences effectively and efficiently on social media.

REFERENCE

- Almurayh, A., & Alahmadi, A. (2022). The Proliferation of Twitter Accounts in a Higher Education Institution. *IAENG International Journal of Computer Science*, 49(1).
- Amor, B. R., Vuik, S. I., Callahan, R., Darzi, A., Yaliraki, S. N., & Barahona, M. (2016). Community detection and role identification in directed networks: understanding the Twitter network of the care. data debate. *In Dynamic networks and cyber-security* (pp. 111-136).
- Barnes, N. G. and Mattson, E., 2009 N.G. Barnes, E. Mattson Social media and college admissions: The first longitudinal study Center For Marketing Research (2009)
- Bedi, P., & Sharma, C. (2016). Community detection in social networks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 6(3), 115-135.
- Binkley, D., Heinz, D., Lawrie, D., & Overfelt, J. (2014, June). Understanding LDA in source code analysis. *In Proceedings of the 22nd international conference on program comprehension* (pp. 26-36).
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Jmlr.Org. Retrieved from <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=https://githubhelp.com>
- Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10), P10008.
- Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22.
- Chen, S.-H., Santoso, A., Lee, Y.-S., & Wang, J.-C. (2015). Latent dirichlet allocation based blog analysis for criminal intention detection system. 2015 *International Carnahan Conference on Security Technology (ICCST)*.
- Coelho, T., & Figueira, A. (2021). Covid-19 impact on higher education institution's social media content strategy. *In Computational Science and Its Applications – ICCSA 2021* (pp. 657–665). Springer International Publishing.
- Daneshjou, R., Shmuylovich, L., Grada, A., & Horsley, V. (2021). Research techniques made simple: Scientific communication using Twitter. *The Journal of Investigative Dermatology*, 141(7), 1615-1621.e1. <https://doi.org/10.1016/j.jid.2021.03.026>
- Du, S., & Gregory, S. (2017). The Echo Chamber Effect in Twitter: does community polarization increase? *In Studies in Computational Intelligence* (pp. 373–378). Springer International Publishing.
- Eidelman, V., Boyd-Graber, J., & Resnik, P. (2012, July). Topic models for dynamic translation model adaptation. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers) (pp. 115-119).
- Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. *Physics reports*, 659, 1-44.

- Geuna, A., & Martin, B. R. (2003). University research evaluation and funding: An international comparison. *Minerva*, 41(4), 277–304. <https://doi.org/10.1023/b:mine.00000005155.70870.bd>
- Grandjean, M. (2016). A social network analysis of Twitter: Mapping the digital humanities community. *Cogent Arts & Humanities*, 3(1), 1171458. <https://doi.org/10.1080/23311983.2016.1171458>
- Gruzd, A., Haythornthwaite, C., Paulin, D., Gilbert, S., & Esteve del Valle, M. (2021). *Uses and gratifications factors for social media use in teaching: Instructors' perspectives*. <https://doi.org/10.32920/14639559.v1>
- Gurini, D. F., Gasparetti, F., Micarelli, A., & Sansonetti, G. (2014). ISCUR: Interest and sentiment-based community detection for user recommendation on twitter. In *User Modeling, Adaptation, and Personalization* (pp. 314–319). Springer International Publishing.
- Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models?. *Information Processing & Management*, 54(6), 1292-1307.
- Hanna, R., Rohm, A., & Crittenden, V. L. (2011). We're all connected: The power of the social media ecosystem. *Business horizons*, 54(3), 265-273.
- Himmelboim, I., Smith, M., & Shneiderman, B. (2013). Tweeting apart: Applying network analysis to detect selective exposure clusters in twitter. *Communication Methods and Measures*, 7(3–4), 195–223. <https://doi.org/10.1080/19312458.2013.813922>
- Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1), 229-247.
- Hric, D., Darst, R. K., & Fortunato, S. (2014). Community detection in networks: Structural communities versus ground truth. *Physical Review E*, 90(6), 062805.
- Hunt, D. (2021). How food companies use social media to influence policy debates: a framework of Australian ultra-processed food industry Twitter data. *Public Health Nutrition*, 24(10), 3124–3135. <https://doi.org/10.1017/S1368980020003353>
- Jabr, N. H. (2011). Social networking as a tool for extending academic learning and communication. *International Journal of Business and Social Science*, 2(12).
- Jelveh, Z., Kogut, B., & Naidu, S. (2018). Political language in economics. *Columbia Business School Research Paper*, (14-57).
- Linhares, C. D., Ponciano, J. R., Pereira, F. S., Rocha, L. E., Paiva, J. G. S., & Travençolo, B. A. (2020). Visual analysis for evaluation of community detection algorithms. *Multimedia Tools and Applications*, 79(25), 17645-17667.
- Linstead, E., Lopes, C., & Baldi, P. (2008). An application of latent Dirichlet allocation to analyzing software evolution. *2008 Seventh International Conference on Machine Learning and Applications*.
- Linvill, D. L., McGee, S. E., & Hicks, L. K. (2012). Colleges' and universities' use of Twitter: A content analysis. *Public relations review*, 38(4), 636-638.

