APPLIED DATA SCIENCE MSC

THESIS PROJECT

PREDICTING THE TYPE OF ENGAGEMENT FOR UNIVERSITIES' TWITTER FEEDS

Author:

Paula Martínez Vidal (2323664) p.martinezvidal@students.uu.nl

Supervisors:

Dennis Nguyen

Mirko Schaefer

TABLE OF CONTENTS

1. ABSTRACT 5

2. INTRODUCTION 6

2.1. MOTIVATION AND CONTEXT 7
2.2. LITERATURE REVIEW 8
2.2.1. Universities' Social Media Communications 8
2.2.2. Science Communication 10
2.2.3. Twitter and Communities 11
2.2.4. User engagement 12
2.3. RESEARCH QUESTION 14

3. DATA & METHODS 15

3.1. EXPLORATORY ANALYSIS 17
3.1.1. Data preparation and methods 17
Topic Clustering 17
Network analysis 19
3.1.2. Selected data exploration results 20
Topic clustering 20
Science / Research Topic 21
Network Analysis 22
3.2. DATA ANALYSIS 24
3.2.1. Data Preparation and Methods 24
Human Selected Features 25
Machine-extracted features 25
Machine Learning and Deep Learning Models 26
3.3. TRANSLATION OF THE RESEARCH QUESTION TO A DATA SCIENCE QUESTION 27
3.4. ETHICAL AND LEGAL CONSIDERATIONS OF THE DATA 27

4. RESULTS 28

4.1. SELECTED ANALYSIS RESULTS 28
Human Explanatory Variables 28
Classifiers Performance 30

5. DISCUSSION AND CONCLUSION 33

6. LIST OF REFERENCES 37

7. APPENDIX 43

7.1. FULL DATA EXPLORATION RESULTS 47
7.2. ATTACHMENTS 52
7.2.1. Tweets Examples 53

LIST OF FIGURES

Figure 1. Analysis Overview 16
Figure 2. Exploratory Analysis Overview 17
Figure 3. Science / Research clusters wordclouds 21
Figure 4. Retweet Network 22
Figure 5. Replies Network 23
Figure 6. Analysis Overview 24
Figure 7. Multimodal Toolkit Structure 26
Figure 8. Percentage of use of Explanatory variables 29
Figure 9. Results based on Macro Avg Precision 31
Figure 10. Results based on Macro Avg Recall 31
Figure 11. Results based on F1 Score 32
Figure 12. Knee detection line chart for Higher Ranked Universities 46
Figure 13. Knee detection line chart for Middle Ranked Universities 46
Figure 14. Knee detection line chart for Middle-Low Ranked Universities 46
Figure 15. Tweet from the Arizona State University 53
Figure 16. Tweet from Utrecht University 53
Figure 17. Tweet from Kansas State University 53

LIST OF TABLES

Table 3.1. Distribution of tweets per account 15
Table 3.2. Clustered accounts according to the 2021 Shanghai Academic Ranking 16
Table 3.3. Number of English tweets per account 16
Table 3.4. Distribution of tweets per clusters 20
Table 3.5.1.3.6 Distribution of classes in the sample 25
Table 7.1. Overview of the datasets from Twitter API used 43
Table 7.2. Human selected variables description 43
Table 7.3. Personalized stopword list 44
Table 7.4. Node table example 44
Table 7.5. Edge table example 44
Table 7.6. Relevant words from the Topic Clustering 45
Table 7.7. Topic Clustering Results 45
Table 7.8. Selected Explanatory variables 47
Table 7.9. Chi-Square Results for High Ranked Universities 48
Table 7.10. ANOVA Results for High Ranked Universities 48
Table 7.11. Chi-Square Results for Middle Ranked Universities 49
Table 7.12. ANOVA Results for Middle Ranked Universities 49

Table 7.13. Chi-Square Results for Middle-Low Ranked Universities 50
Table 7.14. ANOVA Results for Middle-Low Ranked Universities 50
Table 7.15. Complete Results of the Classifiers' Performance 51
Table 7.16. Confusion matrix from the Classifiers Results 51

1. Abstract

Many studies show how to engage with audiences on social media, but a lack of studies shows how universities use social media accounts in the scientific research domain. Therefore, based on the research gap, the present study aims to contribute to the field of predicting the most probable type of engagement (like, retweet, or reply) for ten university official Twitter accounts. Moreover, the study also proposes to find some of the features contributing to this prediction. In order to predict the type of interaction, the research uses a combination of human-selected and machine-extracted features to train three machine learning models (Logistic Regression, Random Classifier, and LightGBM) and a deep learning model (neural network using BERT model). Human selected features are mainly binary variables that contain tweet information, while machine-extracted features are large-dimensional features that we obtain from the texts of the tweets. The results show that by combining both types of features, we can predict the most probable type of engagement and an overview of the features that contribute to this prediction, such as if the tweet contains a hashtag or if the tweet is a reply. Also, the findings show that the best method to predict this engagement is LightGBM and neural networks.

Research and practical implications include helping practitioners to create the content strategy based on the engagement objectives and providing more knowledge to help them understand which features contribute to the type of engagement.

2. Introduction

Social media have changed the way we communicate and interact. They are the primary source of communication for many individuals (Das et al., 2022) and also for businesses, as they can expose their products or services to their consumers.

Interacting and managing effectively these platforms has become a necessary task for organizations to stay visible and relevant to their consumers (Lemon & Verhoef, 2016). They can cost-effectively present their values, mission, and achievements to attract and share content with their audiences. For this reason, managing social media effectively and engaging with its users has become a priority for these organizations (Weale, 2019).

A broad field of research has explored the interaction between customers and (private and public) organizations, also commonly defined as *customer engagement* (Sashi, 2012; Hollebeek et al., 2014; van Doorn et al., 2010). Sashi (2012, p. 253) defines this relationship briefly as "customer engagement turns customers into fans". Comparably, van Doorn et al. (2010, p. 254) describe customer engagement as behavior that surpasses the act of purchasing. *Customer engagement* is defined as a set of "behavioral manifestations", emotions that drive individuals to act— for example, posting a positive review on their social media.

Engaging with the users is not a simple task. It requires time, effort, and a clear and consistent marketing strategy to meet specific goals. A report published by the Economist Intelligence Unit (2007) acknowledged the importance of technology as a tool to maintain an effective relationship with stakeholders. Moreover, they state that level of engagement is the key to success for organizations.

Indeed, social media platforms are relevant tools to build and maintain user engagement. In particular, the present study focuses on how higher education institutions engage with their users. Previous studies state that universities consider it highly important to use social media for promoting content through their channels (Constantinides & Zinck Stagno, 2011; Marcus, 2021). Motta & Barbosa (2018) conclude in their study that all top one-hundred universities examined in Europe and US from Academic Ranking World Universities (ARWU) use online marketing to promote themselves.

These social media tools help organizations connect with their audiences. The primary audiences of universities' social media include potential and former university students and researchers (Brech et al., 2016). Both audiences follow university accounts to be updated about the developing projects and news despite presenting different interests, expectations, and needs. Other universities' social media audiences include students' parents and employees from the institution.

Research has shown that the universities' most used social media platforms are Facebook and Twitter. Motta & Barbosa (2018) noted that the most trendy social media platforms were Twitter and Facebook, followed by Youtube, Instagram, and LinkedIn. Canada et al. (2014)

show the different purposes that best suit Facebook and Twitter. Facebook is more beneficial for broadcasting events on campus, while Twitter is more effective for establishing direct communication and replying to answers. As the present study focuses on the user engagement of university accounts, Twitter is a more suitable platform for the approach.

Twitter offers different types of engagement metrics to users. The main ones are retweet, like, and reply. Similarly, Lund (2019) describes how user engagement metrics for Facebook help measure the post's efficaciousness with the audience. For Twitter, the most used metrics are *likes*, *retweets*, and *replies*. A *retweet* is a tweet initially posted by another account. When the user retweets a post, it will be added to his feed, including the original user of the tweet. A *like* is the most basic metric of engagement that represents that the user liked the content posted by anyone. It is the most effortless way to show that the user engages with the matter. Both metrics are considered positive types of engagement (Vargo, 2016). Kwak et al. (2010) distinguish the *retweet* and the *like* by the sharing behavior that the retweet implies. Whereas the *reply*, responding to a tweet, requires more time and effort. It does not necessarily show that they liked or enjoyed the post, but it shows the sentimental reaction of the stakeholder (Bliss et al., 2012).

Using these engagement metrics, the main goal of this research project is to determine if we can and what factors reliably predict the most probable user behavior in the context of university Twitter feeds. More specifically, the study centers on ten universities accounts for the science and research domain.

The present paper starts with a critical literature review of the key concepts: Universities' Social Media Communication, Science Communication, Twitter and its communities, and User engagement. Lastly, we examine the aim of the research. In the following sections, we present the dataset and sample used and the procedure used for the analysis. Also, we discuss the research outcome and present the study's limitations and future research and practical implications for universities.

2.1. Motivation and context

Utrecht University Faculty of Science department wants to understand what factors increase the participation and engagement of their users. Their main objective is to reach a broader audience and increase engagement with their followers. Another goal is to raise brand awareness and increment the interaction in science and research content. Increasing visibility also enhances and strengthens the university's research reputation in the Twitter science community and the university brand and attracts scientists to their research.

The marketing team of the faculty of science mostly shares science news and peer-review published content. They aspire to engage in scientific discussion and establish Utrecht University as a thought leader in this context. Hence, the main focus of this study is on analyzing science and research content-related posts.

In order to assist them, the selected approach chose to analyze two of their social media accounts on Twitter. @UUBeta, the faculty of science account with approximately 3,9K tweets, and their official higher education account @UniUtecht with roughly 20K tweets. Eight accounts from other US universities are selected based on academic ranking performance to compare how other universities engage with their audiences. In particular, US universities were sampled for their reputation and visibility in global rankings. In addition, technical considerations related to the Natural Language Processing of English texts also played a role.

After the empirical analysis and critical reflection, the paper concludes by giving an overview of the contribution of the present study and the practical implications for universities' Twitter feeds' in the science and research field.

2.2. Literature Review

In this section, the available literature on five relevant topics is discussed: Universities' Social Media Communications, Science Communications, Twitter and Communities, and User Engagement. Each topic contributes to the current research's essential goal and domain.

2.2.1. Universities' Social Media Communications

Higher education has adapted to technology and connects with its audiences through social networking sites such as Facebook, Twitter, and Instagram. Several studies have focused on studying how universities engage with users through social media platforms (Assimakopoulos et al., 2017; Clark et al., 2016; Brech et al., 2017).

Motta & Barbosa (2018) did an analysis of the higher education social media approach from the Top-100 universities in the US and Europe. Their analysis showed that US universities have more followers on Twitter than their Europe counterparts. One of the factors contributing to this fact is that US universities post three times more content than European universities. Nevertheless, both countries consider Twitter as one of the most favorite social media platforms and the most suitable platform for interacting with users.

Interacting with different audiences implies taking into account distinct needs and expectations from the stakeholders. Particularly, for universities, undergraduate students and the scientific community are two of the most relevant stakeholders.
On one side, undergraduates' interest is to get a general vision of the institution and the programs for decision-making when selecting their future studies (Brech et al., 2016). Universities use this platform to portray or broadcast different elements such as student life on campus (Taecharungroj, 2017) and offer study information about their programs. Stvilia & Gibradze (2017) discuss the value of using social media platforms to advertise study programs and share content with students. Moreover, they infer that the type of content posted also affects the degree of engagement.

On the other side, the science community focuses, in particular, on extending its network, sharing published research, and maintaining communication within its circle (Priem &

Costello, 2010; Veletsianos, 2011). Following accounts that broadcast this type of scientific content is essential for them. For instance, Knight & Kaye (2014) studied the usage of Twitter among academics and students through a survey. They showed that researchers are more interested in strengthening their reputation instead of centering their attention on the platform's utility to connect with students. This lack of use is due to an absence of knowledge of the potential benefits.

Similarly, Linvill et al. (2012) and Veletsianos et al. (2017) highlight that university accounts tend to broadcast information to the public, reducing engagement with their audiences instead of using platforms to increase user interactions and provide a two-way communication. Besides, Kimmons et al. (2016) provide evidence of how US universities use Twitter. Their results confirm that they are not using Twitter to its maximum capacity as a communication platform for educational purposes. The findings suggest that their tweets do not invite users to participate; instead, they are monotonous and neutral in sentiment. Increasing the communication and engagement in this platform can offer new possibilities for universities as a social network in the educational context.

Moreover, Veletsianos et al. (2017) mention how these university accounts try to portray an unrealistic and inaccurate vision of the institution because they present only positive or relevant events happening within the institutions. As a result, they suggest that higher institutions use Twitter mainly for recruiting and brand promotion. Likewise, Peruta & Shields (2016) report that the universities' social media uses its accounts mainly as marketing and branding tools. For instance, to promote student life, newly published research, and milestones such as students' graduation or university anniversaries.

In brief, universities' social media can help promote the institution and increase brand awareness. However, universities must understand and be conscious of what social media can offer them and their limitations.

In the case of universities, which are educational institutions, their use of social media needs should differ from other organizations offering products and services. It should be closer to the non-profit trademark (Peruta & Shields, 2017). However, the reality is distorted, as universities feel obliged to promote and advertise their services as any other business due to digitalization (Maresova, Hruska & Kuca, 2020). On top of that, society values the universities' efforts to understand their stakeholders' needs and adaption to meet them. In other words, those universities that fulfill those needs have a competitive advantage. Teh et al. (2011) conclude that the sector with increasing competition and pressure among universities does not make it suitable to use social media as a non-profit organization.

Even though much research has been done about university communications on social media platforms, there are still some gaps in the literature. For example, Veletsianos et al. (2017) argue that there is a lack of studies analyzing user engagement in the high education social media context. Moreover, they propose investigating features that increase user engagement for future studies. Likewise, Peruta & Shields (2016) suggest looking into the features that create more engaging interactions on Facebook. They highlight features such as the hours, days of the posts, and the content posted. Furthermore, Kimmons et al. (2016) mention the limitations of previous studies due to the sample bias, as some only included top-ranked universities in their studies.

Our research fills the current gap in the literature as it contributes to the user engagement in social media literacy in higher education institutions by exploring user engagement and the features that contribute to it. Besides, it broadens the sample used, combining all types of universities (high, middle, and middle-low ranked) to overcome the popularity bias.

2.2.2. Science Communication

Universities use social media to engage in science debates and communicate about research. This links to the topic of science communication. Burns, O'Connor & Stocklmayer (2003, p. 191) define science communication as "the use of appropriate skills, media, activities, and dialogue to produce one or more of the following personal responses to science. "

Science communication plays a crucial role in higher education. It strengthens the universties reputation, and as a result, attracts students and scientist to their institutions. Therefore, universities are expanding their workforce to include individuals that exclusively focus on science communication matters (Trench, 2017).

Trench (2012) explains the duality of science communication for higher education institutions. On one side, it is essential for these powerful entities. They are entitled to transmit wisdom, reach different collectives, and actively engage in science-related discussions. However, on the other side, scientific communication is a susceptible concept for them. It can go against universities' conventional objectives and limitations, causing detriment.

Studies show that social media use can give institutions some advantages. Olvera-Lobo & López-Pérez's (2014) results show that universities use social media platforms to promote scientific communication. In particular, Davis (2014) exposes that Universities' ambitions to participate in promoting science communication include recruiting, strengthening their reputation, and improving the overall satisfaction of stakeholders.

Besides, researchers also use social media in a comparable way. They want to reach more users and give exposure to recently published studies. In addition, they also want to inspire younger generations to study or get involved in science disciplines (Burns, O'Connor & Stocklmayer, 2003).

Conversely, Gruzd, Staves & Wilk (2012) highlight that one of the major benefits for researchers of joining social media is the capacity to form new connections with peers. In particular, junior researchers also use these networks to enhance their professional image. Moreover, Collins & Hide (2010) mention that a positive correlation exists between the active use of social networks for professional use and the support from their surroundings, including institutional entities.

Nevertheless, social media platforms can harm scientific communication (Fox et al., 2021). The reason lies in that anyone can post content online and get massive attention, but it does not mean it is reliable. In other words, users can spread misinformation to their audience without knowledge or background in the field, or even without any proof.

Also, social media posts are static, while scientific knowledge is constantly evolving. For example, new research providing some evidence can be disproved later, while users can still spread outdated information.

Moreover, social media publications can cause oversaturation of the users. In short, it can be a challenging process for the user to critically filter out the relevant information from the large amounts of posts.

Furthermore, we can highlight a gap in the science communication literature among the published studies. For example, Hwong et al. (2017) mention a lack of studies about social media in science communication. Similarly, Fox et al. (2021, p.1630) explain that more studies in science communication with social media platforms are required:

> Research is needed on these new forms of digital media. Although some platforms may be very popular, it is not yet known whether they are effective strategies for knowledge translation, which platforms are most effective, or even how we should evaluate outcomes related to these new approaches for scientific communication.

2.2.3. Twitter and Communities

Many studies have been carried out about Twitter in different domains. From analyzing fake news during the 2016 presidential elections in the US (Grinberg et al., 2019) to using Twitter for detecting real-time events using individuals as sensors (Zhao et al., 2011). Including studies on illness, symptom detection, and how individuals use medication (Paul, & Dredze, 2021), recognizing and analyzing academics on Twitter (Ke, Ahn, and Sugimoto, 2017), and using Twitter to improve learning Marketing in an educational setting (Lowe & Laffey, 2011).

Likewise, Twitter is an essential social media that the science community uses because it is global, practical, and cost-effective advantages, and in addition, it updates information in real-time. Therefore, it is one of the more suitable platfoms for researchers to communicate their studies (Veletsianos, 2011).

Researchers, as well as journal accounts, post their publications, tweeting forthcoming events with the objective to expose the content broadly to a larger audience and promote scientific research. They even use social media platforms to advocate science for younger students (López-Goñi & Sánchez-Angulo, 2017). In addition, scientists use this social media to participate in discussions, share ideas and views, and keep up with the new trends in the field (Ke, Ahn, and Sugimoto, 2017). Consequently, researchers use Twitter in order to divulgate science, and to increase and strengthen their reputation (Knight & Kaye, 2014).

Besides promoting science and research, Twitter helps researchers create networks and establish communities (Mangold & Faulds, 2009). Science discourses involve communicators that can form communities of different scopes and densities in social media networks. Communities and their underlying dynamics are essential subjects in research on social media.

A community can be defined "as a set of people who share sociability, support, and a sense of identity" (Gruzd et al., 2011, p. 1295). Members inside a community use these interactions to learn and engage with the accounts they follow (Soukup, 2006). One step further, Burke (1974, p.227) established a connection between the groups and movements that individuals follow and their own personal identification. They identify that those people, ideas, and movements individuals follow are "one's ways of seeing one's reflection in the social mirror ."
The concept of community translates into online communities in the social media setting. Baym & Jones (1995, p.152) define online communities as "group-specific forms of expression." These online communities have been the focus of many studies to understand social network dynamics on Twitter (Gruzd et al., 2011; Blight et al., 2017).

Scientists, as well as other users, create networks on Twitter where they can exchange information within its communities. Dron & Anderson (2009) present a study analyzing online social groups. They highlight that the social ties presented in Twitter networks are unique, forming different networks for various individuals. They explain how the networks are formed by strong and weak ties meaning that the type of relationship with the different users varies. Likewise, Gladwell (2010) argues that this social media is mainly built from loose ties, not similar to our real-world network, which is mainly constructed by strong ties.

