University of Zurich[UZH]

Uber Drivers, Activists and Investors: Community Detection and Analysis on Twitter

Tim Strasser
Zurich, Switzerland
12-742-573

Supervisor: Prof. Dr. Anikó Hannák
Date of Submission: October 1st, 2020

MASTER THESIS – Social Computing Group, Prof. Dr. Anikó Hannák

University of Zurich
Department of Informatics (IFI)
Andreasstrasse 15, CH-8050 Zurich, Switzerland

ifi

Master Thesis
Social Computing Group
Department of Informatics (IFI)
University of Zurich
Andreasstrasse 15, CH-8050 Zurich Switzerland
URL: https://www.ifi.uzh.ch/en/scg.html

Abstract

The widespread use and success of platforms of the so-called sharing or gig economy have led to a growing amount of workers whose employment status is often defined by short-term commitments and temporary contracts. Ridesharing companies like Uber and Lyft promote flexible and attractive side-job opportunities, while in reality, their drivers are frequently underpaid, bereft of social security, not enrolled in insurance plans, and report long hours and mental health problems. As opposed to employees of traditional taxi companies, Uber and Lyft workers also have fewer tools to communicate with each other and are only very recently seeing political support of specialized labor unions and organizations advocating for better working conditions. Online social networks like Twitter pose an opportunity for drivers and their organizations to join forces and organize, communicate more effectively, and create a collective identity they naturally lack as independent contractors. In this thesis, I collect and analyze publicly available Twitter data using hashtags related to the IPO (Initial Public Offering) of the ridesharing company Uber on May 10, 2019, and the strikes and protests preceding it. Using a combination of dimensionality reduction via PCA (Principal Component Analysis), k-means clustering, and NLP (Natural Language Processing) techniques, I characterize different communities consisting of news outlets, activists, worker unions, investors or drivers. In further analysis, I identify differences in hashtag usage patterns related to real-life events, topics discussed, and show correlation patterns between expressed sentiment and the number of retweets. Discussions of Uber-related cultural, political, and legal conflicts can be observed in tweets with high emotional scores. The goal of this work is to complement existing branches of sociological, legal, and algorithmic fairness literature.

Die weit verbreitete Nutzung und der Erfolg von Plattformen der so genannten Sharing- oder Gig-Economy hat zu einer wachsenden Zahl von Arbeitnehmenden geführt, deren Beschäftigungsstatus häufig durch kurzfristige Verpflichtungen und zeitlich beschränkte Arbeitsverträge definiert ist. Firmen wie Uber und Lyft werben mit flexiblen und attraktiven Nebenjobs. In Wirklichkeit sind Fahrer:innen häufig unterbezahlt, nicht versichert und melden lange Arbeitszeiten und psychische Probleme. Im Gegensatz zu den angestellten Personen traditioneller Taxiunternehmen haben die Arbeiter:innen von Uber und Lyft weniger Möglichkeiten, miteinander zu kommunizieren, und erst seit kurzem erhalten sie politische Unterstützung von Gewerkschaften und Organisationen, die sich für bessere Arbeitsbedingungen einsetzen. Soziale Netzwerke wie Twitter bieten den Fahrern und ihren Organisationen die Möglichkeit, ihre Kräfte zu bündeln und sich zu organisieren, effektiver zu kommunizieren und eine kollektive Identität zu schaffen, die ihnen als unabhängige Vertragspartner natürlicherweise fehlt. In dieser Arbeit sammle und analysiere ich öffentlich zugängliche Twitter-Daten mit Hilfe von Hashtags im Zusammenhang mit dem IPO (Initial Public Offering) der Ridesharing-Firma Uber am 10ten Mai 2019 und den vorangegangenen Streiks und Protesten. Mit einer Kombination aus Dimensionality Reduction mittels PCP (Principal Component Analysis), K-Means Clustering und NLP (Natural Language Processing) charakterisiere ich verschiedene Gruppen von Nutzer:innen, die aus Nachrichtenagenturen, Aktivist:innen, Gewerkschaften, Investor:innen oder Fahrer:innen bestehen. In einer weiteren Analyse identifiziere ich Unterschiede in den Hashtag-Verwendungsmustern im Zusammenhang mit realen Ereignissen und diskutierten Themen und zeige Korrelationsmuster zwischen ausgedrücktem Sentiment und der Anzahl der Retweets. Eine Diskussion über Uber-bezogene kulturelle, politische und rechtliche Konflikte kann in Tweets mit hohen emotionalem Score beobachtet werden. Ziel dieser Arbeit ist es, bestehende Literatur im Bereich Soziologie, Recht und algorithmische Fairness zu ergänzen.