- Lyu, J. C., Han, E. L., & Luli, G. K. (2021). COVID-19 vaccine-related discussion on Twitter: Topic modeling and sentiment analysis. *Journal of Medical Internet Research*, 23(6), e24435. <https://doi.org/10.2196/24435>
- Kang, J. H., & Lerman, K. (2012, July). Using lists to measure homophily on twitter. In *Workshops at the twenty-sixth AAAI conference on artificial intelligence*.
- Khan, B. S., & Niazi, M. A. (2017). Network community detection: A review and visual survey. In arXiv [cs.SI]. <http://arxiv.org/abs/1708.00977>
- Madhusudhan, M. (2012). Use of social networking sites by research scholars of the University of Delhi: A study. *The International Information & Library Review*, 44(2), 100–113. <https://doi.org/10.1080/10572317.2012.10762919>
- Mazurek, G., Gorska, A., Korzynski, P., & Silva, S. (2022). *Social networking sites and researcher's success*. *Journal of Computer Information Systems*, 62(2), 259–266. <https://doi.org/10.1080/08874417.2020.1783724>
- Mazzarol, T., & Soutar, G. N. (2012). Revisiting the global market for higher education. *Asia Pacific Journal of Marketing and Logistics*, 24(5), 717–737. <https://doi.org/10.1108/13555851211278079>
- McCallum, A., Corrada-Emmanuel, A., & Wang, X. (2005). Topic and role discovery in social networks.
- McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1), 415–444. <https://doi.org/10.1146/annurev.soc.27.1.415>
- Mewburn, I., & Thomson, P. (2013). Why do academics blog? An analysis of audiences, purposes and challenges. *Studies in Higher Education*, 38(8), 1105–1119. <https://doi.org/10.1080/03075079.2013.835624>
- Newman, M. E. J. (2002). Mixing patterns in networks. In arXiv [cond-mat.stat-mech]. <http://arxiv.org/abs/cond-mat/0209450>
- Newman, M. E. J., & Girvan, M. (2003). Finding and evaluating community structure in networks. In arXiv [cond-mat.stat-mech]. <http://arxiv.org/abs/cond-mat/0308217>
- Orman, G. K., Labatut, V., & Cherifi, H. (2012). Comparative evaluation of community detection algorithms: a topological approach. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08), P08001.
- Paul, J., Parameswar, N., Sindhani, M., & Dhir, S. (2021). Use of microblogging platform for digital communication in politics. *Journal of Business Research*, 127, 322–331. <https://doi.org/10.1016/j.jbusres.2021.01.046>
- Pei, Y., Chakraborty, N., & Sycara, K. (2015, June). Nonnegative matrix tri-factorization with graph regularization for community detection in social networks. In *Twenty-fourth international joint conference on artificial intelligence*.
- Peng, T.-Q., Liu, M., Wu, Y., & Liu, S. (2016). Follower-followee network, communication networks, and vote agreement of the U.s. members of congress. *Communication Research*, 43(7), 996-1024. <https://doi.org/10.1177/0093650214559601>