De Melo Maricato and de Castro Manso (2022) analyzed the science communities in Brazil and found that most of the profiles are individual rather than from institutions. They also point out a significant untapped potential because institutional accounts have higher average rates of followers than individual accounts and the highest mentions in research published. Based on the facts, they highlight that high educational accounts have more potential for outreach and impact to promote and spread science content through Twitter. Thefore, universities' accounts can take advantage of these findings to increase their reputation.

In short, while a large field of research has examined how scholars use social media platforms to promote their research and share publications, Veletsianos (2011) states that the number of studies in literacy centered on how higher education institutions use social media for research purposes is scant.

2.2.4. User engagement

Universities and companies use social media platforms to promote and enhance their brand image through relationships with stakeholders. This relationship is a form of user engagement. More precisely, Attfield et al. (2011, p. 2) define *user engagement* as "the emotional, cognitive and behavioural connection that exists, at any point in time and possibly over time, between a user and a resource."

User engagement has been widely explored by researchers in many domains such as Artificial intelligence, Politics, and Marketing. For instance, DeMasi et al. (2016) present a study highlighting how hashtag usage affects user engagement in different settings and communities. The results show that hashtag usage is more efficacious in communities formed by strong ties. In a different setting, Hu et al. (2015) built a recommender system to suggest Twitter news based on user engagement. Some important factors that contribute to these predictions are interests, geolocalization, and network structure. Another example is Siyam, Alqaryouti & Abdallah (2019), which use the government's tweets to predict the civilians' engagement. They use some features from the tweets (day, time) to help predict this engagement.

In the case of Twitter, Muñoz-Expósito, Oviedo-García & Castellanos-Verdugo (2017) review metrics for customer engagement. According to their study, the most used type of engagement metrics are *likes*, *retweets*, and *replies*. Besides, they expose the importance of designing an engagement strategy using the right metrics accordingly.

Other prior investigations found that question marks are valuable predictors of engagement as they invite or call to answer the post in the science social media domain (Hwong et al., 2017). Similarly, Suh et al. (2010) came across that external links mentioned in the tweet and hashtags increase the chance of being retweeted. Besides, the use of hashtags also helps the formation of communities based on a common interest (Su et al., 2017).

Consequently, these metrics contribute to strengthening and maintaining the relationship with the users. In particular, the present study has some foundations based on a previous study (Dai & Wang, 2021). Their study context covers social media posts, particularly Weibo, from three Telecommunication state-owned companies in China.

They predict the type of engagement (like, share, and comment) based on human and machine-extracted features. Human features are selected variables that contain information from the posts, such as if the post was published at the weekends or if the post contains question marks, while machine-extracted features represent the transformations and vectorizations of the post using a bag of words and neural networks. In order to predict the outcome, they use three machine learning (ML) algorithms (Logistic Regression, Random Forest, and LightGBM classifiers) and a deep learning model (neural networks using the BERT model). Their results confirm that the use of human and machine-extracted features improves the prediction of the most probable type of engagement.

In a similar way, the present study incorporates elements from Dai & Wang's (2021) research in a different context, higher education in the science and research domain.

As shown above, many studies provide insights into how universities use their social media to engage with students for branding and recruiting, as well as how scientists use their Twitter accounts to promote their research and connect with their audience. However, literacy about how universities use social media for research intentions and which platforms contribute the translation of knowledge in the science communication domain is not yet explored.

2.3. Research question

Based on the previous discussion, the following key concepts are an existing research gap on social media platforms in science communication and how high educational institutions use social media in the scientific research field. Therefore the selected research questions are two: Is it possible to predict the type of user engagement based on posts published in their Twitter feeds? What are the features that contribute to predicting this engagement type?

Predicting the type of user engagement can help universities provide a clear overview of the most probable form of engagement, give more context, and help them shape their publications according to their objectives, such as increasing visibility or encouraging user interaction. Besides, the study aims to provide insights into the features contributing to this prediction. Those features also provide understanding for practitioners on what features affect the type of engagement.

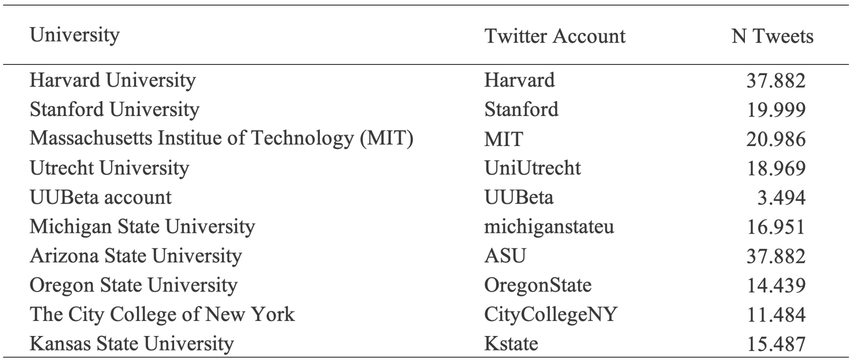
In order to predict the type of user engagement (like, retweet or reply) from ten official university accounts, machine learning (ML) algorithms and deep learning algorithm with neural networks are used.

Therefore, the study implications include practical applications for higher educational institutions and research implications for contributing to university social media use in the science and research context literature.

3. Data & Methods

In order to predict user engagement, we use the Twitter API to collect the tweets from the ten accounts from different universities. Our motivation to collect the data from the API using different accounts is due to a lack of data volume from the initial data provided by the Utrecht Science Faculty Department. Also, including other higher education institutions to provide a general overview of other universities.

The data extracted from the API is saved in four separated CSV files: the tweets, replies, the users, and the attachments (media content). **Table 7.1** presents an overview of the datasets and variables used. The datasets are combined to obtain the final dataset that includes all the accounts.



University	Twitter Account	N Tweets
Harvard University	Harvard	37.882
Stanford University	Stanford	19.999
Massachusetts Institue of Technology (MIT)	MIT	20.986
Utrecht University	UniUtrecht	18.969
UUBeta account	UUBeta	3.494
Michigan State University	michiganstateu	16.951
Arizona State University	ASU	37.882
Oregon State University	OregonState	14.439
The City College of New York	CityCollegeNY	11.484
Kansas State University	Kstate	15.487

Table 3.1. *Distribution of tweets per account*

The accounts selected are in **Table 3.1**.The choice to select those universities is based on two criteria: they are included in the Shanghai Academic Research Ranking and have an official active Twitter account (Shanghai Ranking, 2021).

For the first criteria, a selection of these universities is based on the ranking performance to get three separate groups: high, middle, and middle-low performance universities. Using this criterion, we avoid only studying elite institutions because the result could be biased due to the popularity of high-performance institutions.

For the second criterion, universities with an official Twitter account actively used were selected. By actively use of the account, the measure decided is a minimum threshold of 10.000 tweets per account. In other words, all selected accounts have more tweets than the selected threshold except for the Utrecht Faculty of Science (@UUBeta) account, which only contains around 3,4K tweets. Also, apart from @UUBeta, all accounts chosen are the official university accounts. Official accounts provide more visibility and outreach, as they tend to have more significant amounts of tweets and followers than secondary accounts from the universities (faculty departments).

Using these accounts, the study aims to predict the most probable type of engagement. In order to predict this metric, we gathered all the available tweets from the ten mentioned accounts. In total, we collected 204.181 tweets from different accounts, see **Table 3.1**. The sample includes tweets from March 26, 2006, until the day the data was collected, May 20, 2020.

In order to categorize these accounts, the 2021 Shanghai Academic Ranking is used. The main criteria to divide and cluster these accounts are academic performance, the total score, and the alumni score into three main groups: high-performance universities (HU), middle-performance universities (MU), and middle-low performance universities (LU). The first cluster includes 85.475 tweets, the second 77.296 tweets, and the last 41.410 tweets, as presented in **Table 3.2.**

World rank	University	Twitter Account	Country	National / Regional Rank	Total Shanghai Score	Alumni	Cluster	Amount of Tweets
1	Harvard University	Harvard	US	1	100		1	85.475 tweets
2	Stanford University	Stanford	US	2	75,9	45		
4	Massachusetts Institue of Technology (MIT)	MIT	US	3	69,5	71,6		
50	Utrecht University	UniUtrecht	Netherlands	1	33	21,4	2	77.296 tweets
-	UUBeta account	UUBeta	Netherlands	-	-	-		
101-150	Michigan State University	michiganstateu	US	41-56	-	7.5		
101-150	Arizona State University	ASU	US	41-56	-	0		
201-300	Oregon State University	OregonState	US	63-89	-	7.5	3	41.410 tweets
401-500	The City College of New York	CityCollegeNY	US	111-129	-	27,7		
401-500	Kansas State University	Kstate	US	111-129	-	0		

Table 3.2. *Clustered accounts according to the 2021 Shanghai Academic Ranking*

All selected universities are based in the US except University Utrecht. US universities are sampled for their reputation and visibility in global rankings. The language selected is English, which, according to Haustein et al. (2018), is the predominant language in social media science posts. In addition, technical considerations related to Natural Language Processing of English texts also played a role.

Therefore, the study only considers English tweets (see in **Table 3.3**) and limits the outreach of local communities that post in other languages. Another limitation in the data collection process is the metrics available to extract from the API. Promoted tweets, other engagement metrics, and direct messages can not be collected using the Twitter Academic Developer account.

University Cluster	University Account	English Sample	University Cluster
High Ranked Universities	Harvard	44.068	84.697
	MIT	20.775	
	Stanford	19.854	
Middle Ranked Universities	ASU	36.591	55.206
	michiganstateu	16.207	
	UniUtrecht	1.374	
	UUBeta	1.034	
Middle-Low Ranked Universities	KState	15.285	39.714
	OregonState	13.470	
	CityCollegeNY	10.959	

Table 3.3. *Number of English tweets per account*

In brief, this section exposes the data wrangling process and methods used to prepare the data for the analysis, the results from the data exploration phase, the methods used for the analysis, the translation of the research question into a data science question, and the ethical and legal considerations of the data.

In order to conduct the research, we follow the **Figure 1**. Analysis Overviewworkflow.

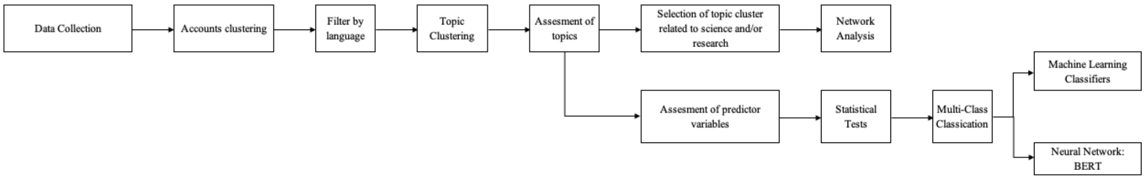


Figure 1. *Analysis Overview*

After selecting the English tweets, we do an exploratory analysis of the data that consists of topic clustering and network analysis. The exploratory analysis aims to understand the data better and give a better context to our research.
Subsequently, we process the data for the analysis and use the multi-class models to classify the tweets. We use three ML algorithms (Logistic Regression, Random Forest, and LightGBM) and multimodal transformers using BERT model.

3.1. Exploratory analysis

3.1.1. Data preparation and methods

In the exploratory phase (see in **Figure 2**), we do a topic clustering to select the tweets clustered as science or research content-related, and a network analysis with the retweets and replies from the ten universtity accounts.

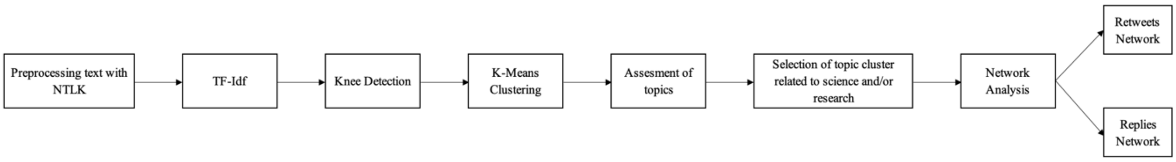


Figure 2*. Exploratory Analysis Overview*

Topic Clustering

Topic clustering is the process of dividing the data into groups according to the topics they contain after preprocessing the texts. In order to prepare the data, we do a preprocessing step using the NTLK library for text mining. Preprocessing helps to clean the data for better performance of the subsequent algorithms and techniques. In our case, we remove all the tweets' mentions, hashtags, links, and emoticons. We also filter out all the digits and symbols that are not alphabetic characters and the retweet heading from the tweet (RT). Apart from all extracting mentioned above, we lowercase and tokenize to group the different representations of the same word in different genres or verbs in a distinct tense. Finally, we additionally pull out stopwords. To clean the text correctly to extract meaningful topics, we add personalized stopwords (see in **Table 7.3**) to the premade list (months, everyday verbs, nouns that do not make any meaningful contribution, and specific words containing the universities name or abbreviations).

In order to get the topics of the tweets and classify them, we use the K-Means algorithm.
K-Means is a commonly used method in various domains such as science, healthcare, and finance (Shukla & Naganna, 2014). K-Means is an unsupervised clustering algorithm that forms non-overlapping groups by partitioning the data based on their characteristics. The number of clusters is determined through the k parameter. K determines the number of centroids and the centers of each cluster. The algorithm recalculates the centroids to minimize the distance between the cluster points and the centroid (Rejito, Atthariq & Abdullah, 2021; 2.3. Clustering, n.d.).

We select K-Means method because it is simple, scalable for large datasets, and generalizes correctly for clusters with different forms (*k-Means Advantages and Disadvantages| Clustering in Machine Learning| Google Developers*, 2021).
Although, one of the main limitations that this clustering method presents is the selection of the number of clusters. Many studies have pointed out how demanding it is to determine the ideal number for this parameter (Abbas, 2008). In order to address it, we use the knee detection method.

We use K-Means algorithm to discover the patterns and commonalities of our data. In order to cluster to get the topics, we first use Weighting Term Frequency Inverse Document Frequency (Tf-Idf) Vectorizer to process and transform the data.
TF-Idf is a method used to check for the connection between the different documents and the terms in each document. It weights the terms depending on the frequency of occurrence in the different documents, giving more value to uncommon words and less to ordinary ones. The method includes the term frequency, how many times a word appears in a single document, while inverse document frequency, the frequency of the term through the different documents (Rejito, Atthariq & Abdullah, 2021).

Our aim of using Tf-Idf is to filter or weigh the uncommonly used words across the documents to find the topics. Using the Tf-Idf, we can create the corpus from the cleaned tweets, vectorize it to make it interpretable for the algorithm, and normalize it for better results. The minimum number of words selected is two, and the maximum document frequency is 38 - 40%. Selecting a lower maximum, we ensure that we get the most distinctive and relevant words in the documents to cluster them into the topics.

In order to obtain the optimal number of clusters for K-Means, we use the knee or kneedle method (Satopaa et al., 2011). It is a commonly used method that locates the optimal number of clusters. This number of optimal clusters is selected as the k parameter in K-Means.

Finally, to plot the result, we use the Principal Component Analysis (PCA) algorithm, a method used for dimensionality reduction. We use it to plot the clusters in two dimensions. Based on the most significant words in each cluster, we manually review the number of clusters obtained, regroup them into larger groups, and label them according to the theme. Then, each tweet is classified into the selected clusters based on the words it contains. This process is iterated for the three university groups (high, middle, and middle-low). The review and assessment of the clusters is done by the author of the study and two extra individuals, non-experts in the data science field. Both revisions are combined to form the regrouped clusters.

From the obtained topic clusters, we choose to focus our research on the associated branch of science-related knowledge, and so do their posts. Due to our purpose being to examine the scientific research content in universities' social media accounts, we only zoom in on the clusters in which the topic is related to the scientific research domain. In order words, we will proceed with the analysis with those tweets clustered in areas of scientific study.

Hence, the network analysis is based on the science and research group selected from the topic clustering. We present two networks: retweets and replies. These networks help us to understand the network's structure. Besides, the purpose is to explore from the selected tweets how these accounts communicate and what kind of account they interact with in science research.

Network analysis

As previously stated, we filter out all the obtained clusters except the science or research-related ones. For the network analysis, we use the R library *network*. It is designed for network analysis tasks. We use this library to plot two directed graphs: retweets and replies.

For each university account cluster (HU, MU, and LU), we do the following process:
To create the retweet network, we use two criteria: filtering tweets that contain the retweet header in the tweet (RT @acount) or filtering the referenced tweets that contain referenced type "retweeted". Both selections are combined after checking for duplicates. In order to extract the nodes and edges for the network, we set the user_name account as a source (the user that has retweeted the tweet) and the mentioned or referenced account (the original creator of the tweet) as the target. These edge tables (one for each university cluster) contain all the retweet interactions, the university cluster (HU, MU, and LU), and the frequency of retweets to these accounts.
We combine all the source and target accounts from the edge list to obtain the node table and remove the duplicates. In **Table 7.4** and **Table 7.5**, we present node and edge table examples.

The second network contains the replies. The replies are answers to a previously published tweet. To create the replies network, we apply one criterion: filtering the referenced tweets that contain referenced type "replied_to". If the tweets are replies, we extract them from the replies tweets dataset to get the target from the tweet. To obtain the user's screen name, we merge the users and replies table (includes all the replies with the user names and the tweet id) with the tweets table. We set the *user_name* account as a source (the user that has replied to the tweet) and the referenced account (the creator of the tweet that has been replied) as the target. The node and edge tables have the same structure as the retweet tables.

By selecting retweeted and replied tweets, we can explore what selected accounts reply to or retweet and their frequency of interaction.
Using the replies, we can detect those users that probably mentioned the official university accounts and got a response from the universities. Or to which account do these universities reply. We can examine if these accounts are using this social media to only broadcast information or if they are trying to interact with their audiences by providing direct responses. Also, we can check what type of account they reply to, such as institutional or personal accounts. While using the retweets, we can identify which accounts the university decides to repost and indirectly associate with themselves.

In summary, exploring the networks of retweets and replies aims to context our research and understand level of interaction with the audience and how often these accounts engage with other users.

3.1.2. Selected data exploration results

This section provides the results of the exploratory analysis, which includes the topic clustering and network analysis.

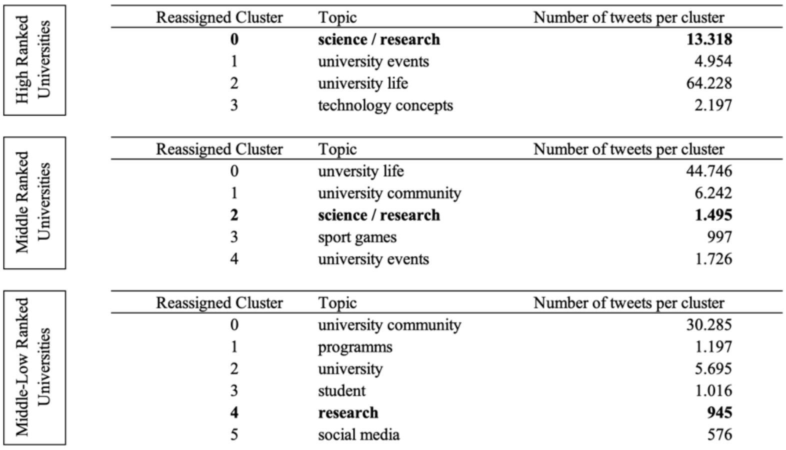
Topic clustering

We perform topic clustering for the groups created: high, middle, and middle-low university accounts (HU, MU, and LU). Using the knee method to find the optimal number of clusters, we find that the suitable number of clusters is eight for HU, nine for MU, and ten for LU. In **Table 7.7,** we present a summary table of the topic clustering and the selected labels, and in **Figure 12**, **Figure 13**, and **Figure 14**, we show the knee detection graph for each cluster.