Acknowledgments

I would like to thank my supervisor Prof. Dr. Anikó Hannák from the Social Computing Group at the Department of Informatics at UZH and Eszter Bokányi, research fellow at the CIAS Neti Lab of Corvinus University Budapest, for their continuous support, guidance, feedback and valuable input throughout the whole development of this thesis.

Contents

Chapter 1

Introduction

“Gig economy“, “platform economy“ or “platform capitalism“: Several terms seem to exist to describe new interrelated business models adopted by companies like Amazon, Airbnb, Uber, or Baidu. The former is used to describe forms of temporary work and online labor markets that act as a mediator between people offering and demanding a variety of services [1], including accommodation (AirBnb), caregiving (Care.com), food delivery (Foodora), mini-tasks (Amazon Mechanical Turk) as well as creative work and design (99Designs). Statistics show the increasing importance of such companies: 36% of U.S. workers have participated in the gig economy through either a primary or a secondary job [2]. The share of gig workers in U.S. businesses has risen by 15% since 2010 [3].

Many who are enthusiastic about the success of the platform and gig economy argue that these firms serve the interest of increasing productivity, reducing costs, raising efficiency, and furthering the general development towards a digital economy. Companies themselves often emphasize supposed benefits for customers as well as workers. Uber for example promotes working for its ride-hailing service by emphasizing the freedom to “earn anytime, anywhere“, “choose how and when you want to get paid“ and “set your own schedule“ while presenting alleged upsides of working with “no office or boss“ [4]. Others highlight negative outcomes of a new non-traditional labor market characterized by short-term working relationships, severe occupational and platform-based vulnerabilities, information asymmetries, a culture of surveillance, and forms of precarity such as a lack of health insurance, collective bargaining, and career training as well as harmful effects on mental health [2, 5, 6, 7].

One specific group of gig workers, namely drivers of ridesharing companies such as Uber and Lyft, has increasingly been the topic of discussion and research. A study found that the median hourly income of Uber drivers is $8.55 with 54% of drivers earning less than the U.S. minimum wage while 8% of drivers are actually making a net loss [8]. Another source says that the median monthly income for Uber drivers lies at $155, lower than that of Lyft drivers at $210 per month. 45% of Uber drivers make $0 to $99 per month and 84.0% of them see a monthly revenue of less than $500 [9]. Drivers said in interviews that they feel “left behind“ and bemoan financial straits partly because Uber “[charges] anything they want“ while reporting 70 hours of work a week and an income of $1,200 [10, 11, 12, 13].

Gig companies have also been criticized for their tendency to exploit regulative loopholes when entering a market. After it was founded in March 2009, Uber has received cease-and-desist orders in multiple cases due to its operation in a legal grey area [14, 15, 16]. In London, Uber's license has been removed by Transport for London in 2017. In 2018 it has been given a 15-month license. Another one has been issued in September 2019, which expired in November 2019 [17]. In 2015, Germany, the company's largest European market, banned UberPop, the service where customers are matched with unlicensed drivers using their private cars, by court ruling [18]. The legal position of ridesharing drivers has been discussed continuously, with several courts or judges ruling that drivers are in fact employees entitled to paid holidays and a minimum wage [12, 19, 20].

In May 2019, Uber went public at an event that some described as "the biggest tech IPO in years" [21], with a valuation of $120 billion earlier that year. Thousands of Uber and Lyft drivers struck ahead of the company's public offering in major cities including New York, Los Angeles, San Francisco, Chicago as well as in cities in the United Kingdom and Australia [22, 23, 24].

Online social networks in general and Twitter have increasingly been used in the context of political activism, communication, and organizing due to their ability of building networks, reaching new supporters, and quickly spreading information. For ridesharing drivers and organizations which represent them, social networks like Twitter pose an opportunity to join forces, organize and communicate more effectively and create a collective identity they naturally lack as independent contractors. Additionally, monitoring activity on Twitter is more and more of importance for companies due to social media platforms' rise in popularity for users to state their experience and opinions as consumers.