- Qomariyah, S., Iriawan, N., & Fithriasari, K. (2019). Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis. THE 2ND INTERNATIONAL CONFERENCE ON SCIENCE, MATHEMATICS, ENVIRONMENT, AND EDUCATION.
- Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1), 209-228.
- Reuben, R. (2008). The use of social media in higher education for marketing and communications: A guide for professionals in higher education.
- Ruiz, J., Featherstone, J. D., & Barnett, G. A. (2021, January). Identifying vaccine hesitant communities on twitter and their geolocations: a network approach. *In Proceedings of the 54th Hawaii international conference on system sciences* (p. 3964).
- Rutter, R., Roper, S., & Lettice, F. (2016). Social media interaction, the university brand and recruitment performance. *Journal of Business Research*, 69(8), 3096–3104. <https://doi.org/10.1016/j.jbusres.2016.01.025>
- Sanandres, E., Abello, R., & Madariaga, C. (2020). Topic modeling of twitter conversations: The case of the national university of Colombia. *In Studies in Classification, Data Analysis, and Knowledge Organization* (pp. 241–251). Springer International Publishing.
- Sayce, D. (2019, December 3). The Number of tweets per day in 2020. David Sayce. <https://www.dsayce.com/social-media/tweets-day/>
- Singh, L. (2022). Computer science educators’ use of Twitter for conference engagements: A grounded theory analysis. *The Journal of Social Media for Learning*. <https://doi.org/10.24377/LJMU.JSML.ARTICLE496>
- Surian, D., Nguyen, D. Q., Kennedy, G., Johnson, M., Coiera, E., & Dunn, A. G. (2016). Characterizing twitter discussions about HPV vaccines using topic modeling and community detection. *Journal of Medical Internet Research*, 18(8), e232. <https://doi.org/10.2196/jmir.6045>
- Syed, S., & Spruit, M. (2017, October). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. *In 2017 IEEE International conference on data science and advanced analytics (DSAA)* (pp. 165-174). IEEE.
- Tremayne, M. (2014). Anatomy of protest in the digital era: A network analysis of Twitter and occupy Wall Street. *Social Movement Studies*, 13(1), 110–126. <https://doi.org/10.1080/14742837.2013.830969>
- Twitter (2021). Twitter Annual Report. https://s22.q4cdn.com/826641620/files/doc_financials/2021/ar/FiscalYR2021_Twitter_Annual_Report.pdf
- Veletsianos, G., & Kimmons, R. (2012). Assumptions and challenges of open scholarship. *The International Review of Research in Open and Distributed Learning*, 13(4), 166. <https://doi.org/10.19173/irrodl.v13i4.1313>

- Viegas, R. R., & Xavier, L. B. (2021). The Political Use of Twitter by the Federal Prosecution Service in Brazil. *In Proceedings of the IAPSS Virtual World Congress*.
- Wekerle, C., Vakili, N., Stewart, S. H., & Black, T. (2018). The utility of Twitter as a tool for increasing reach of research on sexual violence. *Child Abuse & Neglect*, 85, 220–228. <https://doi.org/10.1016/j.chiabu.2018.04.019>
- Yin, C., Zhu, S., Chen, H., Zhang, B., & David, B. (2015). A method for community detection of complex networks based on hierarchical clustering. *International Journal of Distributed Sensor Networks*, 11(6), 849140.
- Zachos, G., Paraskevopoulou-Kollia, E.-A., & Anagnostopoulos, I. (2018). Social media use in higher education: A review. *Education Sciences*, 8(4), 194. <https://doi.org/10.3390/educsci8040194>
- Zhao, J., Lui, J. C. S., Towsley, D., Guan, X., & Zhou, Y. (2011). Empirical analysis of the evolution of follower network: A case study on Douban. *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*.
- Zhao, Y. (2013). Analysing twitter data with text mining and social network analysis. *In 11th Australasian Data Mining Conference (AusDM 2013)* (pp. 41-47).
- Zimba, O., & Gasparyan, A. Y. (2021). Social media platforms: a primer for researchers. *Reumatologia*, 59(2), 68–72. <https://doi.org/10.5114/reum.2021.102707>