To look closer into each university cluster, the first group (HU) includes three top-ranked universities: Harvard, Stanford, and MIT. In order to assess and reorganize the clusters, we plot the most influential words from each cluster. See some of the relevant words in each cluster in **Table 7.6**. Some of the clusters found are similar, so we reduce the initial clusters to four. The main topics included are science/research, university events, campus life, and technology.

Science and research contain terminology related to science, labs, and science research. At the same time, university events cover common words for graduation programs, ceremonies, and celebrations.

Campus life exposes terms more familiar with the day-to-day life of the students from the university, including programs, lectures, and related words for learning. The technology topic is a fuzzy cluster that includes a range of words connected to technologies.



High Ranked Universities

Reassigned Cluster	Topic	Number of tweets per cluster
0	**science / research**	**13.318**
1	university events	4.954
2	university life	64.228
3	technology concepts	2.197

Middle Ranked Universities

Reassigned Cluster	Topic	Number of tweets per cluster
0	unversity life	44.746
1	university community	6.242
2	**science / research**	**1.495**
3	sport games	997
4	university events	1.726

Middle-Low Ranked Universities

Reassigned Cluster	Topic	Number of tweets per cluster
0	university community	30.285
1	programms	1.197
2	university	5.695
3	student	1.016
4	**research**	**945**
5	social media	576

Table 3.4. *Distribution of tweets per clusters*

The cluster we analyze is the science/research one presented in **Table 3.4**, which aligns with our objectives and purpose. This cluster contains 13.318 tweets.

The second group (MU) includes three middle-ranked universities: Utrecht University and the Science Faculty, Michigan, and Arizona State Universities. Although the optimal number selected is nine, we review the clusters and group them. The final number of clusters is five. Among the different topics, we can highlight similar topics to the other HU institutions, such as university life, science and research, and university events.

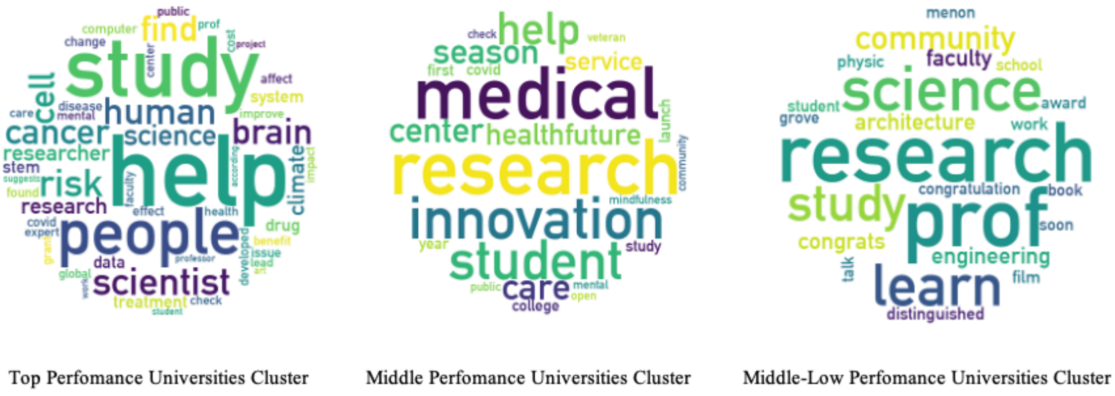
The most specific topics we find are the university community and sports games. In the university community, we find a sense of belonging associated with positive emotions and a knitted cultural group. At the same time, the sports game topics include the university's pets and several sports. This group contains 1.495 tweets selected for science/research context.

The last group (LU) includes three middle-low ranked universities: Oregon and Kansas State Universities and City College of NY. In this particular cluster, we obtain ten topics as the ideal k. After checking the clusters, we reduce them to six groups. Again, some of these topics are comparable to the other groups (HU and MU). For instance, the university community, student research, and university life. In this case, we find two distinct topics: social media and university programs. The first includes social networks and their usage in the university promotion life topics. On the contrary, university programs include various disciplines that universities offer to students. We select the related research/science one from these six clusters incorporating 945 tweets.

Science / Research Topic

We present below three different word clouds containing some of the words clustered as science or research domain. We can spot a wide variety of related words; some of them are common in all distinctive clusters.

One of the limitations of using the topic clustering for the tweets is that we drastically reduce the volume of data for the analysis phase. As presented in **Table 3.4**, for the MU and LU university clusters, the number of tweets related to the science and research domain is lower compared to the HU universities.



Top Perfomance Universities Cluster

Middle Perfomance Universities Cluster

Middle-Low Perfomance Universities Cluster

Figure 3. Science / Research clusters wordclouds

Network Analysis[1]

After selecting the scienfic cluster, we plot two separate networks for better knowledge of the data. We present a network for retweets and an additional one for replies.

Using the retweets network[2], we can briefly observe which ones are the accounts that they choose to get implicated. We use different node colors for the different groups (high, middle, and middle-low universities). Besides, we can see the labels indicate the node account and the degree of centrality each node has according to the size. The degree of centrality is calculated based on the total amount of links connected to the node.

The network structure is a scale-free network. We can spot that the universities' accounts are large hubs connected to different accounts. While High Ranked Universities and Michigan State Universities are connected, the rest of the universities have their own retweet clusters, forming individual components and not interacting with any other account from the present study. Therefore, we can see seven components in the network. Besides, HU cluster form a community as they are connected to each other.

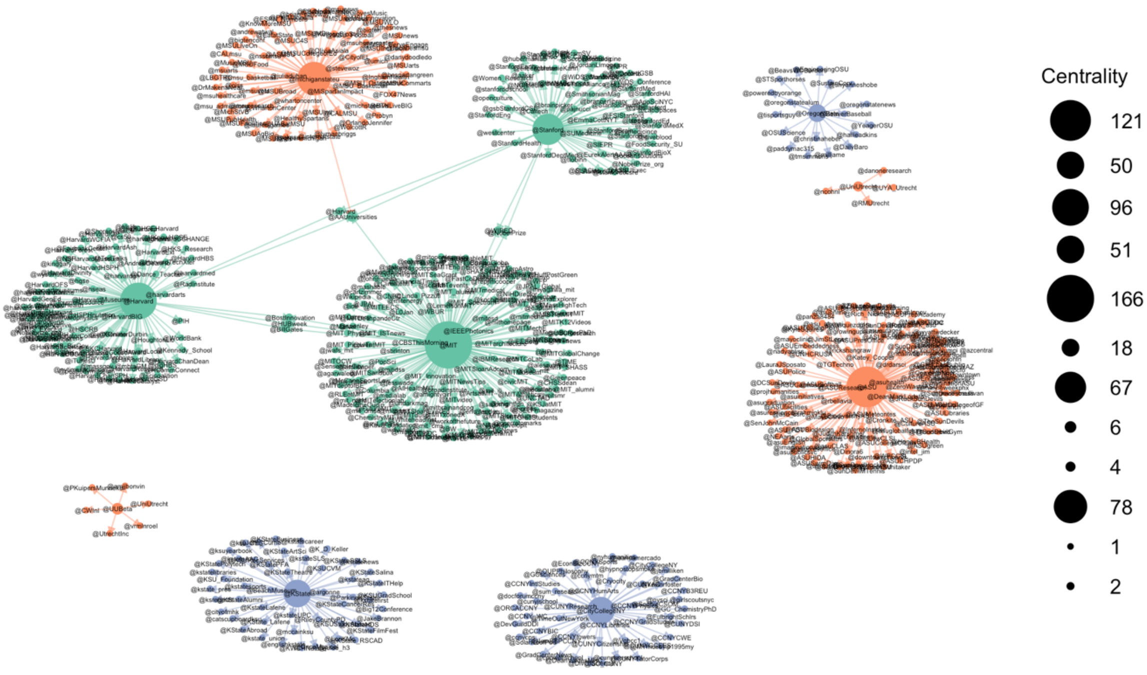


Figure 4. *Retweet Network*

HU contain a more significant amount of retweets than smaller ones (MU and LU). As we can spot in the **Figure 4.** Retweet Network[1], some of the retweeted accounts are secondary accounts of the universities, such as @StanfordHealth, @CCNYLibraries, or @KStateAbroad. They also engage with other representative accounts such as @Kennedy_School, @WorldBank, or @openculture. In the center of the retweets graph (**Figure 4**), we can observe

[1] Due to the interpretability of the graph, the frequency of interaction is not present in the graphs.

[2] In the **7.2** Attachments, it is provided the full-size page of **Figure 4** the retweet network, for better interpretability.

an interaction between some shared accounts. For instance, Stanford and MIT retweet content posted by @WIRED and @NobelPrize. @WIRED is a science/technology account where they post articles from realized studies. In contrast, @NobelPrize is the official account for the Nobel prize awards, where they post new updates and informative content.

Besides, Harvard and MIT retweet tweets from the same accounts such as @BostInnovation, and @BillGates. Bost Innovation is an old account from @BostInno, a company that gives visibility to local startups, technology, and innovation (BostInno [@BostInno], n.d). On the contrary, Bill Gates is a known software engineer and entrepreneur renowned for creating Microsoft company.

Lastly, it is interesting to highlight the retweet connection between Harvard, MIT, Stanford, and Michigan State University with the @AAUniverstities account. The American Association of Universities is the official account for an organization formed by 65 research higher institutions with a common aim for research innovation (AAU [@AAUniversities], n.d). This retweet network helps to context the universities' social networks in the science/research topics.

Meanwhile, the replies network only contains nine individual components not connected to each other. The node size also represents the degree of centrality of the nodes. The network structure is also a free-scale network formed by individual components.

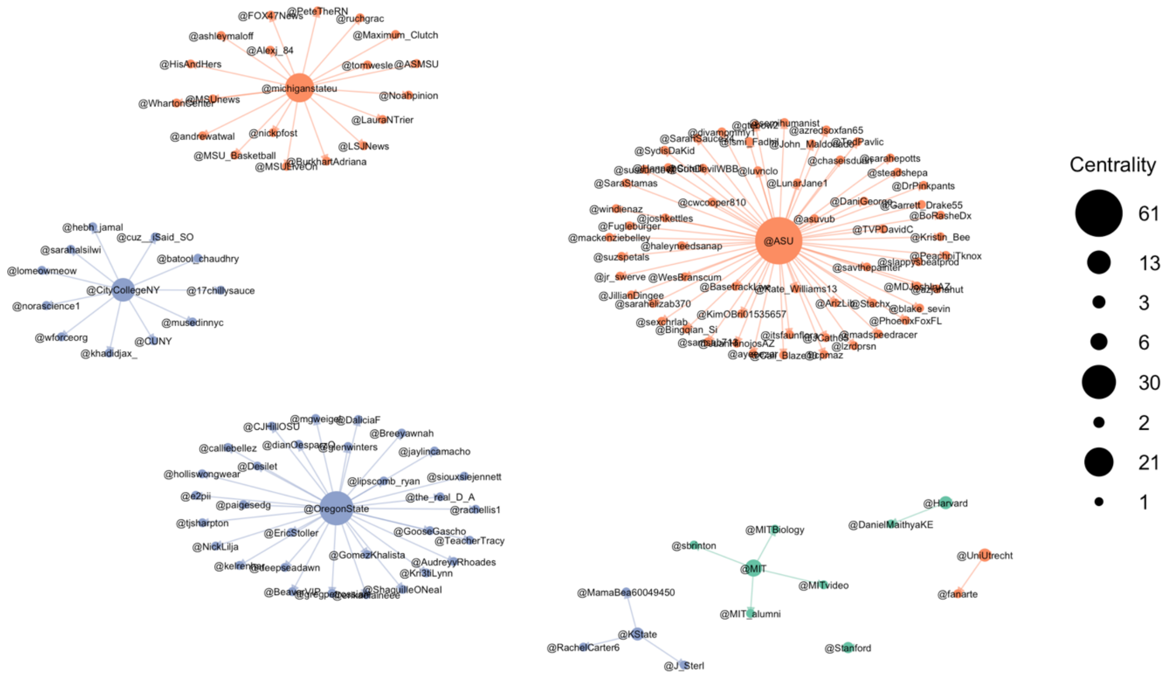


Figure 5. Replies Network

Besides, in the reply Network (**Figure 5),** we can observe which accounts interact more with the users except for Utrecht Faculty of Science, that does not appear in the graph. We can spot that while most of the top universities (HU) do not have many replies to other users, MU and LU like Arizona State or Oregon State Universities are encouraging more interaction and, therefore, a two-way communications instead of only broadcasting information. We can spot that mainly these users that these official accounts replies are mainly individual accounts rather than institutions, except MIT, which mainly replies to secondary accounts. These interactions can be from students or individuals requesting information from the higher institution. Some examples are provided in section **7.2.1.Tweets Examples**.

The main insights from the exploratory analysis are visualizing the free-scale network structures formed in both cases by individual components and in the retweet network also by different components' interaction. The retweet network offers interaction between accounts, and the main retweeted accounts are related to secondary accounts from the universities or relevant institutions or public figures related to the research and science university domain. In contrast, the reply network contains more replies from MU and LU than the HU ones. The replied accounts are mainly secondary accounts from the institution or Twitter personal accounts.

In brief, we can spot that the retweet network contains more interactions than the replies. It can be due to it being a more straightforward form of interaction or can also be due to the data sampling chosen. We need to acknowledge that both networks are used for qualitative analysis with the purpose of giving more context to the study. The science/research context restricts the samples used for both networks. Consequently, the rt network includes approximately 2.400 tweets and the replies network 275 tweets.

3.2. Data Analysis

3.2.1. Data Preparation and Methods

After the exploratory phase, we proceed to process the data for analysis (**Figure 6**).

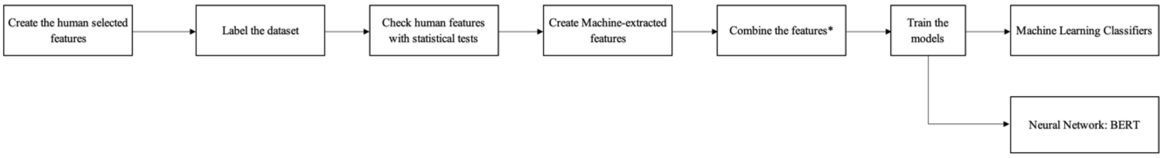

Create the human selected features
Label the dataset
Check human features with statistical tests
Create Machine-extracted features
Combine the features*
Train the models
Machine Learning Classifiers
Neural Network: BERT

***Figure 6**. Analysis Overview*

The analysis aims to predict the type of engagement (*like*, *retweet*, or *reply*) for each university cluster (HU, MU, and LU) using the post from Twitter. Also, to find what are the features that contribute to the prediction, we use two types of features: human-selected and large-dimensional features that we obtain from the texts of the tweets (machine-extracted features). The features and methods used are based on a previous study (Dai & Wang, 2021) that incorporates human and machine-extracted features from three Chinese telecom state-owned companies' posts on Weibo. The present research is based on some of the foundations in the previous research.

We can divide the analysis into two parts: creation, and process of the features, and training of the models.

In the first part, the main overview of the analysis is to create the human features, use statistical tests to check the dependence with the outcome variable (type of engagement), create the machine-extracted features, and combine the two types of features (human and machine-extracted). We combined both types of features based on the Dai & Wang's (2021) results, where they state that combining the features increase the accuracy of the prediction (type of engagement).

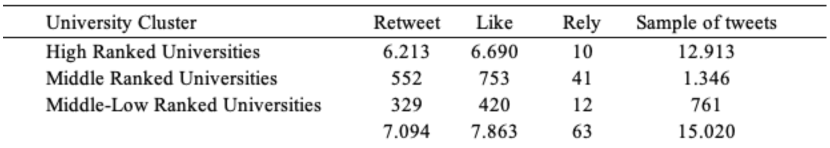
In the second part, we train four selected models (Logistic Regression, Random Forest, LightGBM, and multimodal transformers using BERT) to classify the tweets into the most probable type of engagement. In order to train the classifiers using supervised methods, we need first to label the dataset.

Human Selected Features

We select thirty-six human variables that contain information about the tweet content, the account, or the interaction of the tweet. The human-selected features are mainly binary/ dummy variables, except for the length variable, which is a numeric variable containing the tweet's length.

In order to create them, we check for duplicate tweets and the types of the variables, and we add the tweet's length as a new variable. We process the text, delete emojis, URLs, mentions, hashtags, and stopwords, tokenize the words, and create the potential predictor variables. These thirty-six selected variables are presented in **Table 3.5**. Most of the main human-selected variables used are based on the Dai & Wang's (2021) study, such as if it contains question marks, hashtags, or the time of the day. Others are created for our research, such as if the content is broadcasted or the universities' accounts.

After preparing the dummy variables, we use the higher type of engagement as the label for supervised ML algorithms. It is a multi-class classification problem where the predicted classes are *like*, *retweet* or *reply*. If the tweet selected has no engagement (zero *likes*, *retweets*, or *replies*), we discard it, as we are interested in predicting the engagement. **Table 3.5** contains the label distribution of the sample.



University Cluster	Retweet	Like	Rely	Sample of tweets
High Ranked Universities	6.213	6.690	10	12.913
Middle Ranked Universities	552	753	41	1.346
Middle-Low Ranked Universities	329	420	12	761
	7.094	7.863	63	15.020

Table 3.5.1.3.6 *Distribution of classes in the sample*

Subsequently, we use Chi-Square and ANOVA tests to check the statistical dependence between the variables.

Chi-Square Test is a statistical test used to examine if there is any correlation between two binary variables. It uses the null hypothesis to check if the variables are independent (we fail to reject the null hypothesis) or dependent (we reject the null hypothesis). Similarly, ANOVA (Analysis of Variance) to test the correlation between two variables that are not categorical (numeric). After, obtaining all the selected human features, we process the large-dimensional features that we obtain from the texts of the tweets.

Machine-extracted features

Machine-extracted features are large-dimensional features that we obtain from the texts of the tweets. The machine-extracted features are created differently for the ML classifiers and the network analysis model. We use a bag of words for the ML classifiers to create a representative matrix that contains all the words in the corpus. The minimum term frequency selected is five. Once we obtain the matrix, we combine it with the human features to obtain the final representation for training the ML models.

Alternatively, we use multimodal transformers for the neural networks that use BERT to process the text features. These multimodal transformers combine all features (text, categorical and numerical) to train the model for the classification task.

Machine Learning and Deep Learning Models

The models selected for ML classifiers are Logistic Regression (LR), Random Forest (RF), and LightGBM (LGBM). LR is a statistical method conventionally used in statistics, ML, and data science. It is more commonly used as a binary classifier than a multiclass classifier. However, we find some studies that use them for classifying categorical data, such as Dai & Wang (2021) and Pranckevicius & Marcinkevicius (2016).

RF is a randomized decision tree algorithm (*1.11. Ensemble methods*, 2022). The decision tree model is based on the concept of the wisdom of crowds (Yiu, 2019). The idea behind it is that multiple individual trees combined perform better than a single one.

Light Gradient Boosting Framework is also a decision tree algorithm based on the wisdom of crowds (*Welcome to LightGBM's Documentation!- LightGBM 3.3.2.99 Documentation*, 2022). Remarkably, the advantages of this method for our case are rapid training speed and superior accuracy.

Whereas for the deep learning model (neural network), we use multimodal transformers (**Figure 7**). The multimodal-transformers is a toolkit that allows choosing between different pre-trained transformers for the text features. Then, it combines all features (text, categorical and numerical) using neural networks to perform the selected task (Gu, n.d.). It "integrates well with Hugging Face's existing API, such as tokenization and the model hub, which allows easy download of different pre-trained models" (Gu & Budhkar, 2021, p.69). Between the available pre-trained models, it includes the BERT model.

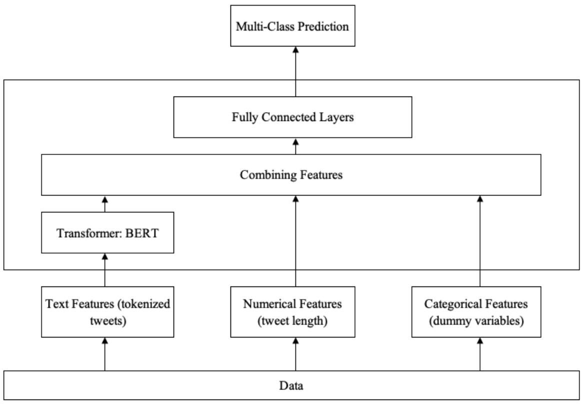

Multi-Class Prediction
Fully Connected Layers
Combining Features
Transformer: BERT
Text Features (tokenized tweets)
Numerical Features (tweet length)
Categorical Features (dummy variables)
Data

***Figure** 7. Multimodal Toolkit Structure*[3]

[3] Adapted from Gu & Budhkar (2021)

Bidirectional Encoder Representations from Transformers (BERT) model is a pre-trained model used for NLP (Cosimo, 2022). In our case, we use it as a multi-nomial classifier. We use the transformer pre-trained model to transform the tweet's text into high-dimensional vectors. These high-dimensional vectors contribute to predicting the type of engagement.

For the model training, we split the data into training (70%), validation (15%), and testing (15%) before building the models. For the first three classifiers, we knit together the human and large dimension features, whereas, for the multimodal transformers, we introduce them separately. Overall, we proceed with the analysis repeatedly for our university clusters (HU, MU, and LU) to find the differences and similarities among groups.

3.3. Translation of the research question to a data science question

Based on the research gap provided in the literature review, the present study focuses on the social media platforms in science communication and how high educational institutions use social media in the scientific research field context. Therefore, the research question of our study focuses on two main points: Is it possible to predict the type of user engagement based on posts published in their Twitter feeds? What are the features that contribute to predicting this engagement type?

As a result, the translation of the research questions into a data science question is the following. First, can ML and deep learning algorithms (neural networks) be used to predict tweets that include science research content? The data from the science-research clustered is divided by the most probable type of engagement: retweets, likes, or replies. After, three ML classifiers are trained: LR, RF, and LGBM. Furthermore, a pre-trained multimodal transformer that integrates fine-tuned BERT is trained. The metrics used to assess these predictions are accuracy, precision, recall, and f1-score.

In the second place, what are the features that correlate with the outcome of the prediction? In order to get these features, human-selected and machine features from the tweet's text are used.

In the third place, a comparison of the different university clustered accounts (HU, MU, and LU) is provided, which includes an exploratory analysis and the prediction results to interpret the research outcome. In particular, we use topic clustering, network analysis, and the results showing the similarities and dissimilarities.

3.4. Ethical and legal considerations of the data

We used the Twitter API to collect the data from the ten universities. This data includes all published tweets in their feeds and other practical information to be analyzed, such as the day and time of the publication creation, the engagement metrics or the retweets, and their replies from their feeds.

It is widely extended to use this Twitter data for research end. However, using this data for research purposes has some ethical considerations. As Williams, Burnap & Sloan (2017) state in their study, every one of us accepts the terms and conditions before creating a Twitter account. When we accept these conditions, we are also giving away all of our personal information published in our account to anyone with access to a Twitter developer account. Or an internet connection (Ahmed, Bath & Demartini, 2017). This issue raises an ethical concern as users are unaware that external parties use their data for different purposes (research, market segmentation). According to their research (Williams, Burnap & Sloan, 2017), 80% of the individuals that participated in their survey expressed that they thought the platform would ask for their consent in case someone used their data. Whereas 90% of them supposed their data would be used anonymously. This fact reflects a lack of knowledge from users on how their data is being treated and by whom.
Although, asking for informed consent is indeed challenging for researchers to request from every user due to the large amounts of data used (Ahmed, 2017).

Besides, we ensure ethical handling of the data as the data is anonymized, demographic data is not included, and the data is used to provide aggregate results, not a qualitative analysis, except in the exploratory phase of the analysis.

The only information we post without anonymity is the screen name of the mentioned user accounts and public tweets to give some examples in the exploratory phase to add more understanding of the data. This information is used in the exploratory analysis to give context to the data and show how universities communicate through Twitter. In other words, it is used to get a better understanding of how university accounts provide information and to whom, as well as to compare and contrast how the different types of universities (according to academic performance) use their Twitter accounts.

Lastly, we do not manipulate or alter the tweets to follow Twitter's User Development Policy as it states to publish the tweets without changing them. We only use preprocessing techniques to prepare the text for the algorithms, but these tweets are not explicitly used to be quoted or mentioned in the analysis.

4. Results

4.1. Selected analysis results

Human Explanatory Variables

After the data analysis, from the thirty-six created explanatory variables, we check which ones correlate with the type of engagement for each university cluster: High Ranked Universities (HU), Middle Ranked Universities (MU), and Middle-Low Ranked Universities (LU). The complete description table of the explanatory variables can be found in **Table 7.2**.

After the statistical tests[4] and checking the variables that correlate with the type of engagement, the results show which dependent variables help to contribute to the type of engagement prediction for each university cluster (HU, MU, and LU). See **Table 7.8.** Selected Explanatory variables for the complete table that contains all the dependent variables.

The first group (HU) includes a total of nineteen out of thirty-six explanatory variables. Similarly, the second group (MU) includes twenty explanatory variables for predicting the type of most probable engagement. The last set (LU) includes the lowest number of predictor variables for the prediction, with fourteen.

Figure 8 shows the percentages of use of the explanatory variables. Of the total amount of human selected variables (n = 36), nine of thirty-six (25%) are used for predicting the engagement type for all university clusters (HU, MU, and LU). Five variables (13,9%) contribute to the prediction in two out of three university clusters, sixteen variables (44,4%) for one university cluster, and six (16,7%) are not included in any cluster because they do not correlate with the target variable.

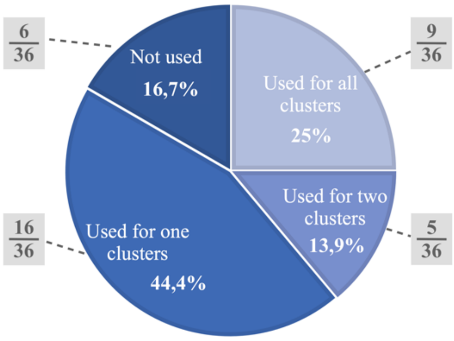


Figure 8. *Percentage of use of Explanatory variables*

The following variables (16,7%) are not used because they do not have any dependence on the type of engagement for any cluster (HU, MU, and LU). This group includes *has_emoji, has_underlines, user_is_uniutrecht, is_morning, is_wednesday,* and *is_thursday*.

Alternatively, it is relevant to highlight that some of these selected variables contribute to predicting the three university clusters (HU, MU, and LU). This group (25%) includes the features: *has_hashtag, is_retweet, has_link, has_exclamation_mark, is_reply, has_images, is_afternoon,* and *length*.

The rest of the variables are explained by university clusters:
For the High Ranked Universities (HU), other features that help to predict the type of engagement are the accounts variables for *user_is_stanford*, *user_is_harvard*, and *user_is_mit*. For the tweets' content *has_mention*, *has_question_mark*, and *broadcast_content*. Moreover, including attributes attached such as *has_gif* or *has_video* features, or if the tweet was

[4] Chi-Square and ANOVA tests measure the relation between the variables using p-value.

published was related to *is_night* or *is_monday*. See the full results of the statistical test in **Table 7.9** and **Table 7.10**.

For the Middle Ranked Universities (MU), more temporal features contribute to the prediction, such as *is_tueday, is_saturday, is_sunday, is_weekend,* and *is_night*. Other attributes related to the tweet's content and the attached media also contribute to predicting the type of interaction, such as *has_video*, *has_mention*, *has_question_mark*, and *broadcast_content.* For the accounts variables, *user_is_uubeta*, *user_is_asu*, and *user_is_michiganstateu* are correlated with the prediction of the type of engagement. Notably, *user_is_uniutrecht* does not correlate with the interaction type prediction. It is an independent variable as the p-value is 2,958E-01 for the Chi-Square test. See the results of the statistical test in **Table 7.11** and **Table 7.12**.

Lastly, for Middle-Low Ranked Universities (LU), the temporal content almost does not correlate with the prediction except for *is_thursday*. From the different accounts, *user_is_oregonstate, user_is_collegeny,* and *user_is_kstate* are dependent on the type of interaction. Lastly, the feature *has_question_mark* is also related to this cluster's prediction. See the full results of the statistical test in **Table 7.13** and **Table 7.14**.

Classifiers Performance

After testing the explained variables and selecting those that have a correlation with the type of engagement for each university cluster (HU, MU, and LU) (see **Table 7.8**), the trained models' performance is shown. These human explanatory variables, in combination with the machine-extracted variables[5], are used to train four different models (LR, RF, LGBM, and multimodal transformers using BERT) for each cluster (HU, MU, and LU). See the Analysis Overview scheme of the process in **Figure 6**.

In order to assess the performance of the methods, precision, recall, and f1 score metrics are used. In particular, average macro metrics are selected because they do not consider the imbalances in the dataset. All classes are equally weighted for calculating the average. On the contrary, weighted avg metrics are more sensible for imbalanced data because they take into account the relative contribution instead of the equal contribution. The Macro metrics consider all classes equally relevant, which is the present case. The results aim to provide an overview based on predicting the type of engagement where all classes (like, retweet, and reply) are uniformly relevant, and using weighted avg would bring misleading conclusions because the scores are higher than with the macro avg metrics.

The complete results of Macro and Weighted metrics are in **Table 7.15** and the confusion matrix in **Table 7.16**. The results show that macro avg metrics are more restrictive than weighted avg ones. Besides, the table shows a significant imbalance in one of the classes: reply. This inequality negatively affects the classifier's performance. In order to prevent it, some parameters are modified to lower the imbalance, such as selecting multinomial in the multiclass

[5] large-dimensional features that we obtain from the texts of the tweets

or balanced in the weight of classes. Moreover, HU dataset has the larger imbalance of classes, see the distributions in **Table 3.5**. Even though the imbalance is still present, the analysis and the data can give insights.

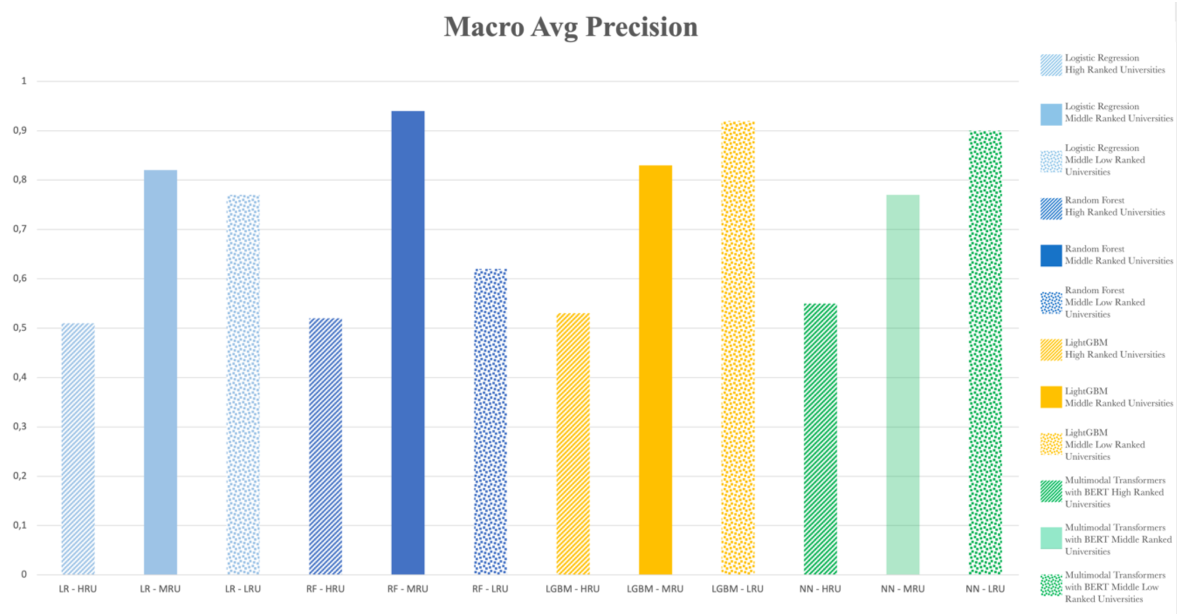


Figure 9. Results based on Macro Avg Precision

According to Macro Avg Precision Metric, **Figure 9** shows that the methods that work better for each cluster are Multimodals Transformers for the High Ranked Universities, Random Forest for the Middle Ranked Universities, and LightGBM for Middle-Low Ranked Universities. Whereas **Figure 10** shows that the best suits High Ranked Universities is MultiModals Transformers. For Middle Ranked Universities is LightGBM, and for the Middle-Low Ranked Universities is Logistic Regression.

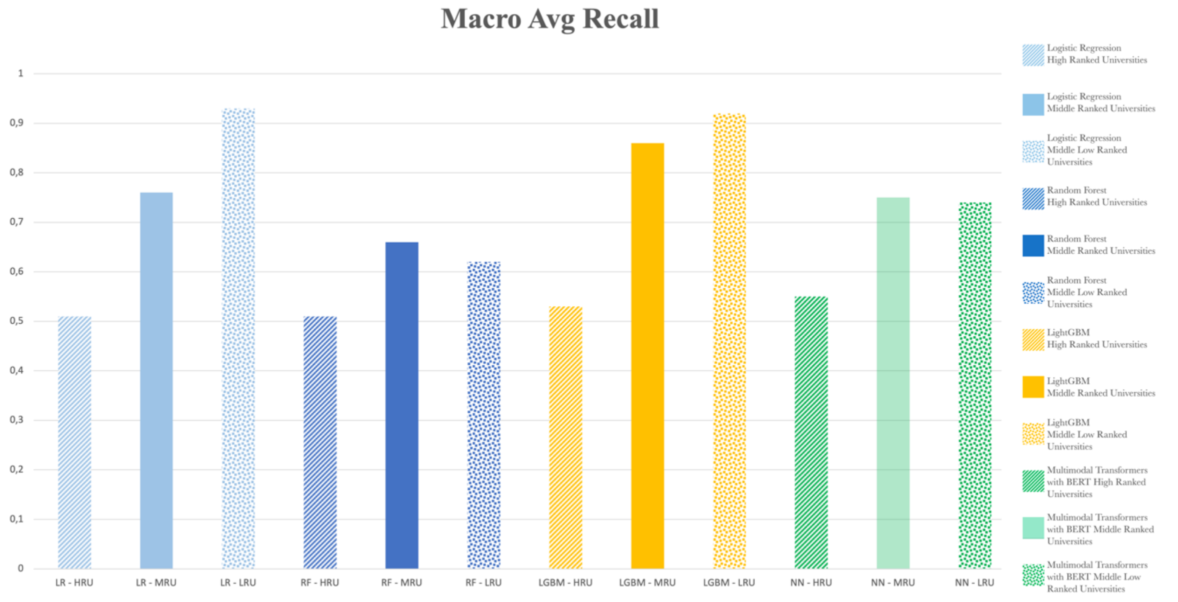


Figure 10. Results based on Macro Avg Recall

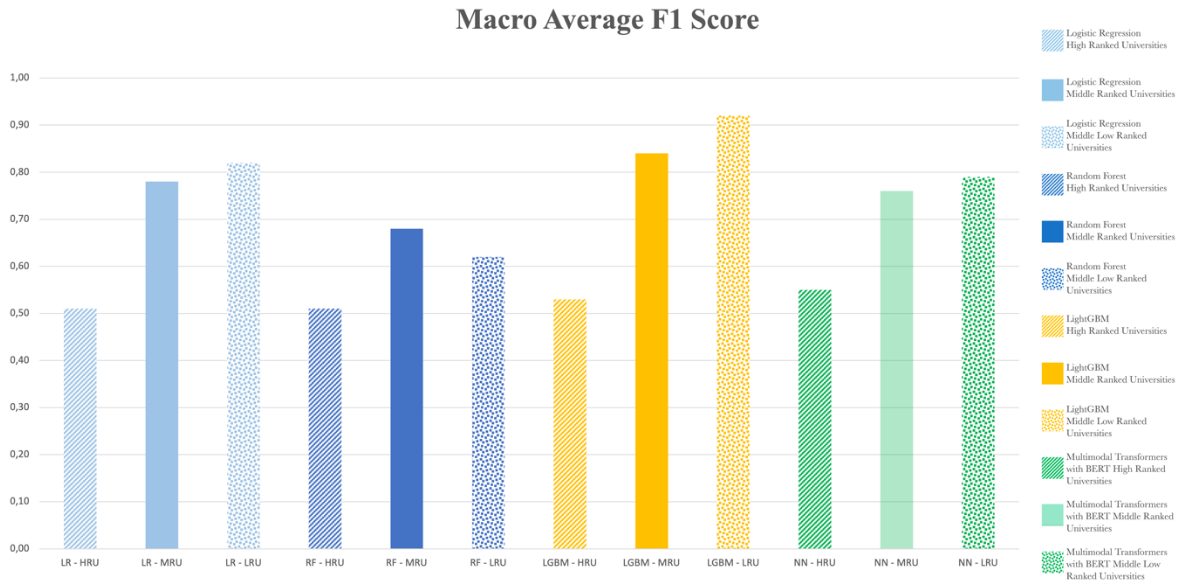


Figure 11. *Results based on F1 Score*

Finally, Based on **Figure 11** presents that the Middle-Low Ranked Universities have the highest Macro F1 Score of 0.92 using the LightGBM classifier, as well as the Middle Ranked Universities with 0.84 scores, and for High Ranked Universities Multimodal Transformers has the higher score with 0.55. The last group has a more imbalanced dataset due to the sample used, and as a result, it is more complex for the different classifiers to train the model and get accurate predictions.

Overall, the method that works better is Light Gradient Boosting Machine, while for more accentuated imbalanced classes, the multimodal transformers with BERT are more suitable.

5. Discussion and Conclusion

The present study aims to predict the most probable type of engagement based on the human and machine-extracted features from ten universities' Twitter feeds. In particular, the study focuses on the science and research domain. Apart from predicting the type of interaction, the research aims to show what features help to predict engagement across the university clusters (High, Middle, and Middle-Low Ranked Universities). Moreover, the objectives include an overview of the similarities and differences of those features across the universities' clusters.

The research uses four multi-class models to predict the engagement (like, retweet, or reply) of the clusters (HU, MU, and LU). The models are Logistic Regression, Random Forest, LightGBM, and multimodal transformers using BERT.
The metrics (based on precision, recall, and f1-score) show the performance of the models for each cluster (HU, MU, and LU). Based on these metrics, the findings confirm that the most engagement type can be predicted based on human and machine-extracted features.
Moreover, the results also show which human features correlate to this prediction for each university cluster. The analysis of these features includes statistical tests such as Chi-Square and ANOVA.

Furthermore, the present study objectives incorporate an exploratory data analysis, including topic clustering and network analysis. The exploratory analysis aims to provide context to the data and explore how universities interact with the platform and audience.

The results from the topic clustering show that the main clusters contain information related to students, university life, science/research, and the university community. Also, the topic clustering provides results that large volumes of tweets are focused mainly on students and not on the science and research field.

From the obtained clusters, the last one is selected because the focus of the study is to provide some knowledge about how universities use social media platforms for research purposes. Therefore, the exploratory analysis ends by providing two network analyses based on the science and research cluster. It includes one network for retweets and another for replies.
The retweet network shows how universities inside the HU cluster interact with each other, while MU and LU do not have that much interaction with other components.
Also, the network shows that many accounts retweeted are mainly secondary accounts from the university or institutions and public figures related to the scientific field.