APPENDIX A

Word clouds of accounts' descriptions in the five selected accounts' communities

UU

UVA

Wordcloud for Users description in UUBeta community

Wordcloud for Users description in uva science community

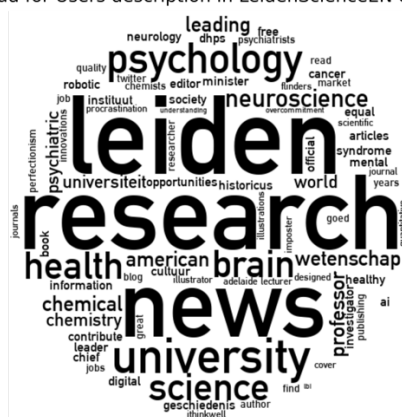


VU

LU

Wordcloud for Users description in VU Science community

Wordcloud for Users description in LeidenScienceEN community



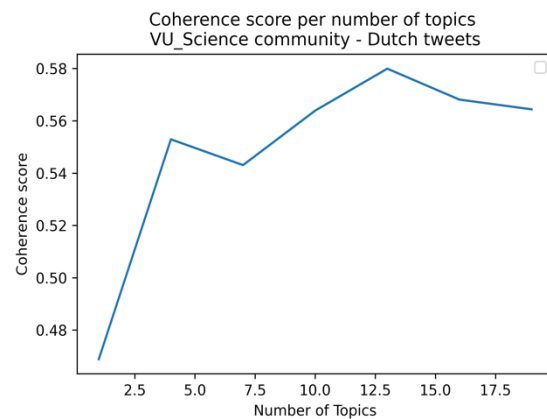
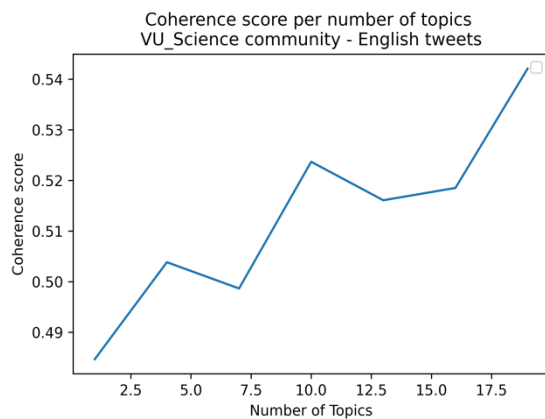
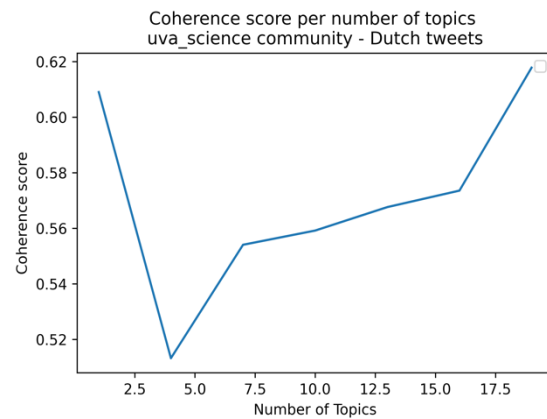
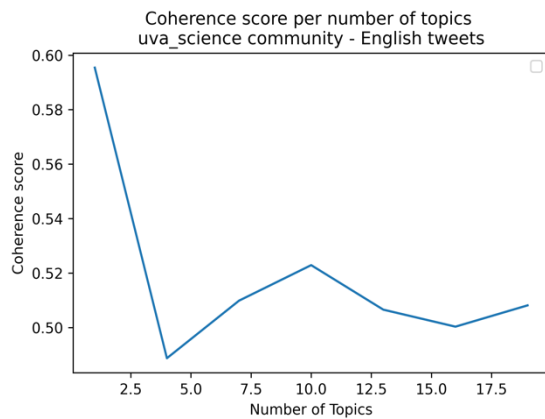
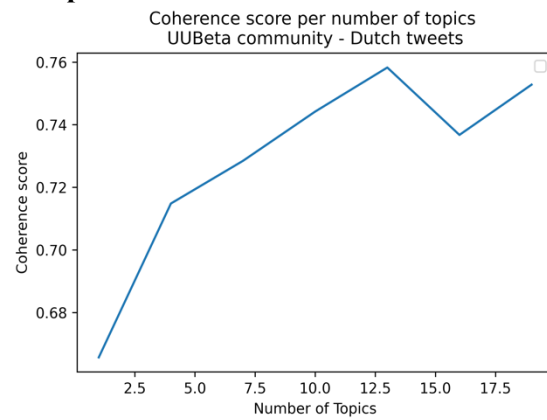
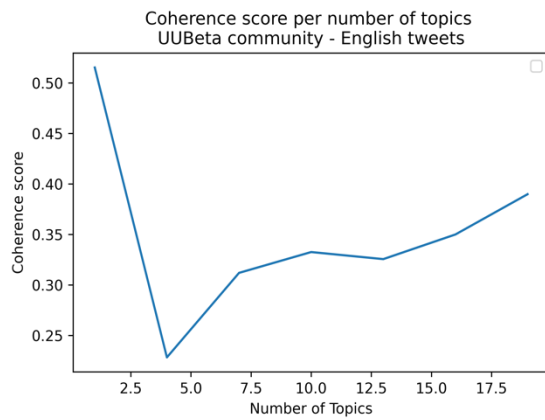
TU Delft