On the contrary, the replies network shows less interaction. It presents that the accounts replied are mainly secondary accounts from the institution or individual Twitter accounts (such as individuals requesting information). Besides, this network shows that MU and LU have a more significant number of replied tweets than HU. In contrast, HU provides more retweets than the other two clusters (MU and LU).

These results from the exploratory analysis can also partly explain the sample used for the data analysis. For instance, results show that retweets are more common ways of engagement than replies, especially for Higher Ranked Universities. The lower number of direct interactions (replies) produces an imbalance in the sample used. Consequently, it reflects in the classifiers' performance. The findings from the data analysis show that the best model for each cluster is Multimodal transformers for Higher Ranked Universities, and LGBM for Middle and Middle-Low Ranked universities. The university cluster with the worst results in the performance metrics is the HU cluster, and the best one is the MU cluster.

Apart from the classifier's training, a selection of human variables that contribute to predicting the most probable type of engagement are tested. The results present the human features that correlate with the outcome variable using statistical tests.
The results show that the MU is the cluster that uses more explanatory variables (twenty out of thirty-six) to predict the engagement type. In contrast, the LU uses less amount of variables (fourteen out of thirty-six), and the HU uses nineteen.
The findings show that if the tweet includes hashtags, mentions, or it is a retweet, these are some examples of human features that contribute to the predictions of the type of engagement. The length of the tweets is also a great predictor.

In comparison, some specific features related to the time and day of the tweets' publication vary across the university clusters (HU, MU, and LU). Similarly, the features related to accounts also differ between the universities' groups. See the complete results in **Table 7.8**.

Besides, machine-extracted features combined for predictions also contribute to increasing the classifiers' predictions, as previous research acknowledges (Dai & Wang, 2021). These machine-extracted features are based on the tweet's text. In combination, both features help to predict the most probable type of engagement.

In addition, the research results support existing studies based on the lack of science or research content from the universities' accounts, as they use mainly the platforms as a recruiting tool for students and less to create and engage with the science and research communities (Davis, 2014). Moreover, it also aligned with previous studies that reflect a lack of interaction between these accounts and their audiences because they use these social media to broadcast content instead of establishing a connection with their followers (Linvill et al., 2012; Veletsianos et al., 2017). Additionally, lined up with De Melo Maricato & de Castro Manso (2022) points out an untapped potential for universities' official accounts to promote their research as they have more outreach and visibility than researchers' accounts. It can help to promote scientific research and position the universities as thought leaders in the community.

The present study aims to create knowledge in this area based on the research gap and contribute to the literacy of the universities' social media for research purposes.

Therefore the selected research questions are two: Is it possible to predict the type of user engagement based on posts published in their Twitter feeds? What are the features that contribute to predicting this engagement type?

As a result, the translation of the research questions into a data science question is the following. Is it possible to predict the university's most probable type of engagement from the tweets using machine and deep learning algorithms? What are the variables that contribute to this prediction?

Using ten university accounts clustered into (HU, MU, and LU), the study provides an exploratory analysis and a data analysis to train different models and confirm that it is possible to predict the type of engagement based on selected features. Those features are human-selected features and machine-extracted features. An overview of the features that contribute to the prediction for each university cluster (HU, MU, and LU) is provided using a statistical test. As a result, the research and data science questions are answered through this study.

Predicting the type of engagement contributes to creating and building networks in selected communities in many ways. Hashtags prediction can be used to strengthen connections using topics that audiences are interested in and increase the visibility of the content published. In comparison, replies can improve the accessibility to the institutions, which lets to customer satisfaction. Connected and interested users contribute to the level of influence and deepen the connection with the users. In addition, likes contribute as the most basic engaging metric to contextualize user interest. Therefore, each engagement metric leads to specific objectives that can be used according to the purpose.

Even though the research provides insights, it also presents several limitations. The first limitation is that the study does not differentiate the type of university (public, private) in the analysis. A distinguishable analysis based on the universities aim and values can be included in future research.

The second limitation is that specific metrics were not used because of the Twitter API limitations, such as if the tweet was promoted, the number of clicks a post got, or the direct messages between users and the official accounts.

The third limitation is that while several universities have more than one official account that provides more specific content to the audiences, the study scope considers only the universities' official main accounts. Therefore, each university's available accounts can be incorporated into future studies.

The fourth limitation of the study due to the time and scope is that the present study does not add the analysis of the links, images, or any other content attached to the tweet's content. It would be interesting to analyze all the multimedia content using convolution neuronal networks for future research to get attachment content. Also, the analysis of the links in the tweets to extract more information about the external references could be added in future research.

Moreover, the network analysis from the exploratory phase presents another research limitation. Due to Twitter API limitations, users who liked, retweeted, and replied to university posts could not be retrieved. The main reason is because of the volume of the dataset and the cap of the number of requests the limited Twitter API was able to retrieve.

Furthermore, the research only considers tweets posted in English. Therefore, local content is not selected for the analysis, especially for Utrecht University and the Faculty of Science, which includes content in the dutch language. Extending the number of languages and comparing engagement behavior across languages is an aim for future studies.

The last limitation is that the sampled use for the analysis contains a significant imbalance among one of the classes, affecting the classifiers' performance but still giving valuable insights. For future research, methods for creating synthetic data to balance the classes will be considered.

On top of that, the study presents some ethical, research, and practical implications. On one side, the present study's ethical implications ensure that the data is used for research purposes to provide generally aggregated results and insights. Moreover, the information provided aligns with Twitter's User Development Policy.

On the other side, the research implications help universities context and understand how they can write their tweets according to the objective they want to archive posting it. Universities aiming to increase their interaction with their audiences can use the present research to understand what features contribute to predicting the type of engagement and get the most probable engagement form based on the tweet published. Universities aiming to expand their social network probably will want to aim for retweet engagement. The reason is that the content gets published in other accounts, increasing the account's visibility and, therefore, the publication's exposure. Research Universities can use these implications to broadcast to larger audiences their studies and promote their Academic research as well as their brand awareness and positioning.

In general, by using their content according to their engagement aim and assessing the factors contributing to this outcome, they can gain visibility in their research and studies. As a result, it attracts more qualified scholars and motivates students to join its community which consequently helps to strengthen the brand image and helps to position itself as a referent in the educational sector, and, overall, enhances the universities reputation.

6. List of references

1.11. Ensemble methods. (n. d.). Scikit-Learn. Retrieved 12 June 2022, from https://scikit-learn.org/stable/modules/ensemble.html#forest

2.3. Clustering. (n.d.). Scikit-Learn. Retrieved 12 June 2022, from https://scikit-learn.org/stable/modules/clustering.html#k-means

AAU [@AAUniversities]. (n.d). AAU (@AAUniversities) [Twitter Account]. Twitter. https://twitter.com/AAUniversities

Abbas, O. A. (2008). Comparisons between data clustering algorithms. *International Arab Journal of Information Technology (IAJIT)*, 5(3).

Ahmed, W., Bath, P., & Demartini, G. (2017). Chapter 4: Using Twitter as a Data Source: An Overview of Ethical, Legal, and Methodological Challenges. *The Ethics Of Online Research*, 79-107. doi: 10.1108/s2398-601820180000002004

Amor, B., Vuik, S., Callahan, R., Darzi, A., Yaliraki, S., & Barahona, M. (2016). Community detection and role identification in directed networks: Understanding the Twitter network of the care.data debate. *Dynamic Networks And Cyber-Security,* 111-136. doi: 10.1142/9781786340757_0005

Assimakopoulos, C., Antoniadis, I., Kayas, O., & Dvizac, D. (2017). Effective social media marketing strategy: Facebook as an opportunity for universities. *International Journal Of Retail &Amp;* Distribution Management, 45(5), 532-549. doi: 10.1108/ijrdm-11-2016-0211

Attfield, S., Kazai, G., Lalmas, M., & Piwowarski, B. (2011). Towards a science of user engagement (position paper). In WSDM workshop on user modelling for Web applications (pp. 9-12).

Barnes, N. G., & Lescault, A. M. (2013). College presidents out-blog and out-tweet corporate CEO's as higher ed delves deeper into social media to recruit students. *University of Massachusetts Dartmouth Center for Marketing Research*, 1-9.

Baym, N. K., & Jones, S. G. (1995). Cybersociety: Computer-mediated communication and community. *Illinois: Sage Publication*

Bélanger, C., Bali, S., & Longden, B. (2014). How Canadian universities use social media to brand themselves. Retrieved 10 June 2022

Bliss, C., Kloumann, I., Harris, K., Danforth, C., & Dodds, P. (2012). Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal Of Computational Science*, 3(5), 388-397. doi: 10.1016/j.jocs.2012.05.001

Blight, M. G., Ruppel, E. K., & Schoenbauer, K. V. (2017). Sense of Community on Twitter and Instagram: Exploring the Roles of Motives and Parasocial Relationships. *Cyberpsychology, Behavior, and Social Networking*, 20(5), 314–319. https://doi.org/10.1089/cyber.2016.0505

BostInno [@BostInno]. (n.d). BostInno (@BostInno) [Twitter Account]. Twitter. https://twitter.com/BostInno

Brech, F., Messer, U., Vander Schee, B., Rauschnabel, P., & Ivens, B. (2016). Engaging fans and the community in social media: interaction with institutions of higher education on Facebook. *Journal Of Marketing For Higher Education*, 27(1), 112-130. doi: 10.1080/08841241.2016.1219803

Burke, K. (1974). *The philosophy of literary form* (Vol. 266). *Univ of California Press.*

Burns, T., O'Connor, D., & Stocklmayer, S. (2003). Science Communication: A Contemporary Definition. *Public Understanding Of Science*, 12(2), 183-202. doi: 10.1177/09636625030122004

Cao, D., Meadows, M., Wong, D., & Xia, S. (2021). Understanding consumers' social media engagement behaviour: An examination of the moderation effect of social media context. *Journal Of Business Research*, 122, 835-846. doi: 10.1016/j.jbusres.2020.06.025

Clark, M., Fine, M., & Scheuer, C. (2016). Relationship quality in higher education marketing: the role of social media engagement.

Collins, E., & Hide, B. (2010). Use and relevance of Web 2.0 resources for researchers. *In ELPUB* (pp. 271-289).

Constantinides, E., & Zinck Stagno, M. (2011). Potential of the social media as instruments of higher education marketing: a segmentation study. *Journal Of Marketing For Higher Education,* 21(1), 7-24. doi: 10.1080/08841241.2011.573593

Cosimo, N. (2022). Fine-Tuning BERT for Text Classification. Retrieved 10 June 2022, from https://towardsdatascience.com/fine-tuning-bert-for-text-classification-54e7df642894

Dai, Y., & Wang, T. (2021). Prediction of customer engagement behaviour response to marketing posts based on machine learning. *Connection Science*, 33(4), 891-910. doi: 10.1080/09540091.2021.1912710

Davis, L. (2014). Outreach Activities by Universities as a Channel for Science Communication. *Communicating Science To The Public*, 161-181. doi: 10.1007/978-94-017-9097-0_10

Das, A. C., Gomes, M., Patidar, I. L., & Thomas, R. (2022, May 26). *Social media as a service differentiator: How to win.* McKinsey & Company. https://www.mckinsey.com/business-functions/operations/our-insights/social-media-as-a-service-differentiator-how-to-win

DeMasi, O., Mason, D., & Ma, J. (2016). Understanding Communities via Hashtag Engagement: A Clustering Based Approach. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1), 102-111. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14746

De Melo Maricato, J., & de Castro Manso, B. L. (2022). Characterization of the communities of attention interacting with scientific papers on Twitter: altmetric analysis of a Brazilian University. *Scientometrics*. https://doi.org/10.1007/s11192-022-04442-2

Dron, J., & Anderson, T. (2009). Lost in social space: *Information retrieval issues in Web 1.5.*

Fox, M., Carr, K., D'Agostino McGowan, L., Murray, E., Hidalgo, B., & Banack, H. (2021). Will Podcasting and Social Media Replace Journals and Traditional Science Communication? No, but... *American Journal Of Epidemiology.* doi: 10.1093/aje/kwab172

Gallaugher, J., & Ransbotham, S. (2010). Social media and customer dialog management at Starbucks. *MIS Quarterly Executive,* 9(4).

Gladwell, M. (2010). Small change. *The New Yorker*, 4.

Grace, D., Ross, M., & Shao, W. (2015). Examining the relationship between social media characteristics and psychological dispositions. *European Journal Of Marketing*, 49(9/10), 1366-1390. doi: 10.1108/ejm-06-2014-0347

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. Science, 363(6425), 374-378. doi: 10.1126/science.aau2706

Gruzd, A., Wellman, B., & Takhteyev, Y. (2011). Imagining Twitter as an Imagined Community. *American Behavioral Scientist*, 55(10), 1294–1318. https://doi.org/10.1177/000276421140937

Gruzd, A., Staves, K., & Wilk, A. (2012). Connected scholars: Examining the role of social media in research practices of faculty using the UTAUT model.

Gu, K. Multimodal Transformers Documentation — Multimodal Transformers documentation. Retrieved 20 June 2022, from https://multimodal-toolkit.readthedocs.io/en/latest/index.html

Gu, K., & Budhkar, A. (2021). A Package for Learning on Tabular and Text Data with Transformers. *Proceedings Of The Third Workshop On Multimodal Artificial Intelligence*. doi: 10.18653/v1/2021.maiworkshop-1.10

Haustein, S., Barata, G., & Alperin, J. P. (2018, June 21). *It ain't where you're from, it's where you're tweeting (Or: Where tweets about scholarly articles come from).* Altmetric. Retrieved June 6, 2022, from https://www.altmetric.com/blog/it-aint-where-youre-from-its-where-youre-tweeting-or-where-tweets-about-scholarly-articles-come-from/

Hollebeek, L. D., Glynn, M. S., & Brodie, R. J. (2014). Consumer Brand Engagement in Social Media: Conceptualization, Scale Development and Validation. *Journal of Interactive Marketing*, 28(2), 149–165. https://doi.org/10.1016/j.intmar.2013.12.002

Hoyer, W.D. and MacInnis, D.J. (1997), *Consumer Behaviour*, Houghton Mifflin, Boston, MA.

Hu, Y., Farnham, S., & Talamadupula, K. (2015). Predicting User Engagement on Twitter with Real-World Events. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1), 168-177. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14638

Hwong, Y., Oliver, C., Van Kranendonk, M., Sammut, C., & Seroussi, Y. (2017). What makes you tick? The psychology of social media engagement in space science communication. *Computers In Human Behavior,* 68, 480-492. doi: 10.1016/j.chb.2016.11.068

k-Means Advantages and Disadvantages | Clustering in Machine Learning | Google Developers. (2021). Retrieved 15 June 2022, from https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages

Khan, R., Qian, Y., & Naeem, S. (2019). Extractive based Text Summarization Using KMeans and TF-IDF. *International Journal Of Information Engineering And Electronic Business*, 11(3), 33-44. doi: 10.5815/ijieeb.2019.03.05

Knight, C., & Kaye, L. (2014). 'To tweet or not to tweet?' A comparison of academics' and students' usage of Twitter in academic contexts. *Innovations In Education And Teaching International*, 53(2), 145-155. doi: 10.1080/14703297.2014.928229

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media?. *Proceedings Of The 19Th International Conference On World Wide Web - WWW '10*. doi: 10.1145/1772690.1772751

Lemon, K., & Verhoef, P. (2016). Understanding Customer Experience Throughout the Customer Journey. *Journal Of Marketing*, 80(6), 69-96. doi: 10.1509/jm.15.0420

López-Goñi, I., & Sánchez-Angulo, M. (2017). Social networks as a tool for science communication and public engagement: focus on Twitter. *FEMS Microbiology Letters*, 365(2). doi: 10.1093/femsle/fnx246

Lowe, B., & Laffey, D. (2011). Is Twitter for the Birds?. *Journal Of Marketing Education*, 33(2), 183-192. doi: 10.1177/0273475311410851

Luqman, A., Cao, X., Ali, A., Masood, A., & Yu, L. (2017). Empirical investigation of Facebook discontinues usage intentions based on SOR paradigm.

Magalhães, J., Pessoa, R., Souza, C., Costa, E., & Fechine, J. (2014). A Recommender System for Predicting User Engagement in Twitter. *Proceedings Of The 2014 Recommender Systems Challenge On - Recsyschallenge '14*. doi: 10.1145/2668067.2668078

Marcus, J. (2021). From Google ads to NFL sponsorships: Colleges throw billions at marketing themselves to attract students. Retrieved 11 June 2022, from https://www.washingtonpost.com/local/education/colleges-marketing-student-recruitment/2021/09/30/b6ddd246-2166-11ec-8200-5e3fd4c49f5e_story.html

Maresova, P., Hruska, J., & Kuca, K. (2020). Social Media University Branding. *Education Sciences,* 10(3), 74. doi: 10.3390/educsci10030074

Motta, J., & Barbosa, M. (2018). Social Media as a Marketing Tool for European and North American Universities and Colleges. *Journal Of Intercultural Management*, 10(3), 125-154. doi: 10.2478/joim-2018-0020

Muñoz-Expósito, M., Oviedo-García, M., & Castellanos-Verdugo, M. (2017). How to measure engagement in Twitter: advancing a metric. *Internet Research*, 27(5), 1122-1148. doi: 10.1108/intr-06-2016-0170

Nur'aini, K., Najahaty, I., Hidayati, L., Murfi, H., & Nurrohmah, S. (2015). Combination of singular value decomposition and K-means clustering methods for topic detection on Twitter. *2015 International Conference On Advanced Computer Science And Information Systems (ICACSIS).* doi: 10.1109/icacsis.2015.7415168

Olvera-Lobo, D., & López-Pérez, L. (2014). Science Communication 2.0. *Information Resources Management Journal*, 27(3), 42-58. doi: 10.4018/irmj.2014070104

Paul, M., & Dredze, M. (2021). You Are What You Tweet: Analyzing Twitter for Public Health. *Proceedings of the International AAAI Conference on Web and Social Media,* 5(1), 265-272. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14137

Peruta, A., & Shields, A. (2017). Social media in higher education: understanding how colleges and universities use Facebook. *Journal Of Marketing For Higher Education*, 27(1), 131-143. doi: 10.1080/08841241.2016.1212451

Pranckevicius, T., & Marcinkevicius, V. (2016). Application of Logistic Regression with part-of-the-speech tagging for multi-class text classification. 2016 *IEEE 4Th Workshop On Advances In Information, Electronic And Electrical Engineering (AIEEE)*. doi: 10.1109/aieee.2016.7821805

Priem, J., & Costello, K. (2010). How and why scholars cite on Twitter. *Proceedings Of The American Society For Information Science And Technology*, 47(1), 1-4. doi: 10.1002/meet.14504701201

Raj, A. (2020). The Perfect Recipe for Classification Using Logistic Regression. Retrieved 11 June 2022, from https://towardsdatascience.com/the-perfect-recipe-for-classification-using-logistic-regression-f8648e267592#:~:text=Logistic%20Regression%20is%20a%20classification%20technique%20used%20in%20machine%20learning,cancer%20is%20malignant%20or%20not).

Rejito, J., Atthariq, A., & Abdullah, A. (2021). Application of text mining employing k-means algorithms for clustering tweets of Tokopedia. *Journal Of Physics: Conference Series*, 1722(1), 012019. doi: 10.1088/1742-6596/1722/1/012019

Sashi, C. (2012). Customer engagement, buyer-seller relationships, and social media. *Management Decision*, 50(2), 253-272. doi: 10.1108/00251741211203551

Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. *2011 31St International Conference* On Distributed Computing Systems Workshops. doi: 10.1109/icdcsw.2011.20

Shanghai Ranking. (2021). *ShanghaiRanking's Academic Ranking of World Universities*. Retrieved Jule 6, 2022, from https://www.shanghairanking.com/rankings/arwu/2021

Shukla, S., & Naganna, S. (2014). A review on K-means data clustering approach. *International Journal of Information and Computation Technology*, 4(17), 1847-1860.