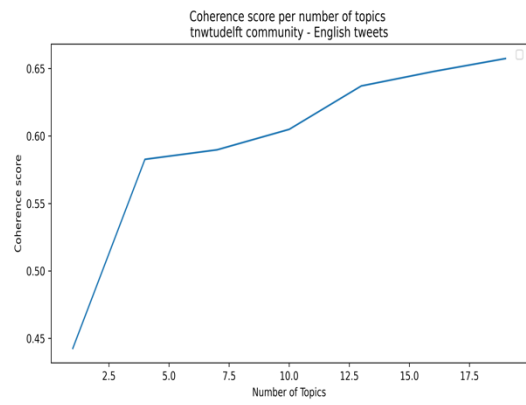
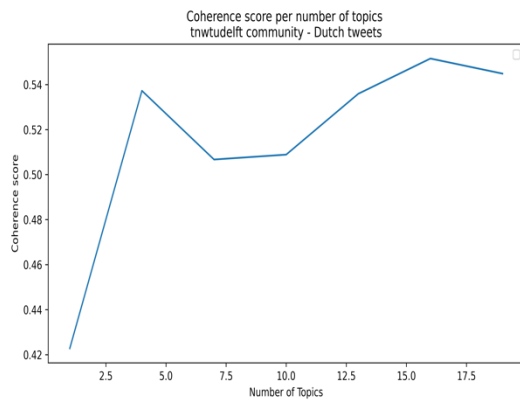
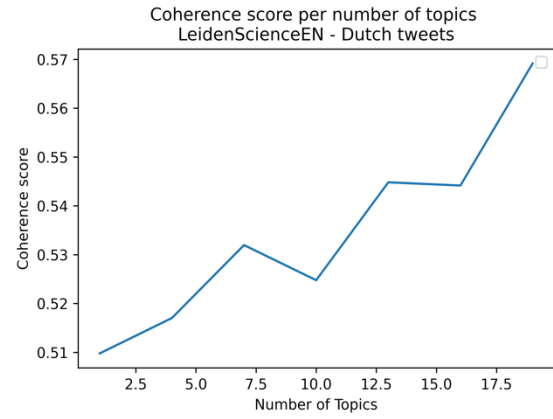
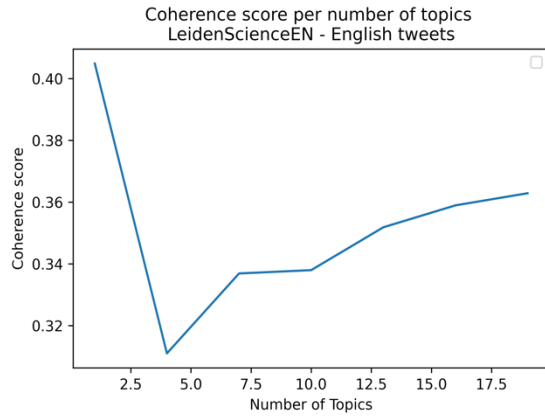
Wordcloud for Users description in tnwtudelft community



APPENDIX B

CV Coherence values for different topic numbers for each model





APPENDIX C

1. Faculty of Science at UU

1. Faculty of Science at UU

Top 10 words in each topic in English tweets

Topic	Label	Words
0	Social Issue	government worker cost price tax crisis pay inflation rise country
1	Sport Media	podcast report sport football club video school story woman news
2	Student Life	love friend work account book student family life news story
3	Politics	law government party country mp deal election war vote rule

Summed probability of each topic in English tweets

Topic	Sum_distribution
topic_0 sum	3.685124
topic_1 sum	3.221739
topic_2 sum	5.284453
topic_3 sum	5.808684

Word cloud of each topic in English tweets

Wordcloud for UUBeta community
English tweets - Topic 0



Wordcloud for UUBeta community
English tweets - Topic 2

Wordcloud for UUBeta community
English tweets - Topic 1



Wordcloud for UUBeta community
English tweets - Topic 3



Top 10 words in each topic in Dutch tweets

Topic	Label	Words
0	Research Project	onderzoek onderzoeker hoogleraar plant wetenschapper miljoen bioloog college interview collega
1	Higher Education Institutions	student vraag wetenschap bta faculteit studie onderwijs film onderzoeker docent
2	Environmental Issue	klimaatverandering water probleem plastic ijs zeespiegelstijging oceaan schimmel soep kaart

Summed probability of each topic in Dutch tweets

Topic	Sum_distribution
topic_0 sum	4.750110
topic_1 sum	4.512661
topic_2 sum	2.737229

Word cloud of each topic in Dutch tweets

Wordcloud for UUBeta community
Dutch tweets - Topic 0



Wordcloud for UUBeta community
Dutch tweets - Topic 1



Wordcloud for UUBeta community
Dutch tweets - Topic 2



2. Faculty of Science at UVA

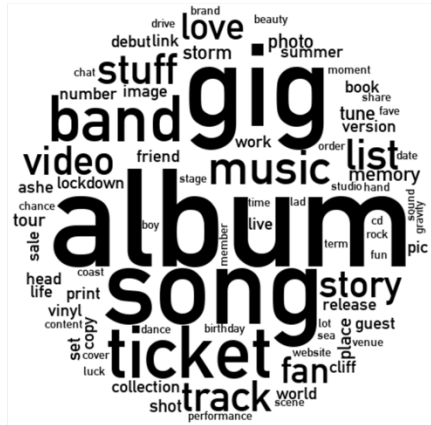
Top 10 words in each topic in English tweets