Siyam, N., Alqaryouti, O., & Abdallah, S. (2019). Mining government tweets to identify and predict citizens engagement. *Technology In Society*, 60, 101211. doi: 10.1016/j.techsoc.2019.101211

Sousa, D., Sarmento, L., & Mendes Rodrigues, E. (2010). Characterization of the twitter @replies network. *Proceedings Of The 2Nd International Workshop On Search And Mining User-Generated Contents - SMUC '10*. doi: 10.1145/1871985.1871996

Soukup, C. (2006). Hitching a Ride on a Star: Celebrity, Fandom, and Identification on the World Wide Web. *Southern Communication Journal*, 71(4), 319–337. https://doi.org/10.1080/10417940601000410

Stvilia, B., & Gibradze, L. (2017). Examining Undergraduate Students' Priorities for Academic Library Services and Social Media Communication. *The Journal Of Academic Librarianship*, 43(3), 257-262. doi: 10.1016/j.acalib.2017.02.013

Suh, B., Hong, L., Pirolli, P., & Chi, E. (2010). Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. *2010 IEEE Second International Conference On Social Computing*. doi: 10.1109/socialcom.2010.33

Taecharungroj, V. (2017). Higher education social media marketing: 12 content types universities post on Facebook. *International Journal Of Management In Education,* 11(2), 111. doi: 10.1504/ijmie.2017.083350

Teh, G. M., & Salleh, A. H. M. (2011). Impact of brand meaning on brand equity of higher educational institutions in Malaysia*. World*, 3(2), 218-228.

The Economist Intelligence Unit. (2007). Beyond loyalty. Meeting the challenge of customer engagement. Retrieved from http://graphics.eiu.com/files/ad_pdfs/eiu_AdobeEngagementPt_I_wp.pdf

Toraman, C., Şahinuç, F., Yilmaz, E., & Akkaya, I. (2022). Understanding social engagements: A comparative analysis of user and text features in Twitter. *Social Network Analysis And Mining*, 12(1). doi: 10.1007/s13278-022-00872-1

Trench, B. (2012). Vital and Vulnerable: Science Communication as a University Subject. *Science Communication In The World,* 241-257. doi: 10.1007/978-94-007-4279-6_16

Trench, B. (2017). Universities, science communication and professionalism. *Journal Of Science Communication*, 16(05). doi: 10.22323/2.16050302 van Doorn, J., Lemon, K. N., Mittal, V., Nass, S., Pick, D., Pirner, P., & Verhoef, P. C. (2010). Customer Engagement Behavior: Theoretical Foundations and Research Directions. Journal of Service Research, 13(3), 253–266. https://doi.org/10.1177/1094670510375599

Veletsianos, G. (2011). Higher education scholars' participation and practices on Twitter. *Journal of Computer Assisted Learning*, 28(4), 336–349. https://doi.org/10.1111/j.1365-2729.2011.00449.x

Veletsianos, G., Kimmons, R., Shaw, A., Pasquini, L., & Woodward, S. (2017). Selective openness, branding, broadcasting, and promotion: Twitter use in Canada's public universities. *Educational Media International*, 54(1), 1–19. https://doi.org/10.1080/09523987.2017.1324363

Volkovs, M., Cheng, Z., Ravaut, M., Yang, H., Shen, K., & Zhou, J. et al. (2020). Predicting Twitter Engagement With Deep Language Models. Proceedings Of The Recommender Systems Challenge 2020. doi: 10.1145/3415959.3416000

Wadhwa, V., Latimer, E., Chatterjee, K., McCarty, J., & Fitzgerald, R. (2017). Maximizing the Tweet Engagement Rate in Academia: Analysis of the AJNR Twitter Feed. *American Journal Of Neuroradiology*, 38(10), 1866-1868. doi: 10.3174/ajnr.a5283

Weale, S. (2019, April 3). *Universities spending millions on marketing to attract students*. The Guardian. https://www.theguardian.com/education/2019/apr/02/universities-spending-millions-on-marketing-to-attract-students

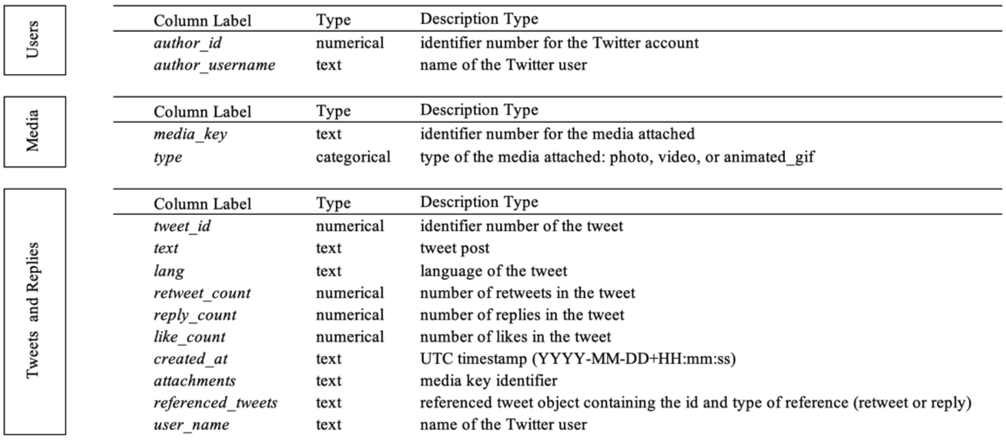
Welcome to LightGBM's documentation! — LightGBM 3.3.2.99 documentation. (2022). LightGBM. https://lightgbm.readthedocs.io/en/latest/

Williams, M., Burnap, P., & Sloan, L. (2017). Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. *Sociology*, 51(6), 1149-1168. doi: 10.1177/0038038517708140

Yiu, T. (2019). Understanding Random Forest. Retrieved 12 June 2022, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Zhao, S., Zhong, L., Wickramasuriya, J., & Vasudevan, V. (2011). Human as Real-Time Sensors of Social and Physical Events: A Case Study of Twitter and Sports Games. ArXiv, abs/1106.4300.

7. Appendix

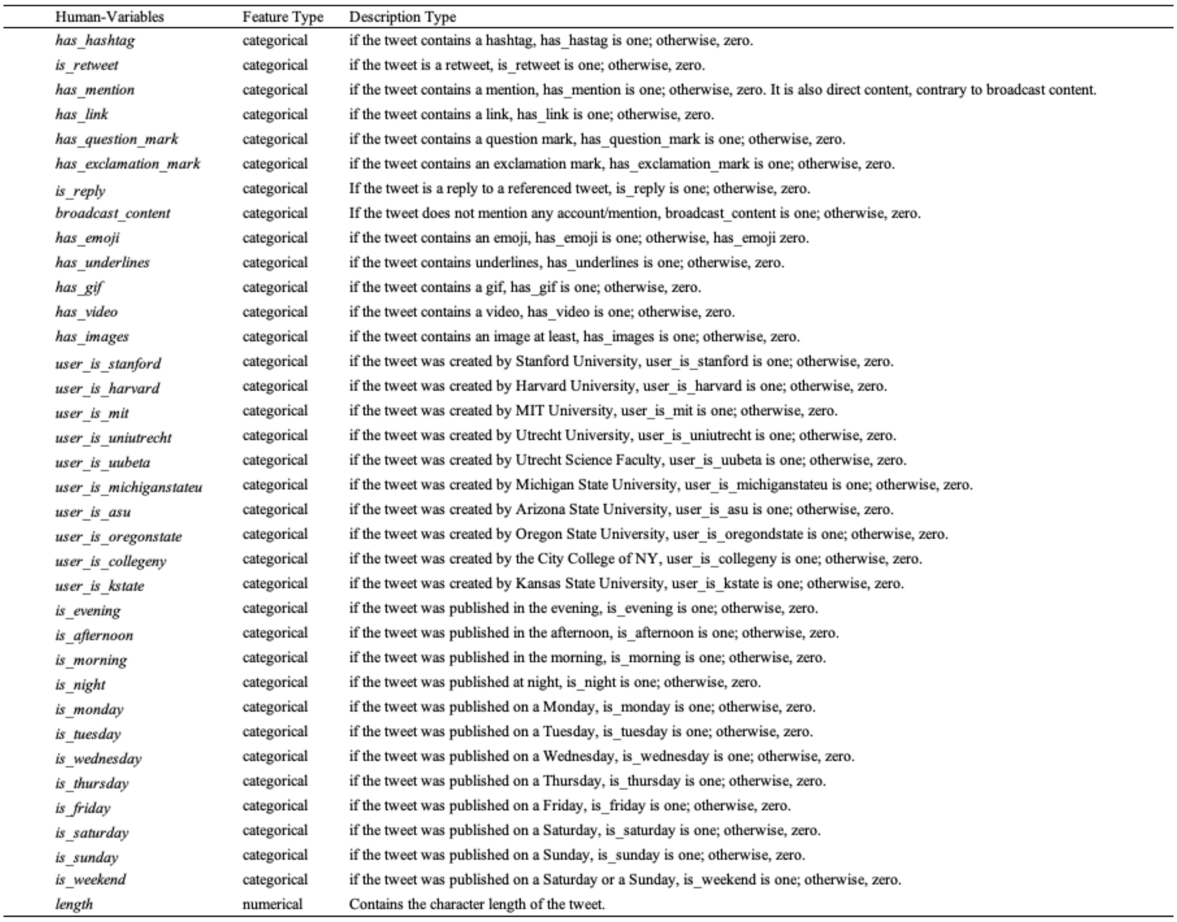


	Column Label	Type	Description Type
Users	*author_id*	numerical	identifier number for the Twitter account
	author_username	text	name of the Twitter user

	Column Label	Type	Description Type
Media	*media_key*	text	identifier number for the media attached
	type	categorical	type of the media attached: photo, video, or animated_gif

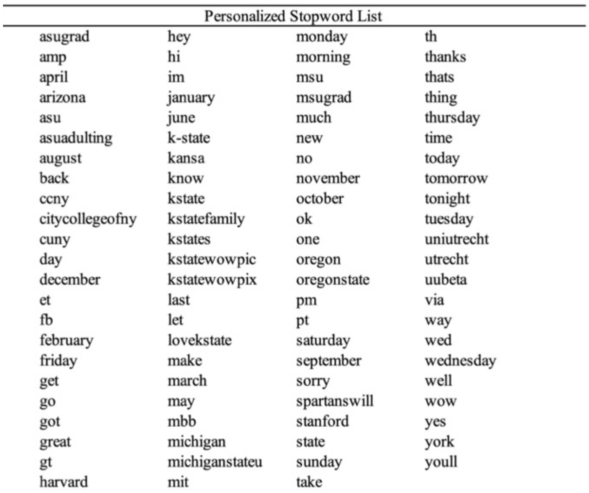
	Column Label	Type	Description Type
Tweets and Replies	*tweet_id*	numerical	identifier number of the tweet
	text	text	tweet post
	lang	text	language of the tweet
	retweet_count	numerical	number of retweets in the tweet
	reply_count	numerical	number of replies in the tweet
	like_count	numerical	number of likes in the tweet
	created_at	text	UTC timestamp (YYYY-MM-DD+HH:mm:ss)
	attachments	text	media key identifier
	referenced_tweets	text	referenced tweet object containing the id and type of reference (retweet or reply)
	user_name	text	name of the Twitter user

***Table** 7.1. Overview of the datasets from Twitter API used*



Human-Variables	Feature Type	Description Type
has_hashtag	categorical	if the tweet contains a hashtag, has_hastag is one; otherwise, zero.
is_retweet	categorical	if the tweet is a retweet, is_retweet is one; otherwise, zero.
has_mention	categorical	if the tweet contains a mention, has_mention is one; otherwise, zero. It is also direct content, contrary to broadcast content.
has_link	categorical	if the tweet contains a link, has_link is one; otherwise, zero.
has_question_mark	categorical	if the tweet contains a question mark, has_question_mark is one; otherwise, zero.
has_exclamation_mark	categorical	if the tweet contains an exclamation mark, has_exclamation_mark is one; otherwise, zero.
is_reply	categorical	If the tweet is a reply to a referenced tweet, is_reply is one; otherwise, zero.
broadcast_content	categorical	If the tweet does not mention any account/mention, broadcast_content is one; otherwise, zero.
has_emoji	categorical	if the tweet contains an emoji, has_emoji is one; otherwise, has_emoji zero.
has_underlines	categorical	if the tweet contains underlines, has_underlines is one; otherwise, zero.
has_gif	categorical	if the tweet contains a gif, has_gif is one; otherwise, zero.
has_video	categorical	if the tweet contains a video, has_video is one; otherwise, zero.
has_images	categorical	if the tweet contains an image at least, has_images is one; otherwise, zero.
user_is_stanford	categorical	if the tweet was created by Stanford University, user_is_stanford is one; otherwise, zero.
user_is_harvard	categorical	if the tweet was created by Harvard University, user_is_harvard is one; otherwise, zero.
user_is_mit	categorical	if the tweet was created by MIT University, user_is_mit is one; otherwise, zero.
user_is_uniutrecht	categorical	if the tweet was created by Utrecht University, user_is_uniutrecht is one; otherwise, zero.
user_is_uubeta	categorical	if the tweet was created by Utrecht Science Faculty, user_is_uubeta is one; otherwise, zero.
user_is_michiganstateu	categorical	if the tweet was created by Michigan State University, user_is_michiganstateu is one; otherwise, zero.
user_is_asu	categorical	if the tweet was created by Arizona State University, user_is_asu is one; otherwise, zero.
user_is_oregonstate	categorical	if the tweet was created by Oregon State University, user_is_oregondstate is one; otherwise, zero.
user_is_collegeny	categorical	if the tweet was created by the City College of NY, user_is_collegeny is one; otherwise, zero.
user_is_kstate	categorical	if the tweet was created by Kansas State University, user_is_kstate is one; otherwise, zero.
is_evening	categorical	if the tweet was published in the evening, is_evening is one; otherwise, zero.
is_afternoon	categorical	if the tweet was published in the afternoon, is_afternoon is one; otherwise, zero.
is_morning	categorical	if the tweet was published in the morning, is_morning is one; otherwise, zero.
is_night	categorical	if the tweet was published at night, is_night is one; otherwise, zero.
is_monday	categorical	if the tweet was published on a Monday, is_monday is one; otherwise, zero.
is_tuesday	categorical	if the tweet was published on a Tuesday, is_tuesday is one; otherwise, zero.
is_wednesday	categorical	if the tweet was published on a Wednesday, is_wednesday is one; otherwise, zero.
is_thursday	categorical	if the tweet was published on a Thursday, is_thursday is one; otherwise, zero.
is_friday	categorical	if the tweet was published on a Friday, is_friday is one; otherwise, zero.
is_saturday	categorical	if the tweet was published on a Saturday, is_saturday is one; otherwise, zero.
is_sunday	categorical	if the tweet was published on a Sunday, is_sunday is one; otherwise, zero.
is_weekend	categorical	if the tweet was published on a Saturday or a Sunday, is_weekend is one; otherwise, zero.
length	numerical	Contains the character length of the tweet.

***Table** 7.2. Human selected variables description*



Personalized Stopword List

asugrad
amp
april
arizona
asu
asuadulting
august
back
ccny
citycollegeofny
cuny
day
december
et
fb
february
friday
get
go
got
great
gt
harvard
hey
hi
im
january
june
k-state
kansa
know
kstate
kstatefamily
kstates
kstatewowpic
kstatewowpix
last
let
lovekstate
make
march
may
mbb
michigan
michiganstateu
mit
monday
morning
msu
msugrad
much
new
no
november
october
ok
one
oregon
oregonstate
pm
pt
saturday
september
sorry
spartanswill
stanford
state
sunday
take
th
thanks
thats
thing
thursday
time
today
tomorrow
tonight
tuesday
uniutrecht
utrecht
uubeta
via
way
wed
wednesday
well
wow
yes
york
youll

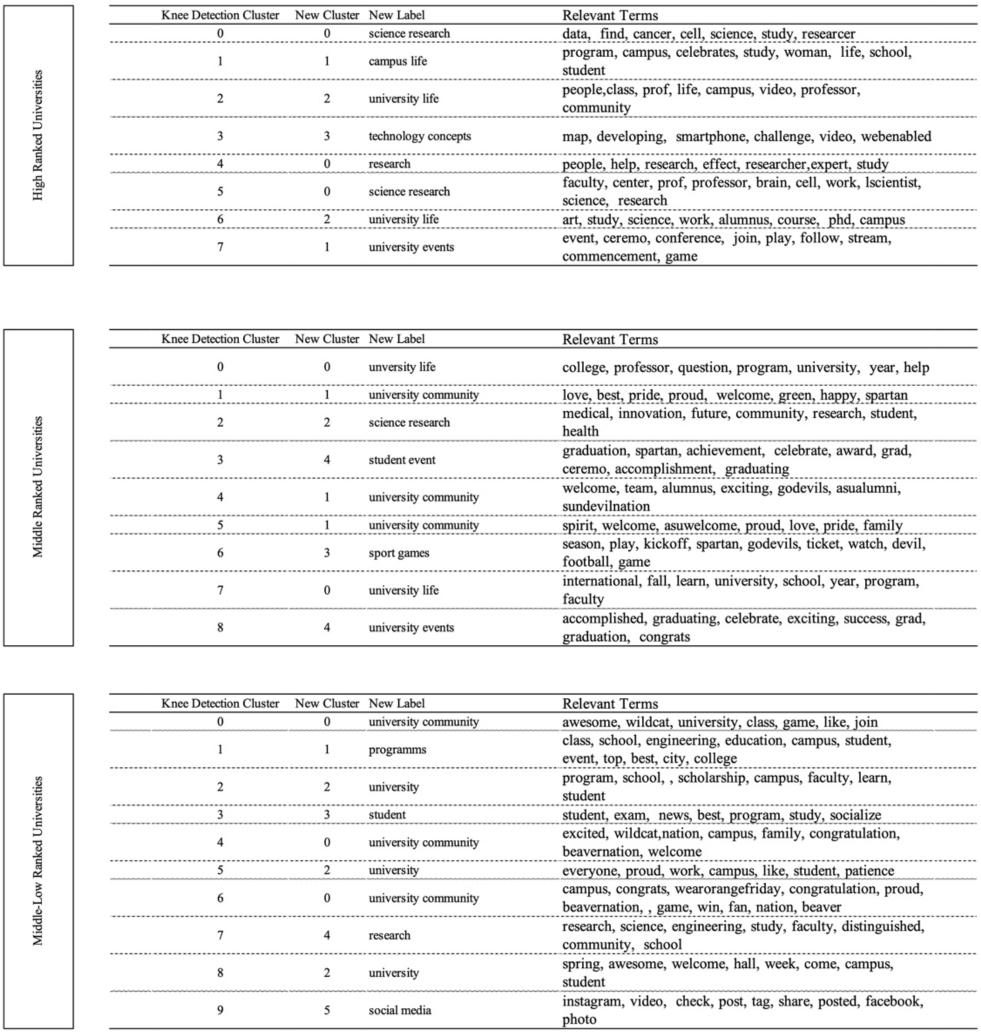
Table 7.3. *Personalized stopword list*

Source	Target	Cluster	Size
@ASU	@ASUBiodesign	2	3
@ASU	@ASUCHPDP	2	1
@ASU	@ASUCollegeOfLaw	2	3
@ASU	@ASUCollegeofGF	2	1
@ASU	@ASUEmbeddedness	2	1