Topic	Label	Words
0	Music Concert	album gig song ticket band music list stuff track video
1	Environmental study	soil paper biodiversity climate phd plant project change research science
2	War criminal	court war judge crime family prosecutor case attack trial justice
3	Covid-19	vaccine rule country test health risk minister lockdown press government
4	Job	city team startup mayor meeting world job tech walk spring
5	General	police bike shot street report view country water film parliament
6	Pandemic Life	book story school woman friend home life kid work child
7	political Campaign	party election leader poll minister listening vote parliament politician light
8	Science	congratulation science research work world student physics art scientist physicist
9	Biking	news cycle path edition bike view story lane dune shot

Summed probability of each topic in English tweets

Topic	Sum_distribution
topic_0 sum	2.389729
topic_1 sum	1.481291
topic_2 sum	1.305958
topic_3 sum	1.129162
topic_4 sum	1.107985
topic_5 sum	1.184194
topic_6 sum	1.539998
topic_7 sum	0.908816
topic_8 sum	2.099605
topic_9 sum	0.853262

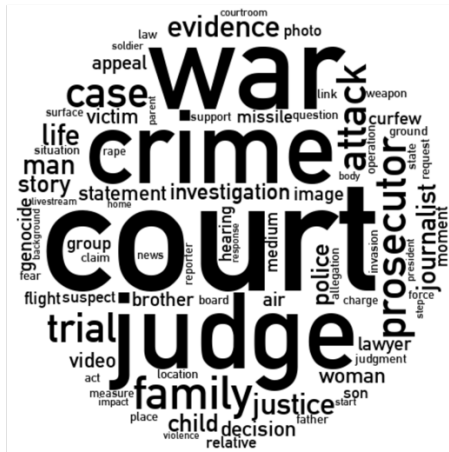
Wordcloud for uva_science community
English tweets - Topic 0



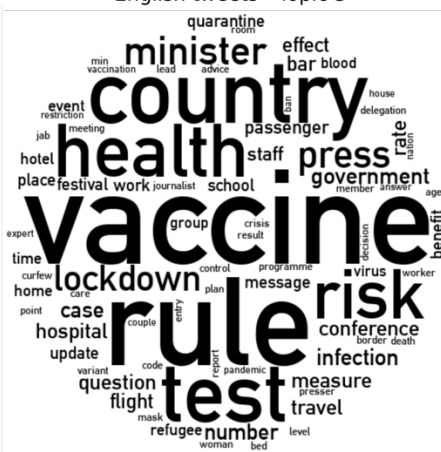
Wordcloud for uva_science community
English tweets - Topic 2

[illegible]

Wordcloud for uva_science community
English tweets - Topic 3



Wordcloud for uva_science community
English tweets - Topic 4



Wordcloud for uva_science community
English tweets - Topic 5



[illegible][illegible][illegible]

Topic	Label	Words
0	Environmental pollution	stikstofprobleem termijn depositie oproep coach niveau stad meten koolstof klimaat
1	Climate Change Impact	droogte besluit agrolobby oerknal overwinning opening winnaar dorp schaal term
2	Social Discussion	vraag gesprek kind minister man vrouw boer motie onderzoek wetenschap
3	General	model moeder politiek wereld slachtoffer snap steun verlies hoofd verantwoordelijkheid

Summed probability of each topic in Dutch tweets

Topic	Sum_distribution
topic_0 sum	0.538689
topic_1 sum	0.500506
topic_2 sum	12.233074
topic_3 sum	0.727730

Word cloud of each topic in Dutch tweets

Wordcloud for uva_science community
Dutch tweets - Topic 0



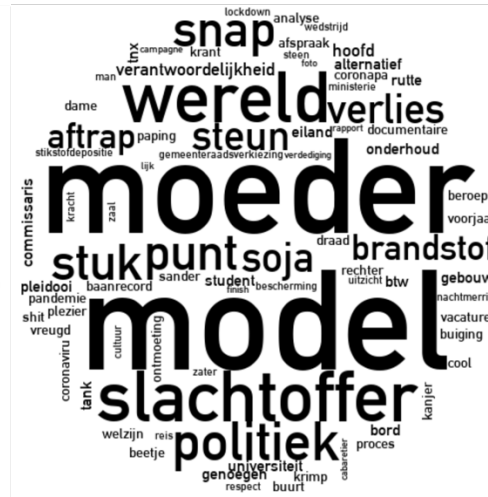
Wordcloud for uva_science community
Dutch tweets - Topic 2



Wordcloud for uva_science community
Dutch tweets - Topic 1



Wordcloud for uva_science community
Dutch tweets - Topic 3



3. Faculty of Science at VU

Top 10 words in each topic in English tweets

Topic	Label	Words
0	Music	album life video foal music song record love track tune
1	Alert Information	time info alert child pass post website kid work brain
2	Show	ticket book show deal tour review sky chance air date
3	Nature	world tiger climate nature forest sea fact animal scientist specie
4	Celestial Observation	credit space image star moon view planet rocket photo launch
5	Astronomy	asteroid article mission impact scientist space planet type rock project
6	Space Event	space event asteroid astronaut mission world program interview panel broadcast

Summed probability of each topic in English tweets

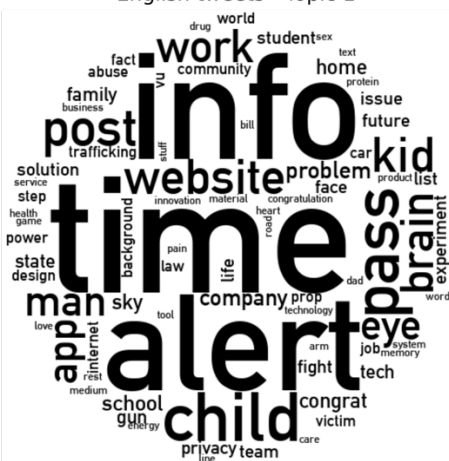
Topic	Sum_distribution
topic_0 sum	1.864914
topic_1 sum	1.557769
topic_2 sum	1.178685
topic_3 sum	2.139306
topic_4 sum	1.874674
topic_5 sum	1.479865
topic_6 sum	1.904787

Word cloud of each topic in English tweets

Wordcloud for VU_Science community
English tweets - Topic 0



Wordcloud for VU_Science community
English tweets - Topic 1



Top 10 words in each topic in Dutch tweets

Topic	Label	Words
0	Website Information	column model boek herkomst hoogleraar diversiteit eer website kost palmolie
1	Climate Change Impact	schade serie hoogtepunt klimaatdoel soja oceaan palmolie klimaatrapport frisdrank buitenland
2	Natural Science Research	natuur onderzoek dier hoogleraar bos actie wereld rapport wetenschapper vraag

Summed probability of each topic in Dutch tweets

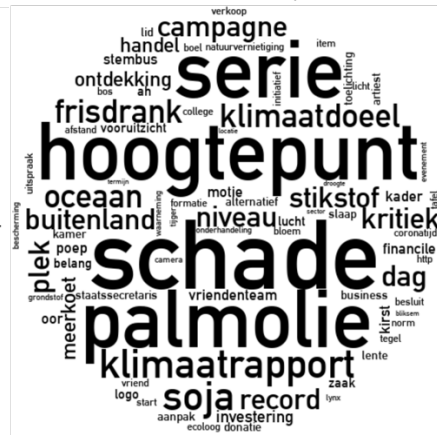
Topic	Sum_distribution
topic_0 sum	0.741833
topic_1 sum	0.430765
topic_2 sum	9.827402

Word cloud of each topic in Dutch tweets

Wordcloud for VU_Science community
Dutch tweets - Topic 0



Wordcloud for VU_Science community
Dutch tweets - Topic 1



Wordcloud for VU_Science community
Dutch tweets - Topic 2



4. Faculty of Science at LU

Top 10 words in each topic in English tweets

Topic	Label	Words
0	Disease Study	health care disorder patient pandemic study research anxiety time depression
1	Book/Author	book time author life story writing email work copy world
2	Doctoral Degree	research time student phd work congratulation paper question article team
3	Neuroscience	brain study cell disease researcher effect risk memory neuron disorder

Summed probability of each topic in English tweets

Topic	Sum_distribution
topic_0 sum	4.974495
topic_1 sum	3.847286
topic_2 sum	8.319250
topic_3 sum	4.858969

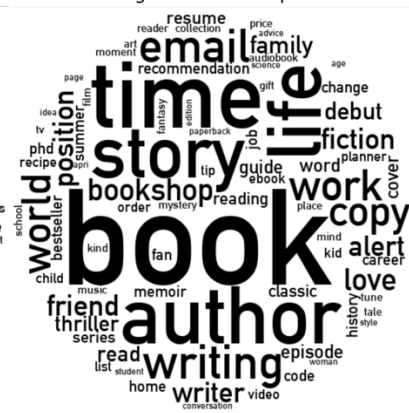
Word clouds of each topic in English tweets