Table 7.4. *Node table example*

Node	Cluster
@ASU	2
@CityCollegeNY	4
@Harvard	0
@KState	4
@MIT	0
@OregonState	4

Table 7.5. *Edge table example*



High Ranked Universities

Knee Detection Cluster	New Cluster	New Label	Relevant Terms
0	0	science research	data, find, cancer, cell, science, study, researcer
1	1	campus life	program, campus, celebrates, study, woman, life, school, student
2	2	university life	people,class, prof, life, campus, video, professor, community
3	3	technology concepts	map, developing, smartphone, challenge, video, webenabled
4	0	research	people, help, research, effect, researcher,expert, study
5	0	science research	faculty, center, prof, professor, brain, cell, work, lscientist, science, research
6	2	university life	art, study, science, work, alumnus, course, phd, campus
7	1	university events	event, ceremo, conference, join, play, follow, stream, commencement, game

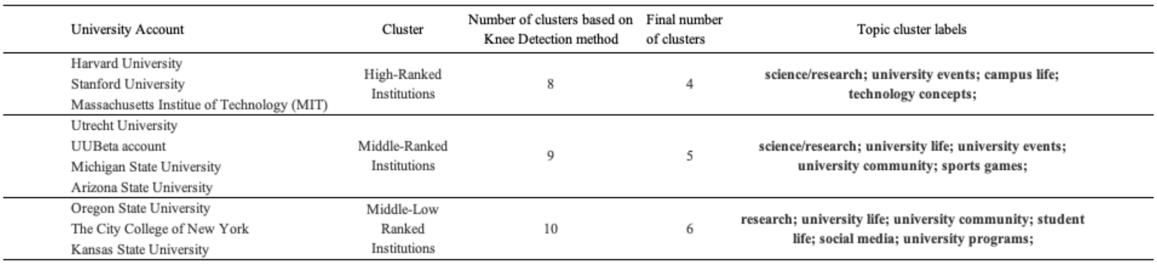
Middle Ranked Universities

Knee Detection Cluster	New Cluster	New Label	Relevant Terms
0	0	unversity life	college, professor, question, program, university, year, help
1	1	university community	love, best, pride, proud, welcome, green, happy, spartan
2	2	science research	medical, innovation, future, community, research, student, health
3	4	student event	graduation, spartan, achievement, celebrate, award, grad, ceremo, accomplishment, graduating
4	1	university community	welcome, team, alumnus, exciting, godevils, asualumni, sundevilnation
5	1	university community	spirit, welcome, asuwelcome, proud, love, pride, family
6	3	sport games	season, play, kickoff, spartan, godevils, ticket, watch, devil, football, game
7	0	university life	international, fall, learn, university, school, year, program, faculty
8	4	university events	accomplished, graduating, celebrate, exciting, success, grad, graduation, congrats

Middle-Low Ranked Universities

Knee Detection Cluster	New Cluster	New Label	Relevant Terms
0	0	university community	awesome, wildcat, university, class, game, like, join
1	1	programms	class, school, engineering, education, campus, student, event, top, best, city, college
2	2	university	program, school, , scholarship, campus, faculty, learn, student
3	3	student	student, exam, news, best, program, study, socialize
4	0	university community	excited, wildcat,nation, campus, family, congratulation, beavernation, welcome
5	2	university	everyone, proud, work, campus, like, student, patience
6	0	university community	campus, congrats, wearorangefriday, congratulation, proud, beavernation, , game, win, fan, nation, beaver
7	4	research	research, science, engineering, study, faculty, distinguished, community, school
8	2	university	spring, awesome, welcome, hall, week, come, campus, student
9	5	social media	instagram, video, check, post, tag, share, posted, facebook, photo

Table 7.6. *Relevant words from the Topic Clustering*



University Account	Cluster	Number of clusters based on Knee Detection method	Final number of clusters	Topic cluster labels
Harvard University Stanford University Massachusetts Institue of Technology (MIT)	High-Ranked Institutions	8	4	**science/research; university events; campus life; technology concepts;**
Utrecht University UUBeta account Michigan State University Arizona State University	Middle-Ranked Institutions	9	5	**science/research; university life; university events; university community; sports games;**
Oregon State University The City College of New York Kansas State University	Middle-Low Ranked Institutions	10	6	**research; university life; university community; student life; social media; university programs;**

Table 7.7. *Topic Clustering Results*

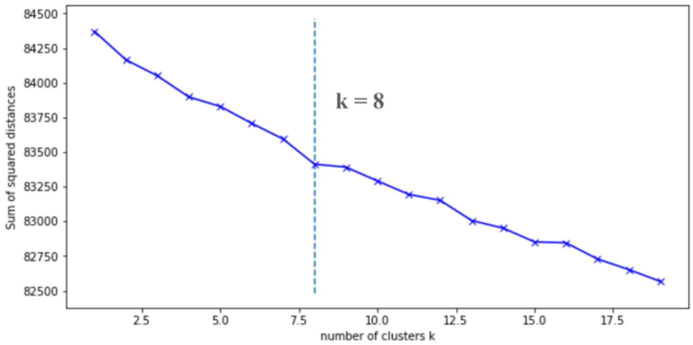


Figure 12*. Knee detection line chart for Higher Ranked Universities*

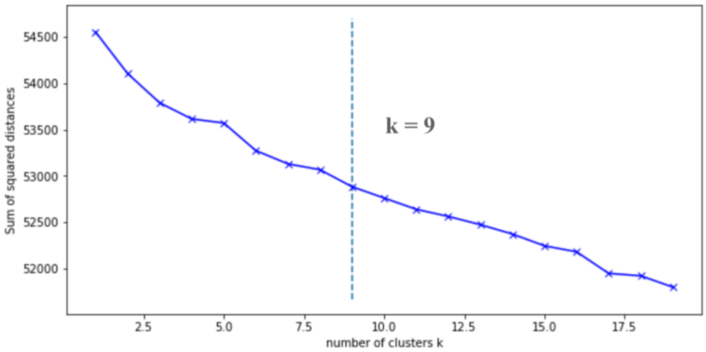


Figure 13*. Knee detection line chart for Middle Ranked Universities*

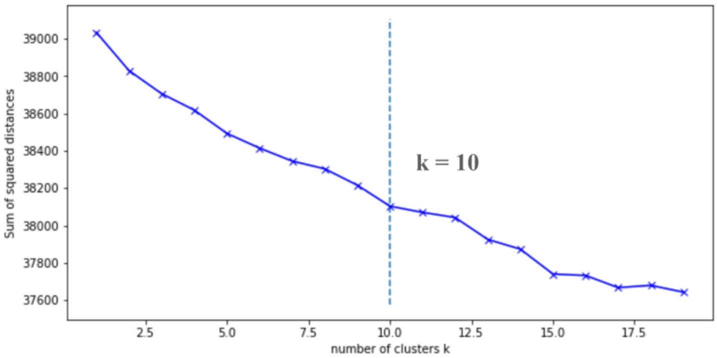
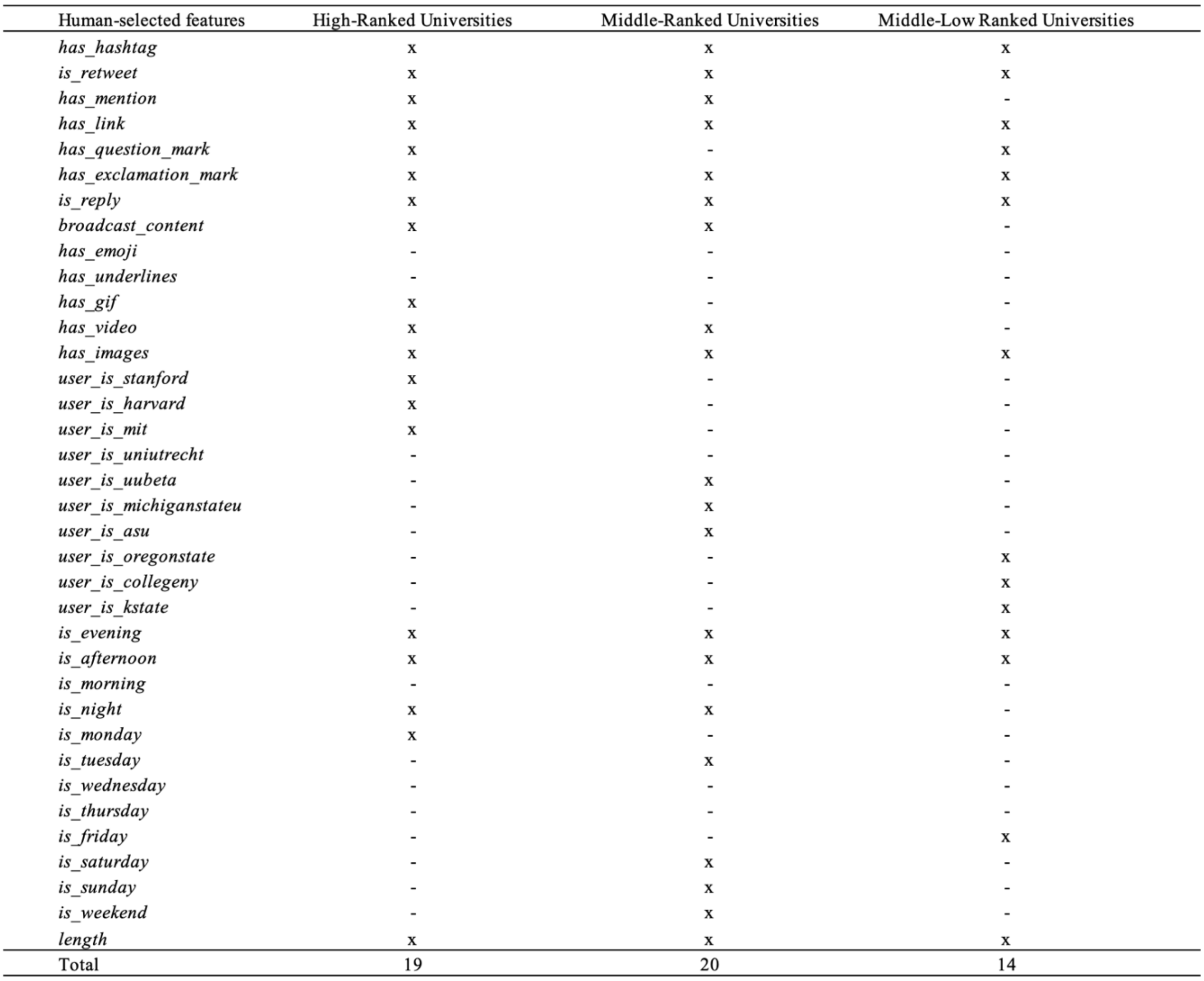


Figure 14. *Knee detection line chart for Middle-Low Ranked Universities*

7.1. Full data exploration results



Human-selected features	High-Ranked Universities	Middle-Ranked Universities	Middle-Low Ranked Universities
has_hashtag	x	x	x
is_retweet	x	x	x
has_mention	x	x	-
has_link	x	x	x
has_question_mark	x	-	x
has_exclamation_mark	x	x	x
is_reply	x	x	x
broadcast_content	x	x	-
has_emoji	-	-	-
has_underlines	-	-	-
has_gif	x	-	-
has_video	x	x	-
has_images	x	x	x
user_is_stanford	x	-	-
user_is_harvard	x	-	-
user_is_mit	x	-	-
user_is_uniutrecht	-	-	-
user_is_uubeta	-	x	-
user_is_michiganstateu	-	x	-
user_is_asu	-	x	-
user_is_oregonstate	-	-	x
user_is_collegeny	-	-	x
user_is_kstate	-	-	x
is_evening	x	x	x
is_afternoon	x	x	x
is_morning	-	-	-
is_night	x	x	-
is_monday	x	-	-
is_tuesday	-	x	-
is_wednesday	-	-	-
is_thursday	-	-	-
is_friday	-	-	x
is_saturday	-	x	-
is_sunday	-	x	-
is_weekend	-	x	-
length	x	x	x
Total	19	20	14

***Table** 7.8. Selected Explanatory variables*

Human Feature	P Value (α = 0.05)	Result	Null Hypothesis
has_hashtag	3,553E-66	Dependent	H0 Rejected
is_retweet	0,000E+00	Dependent	H0 Rejected
has_mention	1,792E-50	Dependent	H0 Rejected
has_link	0,000E+00	Dependent	H0 Rejected
has_question_mark	3,883E+03	Dependent	H0 Rejected
has_exclamation_mark	3,805E+09	Dependent	H0 Rejected
has_gif	8,027E+08	Dependent	H0 Rejected
has_video	2,251E-04	Dependent	H0 Rejected
has_images	4,059E-274	Dependent	H0 Rejected
user_is_stanford	8,147E+09	Dependent	H0 Rejected
is_afternoon	3,130E-03	Dependent	H0 Rejected
is_monday	3,295E-02	Dependent	H0 Rejected
user_is_harvard	6,561E-03	Dependent	H0 Rejected
user_is_mit	3,149E+08	Dependent	H0 Rejected
is_evening	1,153E+02	Dependent	H0 Rejected
is_night	1,583E+11	Dependent	H0 Rejected
is_reply	2,469E+04	Dependent	H0 Rejected
broadcast_content	1,792E-50	Dependent	H0 Rejected
is_weekend	3,433E-01	Independent	H0 Holds True
has_emoji	1,000E+00	Independent	H0 Holds True
has_underlines	1,000E+00	Independent	H0 Holds True
user_is_uniutrecht	1,000E+00	Independent	H0 Holds True
user_is_uubeta	1,000E+00	Independent	H0 Holds True
user_is_michiganstateu	1,000E+00	Independent	H0 Holds True
user_is_asu	1,000E+00	Independent	H0 Holds True
user_is_oregonstate	1,000E+00	Independent	H0 Holds True
user_is_collegeny	1,000E+00	Independent	H0 Holds True
user_is_kstate	1,000E+00	Independent	H0 Holds True
is_morning	1,727E-01	Independent	H0 Holds True
is_tuesday	8,063E-01	Independent	H0 Holds True
is_wednesday	6,797E-01	Independent	H0 Holds True
is_thursday	1,655E-01	Independent	H0 Holds True
is_friday	8,930E-01	Independent	H0 Holds True
is_saturday	5,920E-01	Independent	H0 Holds True
is_sunday	6,372E-01	Independent	H0 Holds True

Table 7.9. *Chi-Square Results for High Ranked Universities*

Human Feature	P Value (α = 0.05)	Result	Null Hypothesis
length	0,000E+00	Dependent	H0 Rejected

Table 7.10. *ANOVA Results for High Ranked Universities*

Human Feature	P Value (α = 0.05)	Result	Null Hypothesis
has_hashtag	7,067E-03	Dependent	H0 Rejected
is_retweet	3,253E-188	Dependent	H0 Rejected
has_mention	2,347E-04	Dependent	H0 Rejected
has_link	1,302E-67	Dependent	H0 Rejected
has_exclamation_mark	6,416E+07	Dependent	H0 Rejected
is_weekend	1,587E-04	Dependent	H0 Rejected
has_video	7,369E+10	Dependent	H0 Rejected
has_images	2,044E-57	Dependent	H0 Rejected
is_afternoon	9,370E-03	Dependent	H0 Rejected
user_is_uubeta	2,927E-02	Dependent	H0 Rejected
user_is_michiganstateu	8,842E+08	Dependent	H0 Rejected
user_is_asu	4,173E+10	Dependent	H0 Rejected
is_evening	8,804E-03	Dependent	H0 Rejected
is_night	1,054E+11	Dependent	H0 Rejected
is_tuesday	1,852E-02	Dependent	H0 Rejected
is_saturday	1,716E-02	Dependent	H0 Rejected
is_sunday	1,496E-02	Dependent	H0 Rejected
is_reply	1,099E-63	Dependent	H0 Rejected
broadcast_content	2,347E-04	Dependent	H0 Rejected
has_question_mark	8,161E-01	Independent	H0 Holds True
has_emoji	1,000E+00	Independent	H0 Holds True
has_underlines	1,000E+00	Independent	H0 Holds True
has_gif	3,845E-01	Independent	H0 Holds True
user_is_stanford	1,000E+00	Independent	H0 Holds True
is_monday	5,100E-01	Independent	H0 Holds True
user_is_harvard	1,000E+00	Independent	H0 Holds True
user_is_mit	1,000E+00	Independent	H0 Holds True
user_is_uniutrecht	2,958E-01	Independent	H0 Holds True
user_is_oregonstate	1,000E+00	Independent	H0 Holds True
user_is_collegeny	1,000E+00	Independent	H0 Holds True
user_is_kstate	1,000E+00	Independent	H0 Holds True
is_morning	1,919E-01	Independent	H0 Holds True
is_wednesday	2,925E-01	Independent	H0 Holds True
is_thursday	9,329E-01	Independent	H0 Holds True
is_friday	1,089E-01	Independent	H0 Holds True

Table 7.11. *Chi-Square Results for Middle Ranked Universities*

Human Feature	P Value (α = 0.05)	Result	Null Hypothesis
length	0,000E+00	Dependent	H0 Rejected

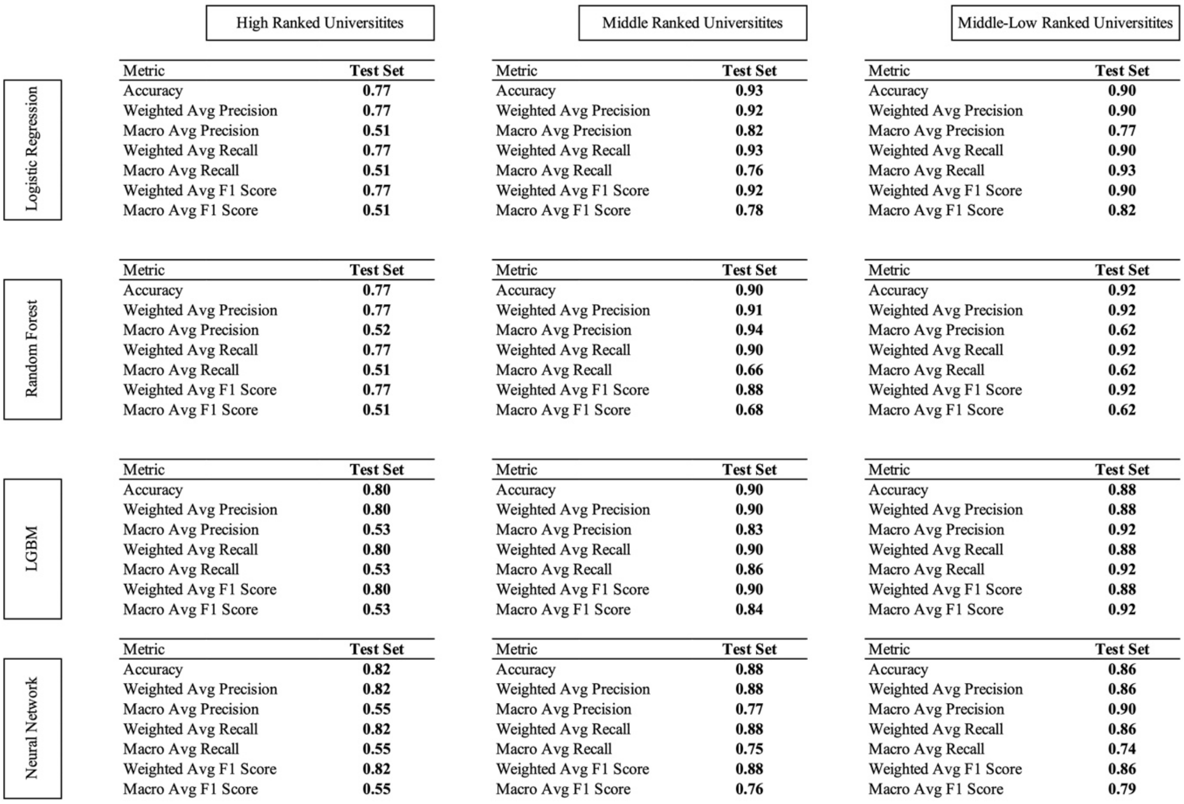
Table 7.12. *ANOVA Results for Middle Ranked Universities*