Wordcloud for LeidenScienceEN community
English tweets - Topic 0

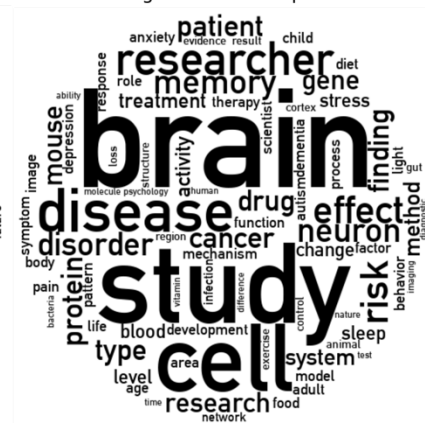
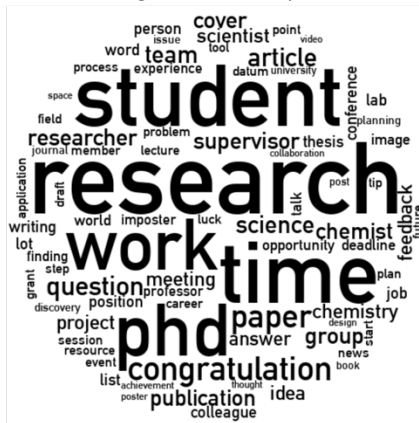


Wordcloud for LeidenScienceEN community
English tweets - Topic 2

Wordcloud for LeidenScienceEN community
English tweets - Topic 1



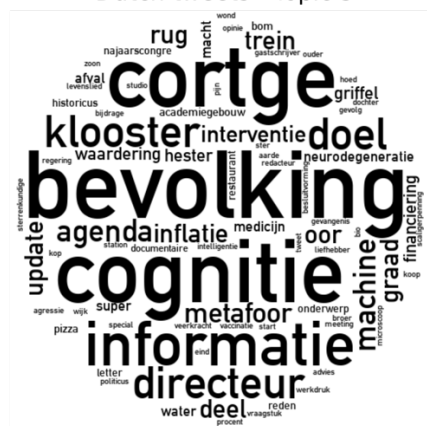
Wordcloud for LeidenScienceEN community
English tweets - Topic 3



Topic	Label	Words
0	General	tijd aanleiding bestwil speer politicus race principe transitie afbeelding kaart
1	Event	ticket kritiek verwachting oerknal storing artikel taart ruimte creativiteit ontdekking
2	Education	student onderzoek wetenschap hoogleraar universiteit wetenschapper boek onderwijs kind vraag
3	Information	bevolking cognitie cortge klooster directeur doel agenda machine metafoor informatie

Topic	Sum_distribution
topic_0 sum	0.493698
topic_1 sum	0.521754
topic_2 sum	9.597146
topic_3 sum	0.387402

Wordcloud for LeidenScienceEN community Dutch tweets - Topic 0 Wordcloud for LeidenScienceEN community Dutch tweets - Topic 1



5. Faculty of Science at TU Delft

Top 10 words in each topic in English tweets

Topic	Label	Words
0	University	student researcher university congratulation phd team cell work research time
1	Research Project	research researcher programme grant project science information funding round application

Summed probability of each topic in English tweets

Topic	Sum_distribution
topic_0 sum	3.737192
topic_1 sum	3.262808

Word cloud of each topic in English tweets



Wordcloud for tnwtudelft community
English tweets - Topic 1



Top 10 words in each topic in Dutch tweets

Topic	Label	Words
0	War News	oorlog land liveblog sanctie president stad update gas wapen leger
1	Study Sponsor	onderzoek student onderzoeker universiteit onderwijs wetenschap wetenschapper vraag minister miljoen
2	Sport Event	sport wk voet uitgangspunt dier beloning metafoor sportzomer community themanummer
3	Construction/ Project	vermogen gebouw draag project aardgas sterrenstelsel verlies migratieachtergrond matchmaking klacht

Summed probability of each topic in Dutch tweets

Topic	Sum_distribution
topic_0 sum	1.509613
topic_1 sum	5.713899
topic_2 sum	0.281900
topic_3 sum	0.494589

Word cloud of each topic in Dutch tweets

Wordcloud for tnwtudelft community
Dutch tweets - Topic 0



Wordcloud for tnwtudelft community
Dutch tweets - Topic 1



Wordcloud for tnwtudelft community
Dutch tweets - Topic 2



Wordcloud for tnwtudelft community
Dutch tweets - Topic 3