Human Feature	P Value (α = 0.05)	Result	Null Hypothesis
has_hashtag	7,067E-03	Dependent	H0 Rejected
is_retweet	3,253E-188	Dependent	H0 Rejected
has_mention	2,347E-04	Dependent	H0 Rejected
has_link	1,302E-67	Dependent	H0 Rejected
has_exclamation_mark	6,416E+07	Dependent	H0 Rejected
is_weekend	1,587E-04	Dependent	H0 Rejected
has_video	7,369E+10	Dependent	H0 Rejected
has_images	2,044E-57	Dependent	H0 Rejected
is_afternoon	9,370E-03	Dependent	H0 Rejected
user_is_uubeta	2,927E-02	Dependent	H0 Rejected
user_is_michiganstateu	8,842E+08	Dependent	H0 Rejected
user_is_asu	4,173E+10	Dependent	H0 Rejected
is_evening	8,804E-03	Dependent	H0 Rejected
is_night	1,054E+11	Dependent	H0 Rejected
is_tuesday	1,852E-02	Dependent	H0 Rejected
is_saturday	1,716E-02	Dependent	H0 Rejected
is_sunday	1,496E-02	Dependent	H0 Rejected
is_reply	1,099E-63	Dependent	H0 Rejected
broadcast_content	2,347E-04	Dependent	H0 Rejected
has_question_mark	8,161E-01	Independent	H0 Holds True
has_emoji	1,000E+00	Independent	H0 Holds True
has_underlines	1,000E+00	Independent	H0 Holds True
has_gif	3,845E-01	Independent	H0 Holds True
user_is_stanford	1,000E+00	Independent	H0 Holds True
is_monday	5,100E-01	Independent	H0 Holds True
user_is_harvard	1,000E+00	Independent	H0 Holds True
user_is_mit	1,000E+00	Independent	H0 Holds True
user_is_uniutrecht	2,958E-01	Independent	H0 Holds True
user_is_oregonstate	1,000E+00	Independent	H0 Holds True
user_is_collegeny	1,000E+00	Independent	H0 Holds True
user_is_kstate	1,000E+00	Independent	H0 Holds True
is_morning	1,919E-01	Independent	H0 Holds True
is_wednesday	2,925E-01	Independent	H0 Holds True
is_thursday	9,329E-01	Independent	H0 Holds True
is_friday	1,089E-01	Independent	H0 Holds True

***Table 7.13**. Chi-Square Results for Middle-Low Ranked Universities*

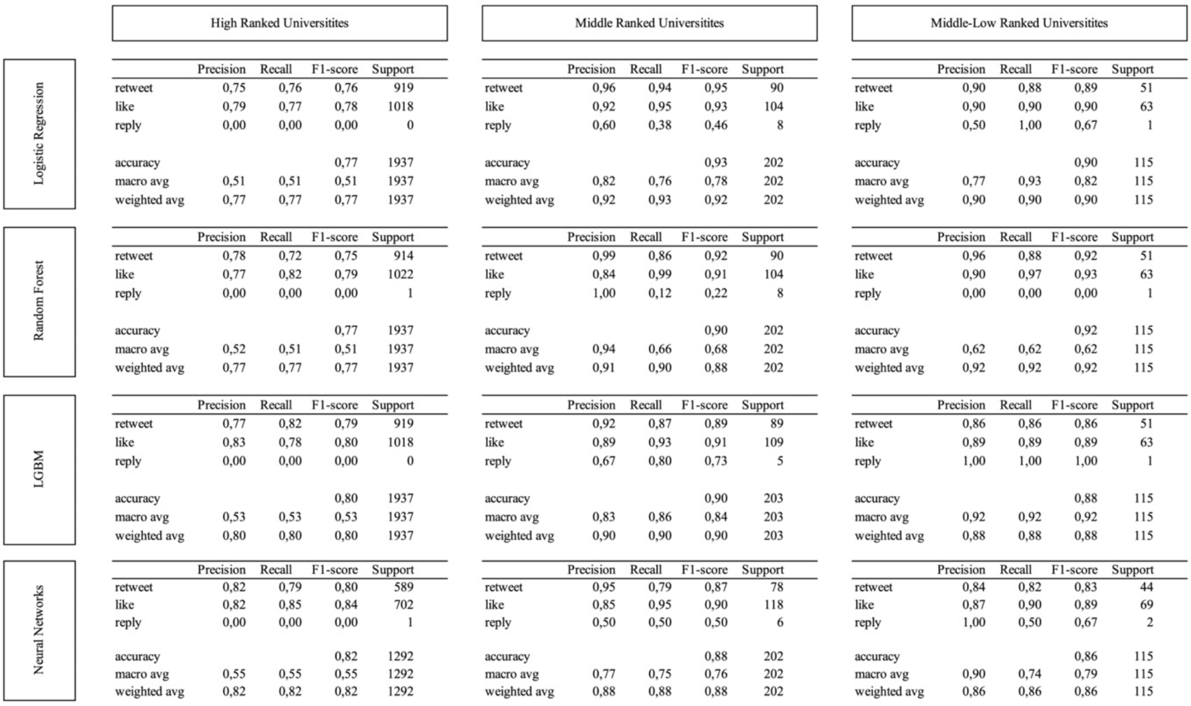
Human Feature	P Value (α = 0.05)	Result	Null Hypothesis
length	0,000E+00	Dependent	H0 Rejected

***Table 7.14.** ANOVA Results for Middle-Low Ranked Universities*



	High Ranked Universitites		Middle Ranked Universitites		Middle-Low Ranked Universitites	
	Metric	**Test Set**	**Metric**	**Test Set**	**Metric**	**Test Set**
Logistic Regression	Accuracy	**0.77**	Accuracy	**0.93**	Accuracy	**0.90**
	Weighted Avg Precision	**0.77**	Weighted Avg Precision	**0.92**	Weighted Avg Precision	**0.90**
	Macro Avg Precision	**0.51**	Macro Avg Precision	**0.82**	Macro Avg Precision	**0.77**
	Weighted Avg Recall	**0.77**	Weighted Avg Recall	**0.93**	Weighted Avg Recall	**0.90**
	Macro Avg Recall	**0.51**	Macro Avg Recall	**0.76**	Macro Avg Recall	**0.93**
	Weighted Avg F1 Score	**0.77**	Weighted Avg F1 Score	**0.92**	Weighted Avg F1 Score	**0.90**
	Macro Avg F1 Score	**0.51**	Macro Avg F1 Score	**0.78**	Macro Avg F1 Score	**0.82**
	Metric	**Test Set**	**Metric**	**Test Set**	**Metric**	**Test Set**
Random Forest	Accuracy	**0.77**	Accuracy	**0.90**	Accuracy	**0.92**
	Weighted Avg Precision	**0.77**	Weighted Avg Precision	**0.91**	Weighted Avg Precision	**0.92**
	Macro Avg Precision	**0.52**	Macro Avg Precision	**0.94**	Macro Avg Precision	**0.62**
	Weighted Avg Recall	**0.77**	Weighted Avg Recall	**0.90**	Weighted Avg Recall	**0.92**
	Macro Avg Recall	**0.51**	Macro Avg Recall	**0.66**	Macro Avg Recall	**0.62**
	Weighted Avg F1 Score	**0.77**	Weighted Avg F1 Score	**0.88**	Weighted Avg F1 Score	**0.92**
	Macro Avg F1 Score	**0.51**	Macro Avg F1 Score	**0.68**	Macro Avg F1 Score	**0.62**
	Metric	**Test Set**	**Metric**	**Test Set**	**Metric**	**Test Set**
LGBM	Accuracy	**0.80**	Accuracy	**0.90**	Accuracy	**0.88**
	Weighted Avg Precision	**0.80**	Weighted Avg Precision	**0.90**	Weighted Avg Precision	**0.88**
	Macro Avg Precision	**0.53**	Macro Avg Precision	**0.83**	Macro Avg Precision	**0.92**
	Weighted Avg Recall	**0.80**	Weighted Avg Recall	**0.90**	Weighted Avg Recall	**0.88**
	Macro Avg Recall	**0.53**	Macro Avg Recall	**0.86**	Macro Avg Recall	**0.92**
	Weighted Avg F1 Score	**0.80**	Weighted Avg F1 Score	**0.90**	Weighted Avg F1 Score	**0.88**
	Macro Avg F1 Score	**0.53**	Macro Avg F1 Score	**0.84**	Macro Avg F1 Score	**0.92**
	Metric	**Test Set**	**Metric**	**Test Set**	**Metric**	**Test Set**
Neural Network	Accuracy	**0.82**	Accuracy	**0.88**	Accuracy	**0.86**
	Weighted Avg Precision	**0.82**	Weighted Avg Precision	**0.88**	Weighted Avg Precision	**0.86**
	Macro Avg Precision	**0.55**	Macro Avg Precision	**0.77**	Macro Avg Precision	**0.90**
	Weighted Avg Recall	**0.82**	Weighted Avg Recall	**0.88**	Weighted Avg Recall	**0.86**
	Macro Avg Recall	**0.55**	Macro Avg Recall	**0.75**	Macro Avg Recall	**0.74**
	Weighted Avg F1 Score	**0.82**	Weighted Avg F1 Score	**0.88**	Weighted Avg F1 Score	**0.86**
	Macro Avg F1 Score	**0.55**	Macro Avg F1 Score	**0.76**	Macro Avg F1 Score	**0.79**

***Table 7.15**. Complete Results of the Classifiers' Performance*



High Ranked Universitites

		Precision	Recall	F1-score	Support
Logistic Regression	retweet	0,75	0,76	0,76	919
	like	0,79	0,77	0,78	1018
	reply	0,00	0,00	0,00	0
	accuracy			0,77	1937
	macro avg	0,51	0,51	0,51	1937
	weighted avg	0,77	0,77	0,77	1937
Random Forest	retweet	0,78	0,72	0,75	914
	like	0,77	0,82	0,79	1022
	reply	0,00	0,00	0,00	1
	accuracy			0,77	1937
	macro avg	0,52	0,51	0,51	1937
	weighted avg	0,77	0,77	0,77	1937
LGBM	retweet	0,77	0,82	0,79	919
	like	0,83	0,78	0,80	1018
	reply	0,00	0,00	0,00	0
	accuracy			0,80	1937
	macro avg	0,53	0,53	0,53	1937
	weighted avg	0,80	0,80	0,80	1937
Neural Networks	retweet	0,82	0,79	0,80	589
	like	0,82	0,85	0,84	702
	reply	0,00	0,00	0,00	1
	accuracy			0,82	1292
	macro avg	0,55	0,55	0,55	1292
	weighted avg	0,82	0,82	0,82	1292

Middle Ranked Universitites

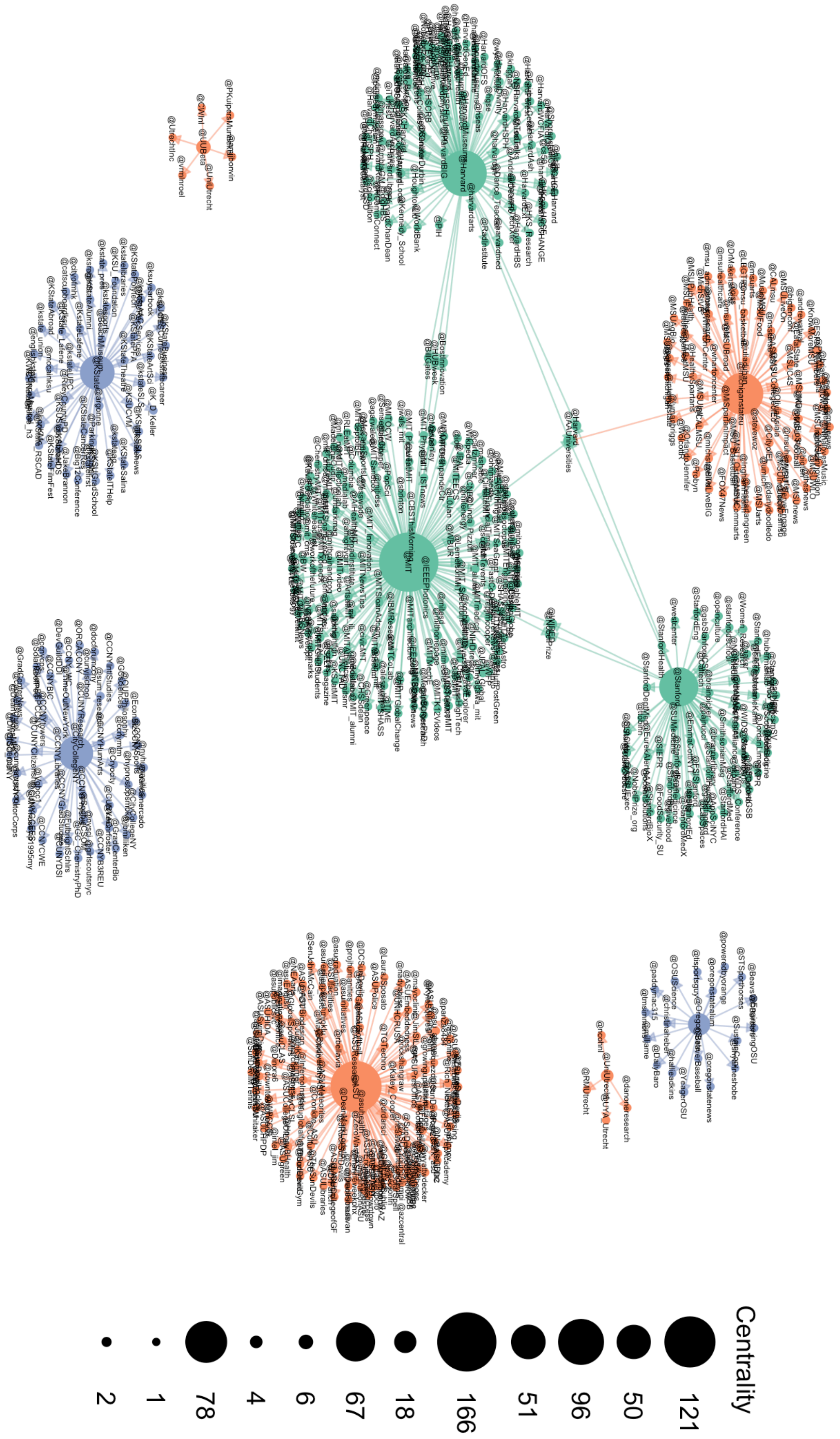
		Precision	Recall	F1-score	Support
Logistic Regression	retweet	0,96	0,94	0,95	90
	like	0,92	0,95	0,93	104
	reply	0,60	0,38	0,46	8
	accuracy			0,93	202
	macro avg	0,82	0,76	0,78	202
	weighted avg	0,92	0,93	0,92	202
Random Forest	retweet	0,99	0,86	0,92	90
	like	0,84	0,99	0,91	104
	reply	1,00	0,12	0,22	8
	accuracy			0,90	202
	macro avg	0,94	0,66	0,68	202
	weighted avg	0,91	0,90	0,88	202
LGBM	retweet	0,92	0,87	0,89	89
	like	0,89	0,93	0,91	109
	reply	0,67	0,80	0,73	5
	accuracy			0,90	203
	macro avg	0,83	0,86	0,84	203
	weighted avg	0,90	0,90	0,90	203
Neural Networks	retweet	0,95	0,79	0,87	78
	like	0,85	0,95	0,90	118
	reply	0,50	0,50	0,50	6
	accuracy			0,88	202
	macro avg	0,77	0,75	0,76	202
	weighted avg	0,88	0,88	0,88	202

Middle-Low Ranked Universitites

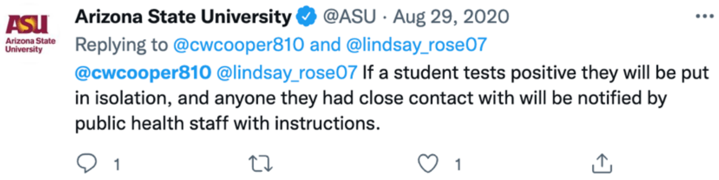
		Precision	Recall	F1-score	Support
Logistic Regression	retweet	0,90	0,88	0,89	51
	like	0,90	0,90	0,90	63
	reply	0,50	1,00	0,67	1
	accuracy			0,90	115
	macro avg	0,77	0,93	0,82	115
	weighted avg	0,90	0,90	0,90	115
Random Forest	retweet	0,96	0,88	0,92	51
	like	0,90	0,97	0,93	63
	reply	0,00	0,00	0,00	1
	accuracy			0,92	115
	macro avg	0,62	0,62	0,62	115
	weighted avg	0,92	0,92	0,92	115
LGBM	retweet	0,86	0,86	0,86	51
	like	0,89	0,89	0,89	63
	reply	1,00	1,00	1,00	1
	accuracy			0,88	115
	macro avg	0,92	0,92	0,92	115
	weighted avg	0,88	0,88	0,88	115
Neural Networks	retweet	0,84	0,82	0,83	44
	like	0,87	0,90	0,89	69
	reply	1,00	0,50	0,67	2
	accuracy			0,86	115
	macro avg	0,90	0,74	0,79	115
	weighted avg	0,86	0,86	0,86	115

***Table 7.16**. Confusion matrix from the Classifiers Results*

7.2. Attachments



7.2.1. Tweets Examples

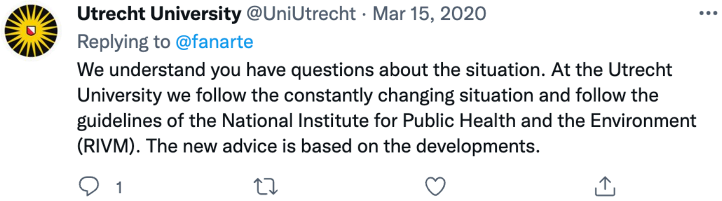

Arizona State University @ASU · Aug 29, 2020

Replying to @cwcooper810 and @lindsay_rose07

@cwcooper810 @lindsay_rose07 If a student tests positive they will be put in isolation, and anyone they had close contact with will be notified by public health staff with instructions.

1 1

***Figure 15**. Tweet from the Arizona State University*

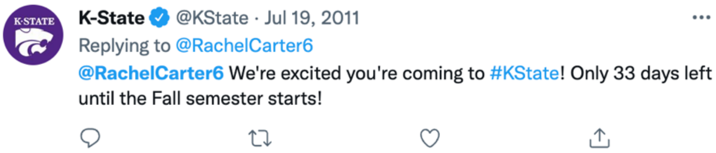

Utrecht University @UniUtrecht · Mar 15, 2020

Replying to @fanarte

We understand you have questions about the situation. At the Utrecht University we follow the constantly changing situation and follow the guidelines of the National Institute for Public Health and the Environment (RIVM). The new advice is based on the developments.

1

***Figure 16**. Tweet from Utrecht University*


K-State @KState · Jul 19, 2011

Replying to @RachelCarter6

@RachelCarter6 We're excited you're coming to #KState! Only 33 days left until the Fall semester starts!

***Figure 17**. Tweet from Kansas State University*