For this thesis, I collected publicly available Twitter data published around the date of Uber's IPO and used NLP techniques and clustering methods to characterize different communities consisting of journalists, investors, activists, driver organizations, unions, and drivers. In further analysis, I identified and interpreted differences in their usage patterns, expressed language, and sentiment. My results show connections between hashtag-related Twitter activity and real-life events. Using valence aware sentiment analysis, I identified a correlation pattern between expressed sentiment and the amount of retweets users receive. Expressions of Uber-related cultural and political conflicts and controversies can be observed in high-valence tweets.

Chapter 2

Background

In this chapter, I give a more detailed insight into the relevant topics of this thesis beyond the scope of the introduction. In Section 2.1, I will describe Uber as a company and give an overview on discussions around the legal position of drivers, the company's business model, safety concerns surrounding sexual harassment and rape allegations against Uber drivers, and accidents involving Uber vehicles, as well as the company's involvement in controversies around its internal work culture. In Section 2.2, I will cover the event of Uber's IPO on May 10, 2019, and some of the company's financial numbers. Finally, in Section 2.3, I will describe the strikes and protests on the days before the IPO and introduce some of the unions and other organizations organizing them and participating in them.

2.1 Uber

Since its founding in 2010, Uber has continuously been the focus of many different controversies and has been criticized for unfairly treating its drivers who often find themselves struggling financially and working long hours. A 2018 survey, answered by more than 2600 Uber drivers, reports a median net income without tips of $13.70, rising only to $14.73 including tips [25]. The survey also reports that 46.4% of drivers stay with the company for less than a year and only 11% drive for 3 years or more. Uber has repeatedly communicated that the flexibility they offer to drivers is incompatible with the drivers' demand to be classified as employees rather than independent contractors, stating that "the laws governing employers require [them] to exert much more control over their employees, monitor, make sure they're taking break times" [26]. In its Security and Exchange filing (S-1 form) filed in April 2019, Uber states that "if (...) we are required to classify Drivers as employees (...) we would incur significant additional expenses for compensating Drivers, potentially including expenses associated with the application of wage and hour laws (including minimum wage, overtime, and meal and rest period requirements)", that these circumstances would "require [them] to fundamentally change [their] business model" and that "[as they] aim to reduce Driver incentives to improve our financial performance, we expect Driver dissatisfaction will generally increase" [27] Uber's CEO Dara Khosrowshahi

has recently authored an opinion piece in which he said that if the company was forced to treat its drivers as employees, the company would only be able to offer full-time jobs to selected drivers in a few cities. He claims that rides would become more expensive while quoting surveys that show that a majority of Uber drivers do not want to be employees. Khosrowshahi then proposes a third way of classifying workers in general emphasizing the need for a new legal framework [28].

Uber has faced various legal challenges since 2009, such as cease-and-desist orders in several cities, including San Franciso in 2010 [14], Portland in [15] as well as South Carolina [16]. Its UberPop service, which matched customers with drivers using their own private vehicles with no requirement for a professional license, has been declared illegal in many countries. UberPop has been discontinued in Zurich as of 2017 and in all of Switzerland as of June 1, 2018, [29, 30]. Uber has also lost legal appeals over drivers' employment rights, meaning that drivers are entitled to paid holidays and a minimum wage and the company has been obligated to pay $20 million in settlements to drivers in California and Massachusetts [19, 20]. In France, Uber suspended UberPOP following violent protests at which taxi drivers flipped over and burnt Uber cars [31]. Protests did not stop, however, as illustrated by events in 2016, where French taxi drivers accused the company of "economic terrorism", burned tires, and blocked traffic in Paris and other French cities, forcing the police to use tear gas against 1,500 protesters [32]. Uber has been facing especially vigorous resistance from taxi drivers in Brazil: In 2016, Rio de Janeiro's mayor signed a bill banning the company in the city, reacting to growing sentiments against Uber drivers. Taxi employees attacked and kidnapped an Uber driver in São Paulo and the head of São Paulo's taxi syndicate told council members that "someone is going to die." [33] In 2016, a protest in Guadalajara, Mexico, turned violent as taxi drivers got in fights with local authorities and Uber supporters after blocking streets with their vehicles [34].

The company has been part of several discussions around safety concerns. Media has raised alarm about the numbers in the company's first safety report published in December 2019, emphasizing the reported 5,981 instances of sexual assault (by drivers against customers), including 464 cases of rape from 2017 to 2018. Uber also reported 97 fatal accidents and 107 total deaths. Although, due to the lack of comprehensive data and its interpretation, it is unclear whether Uber's service can be considered more dangerous than that of traditional taxis [35, 36].

Uber has been involved in multiple controversies, including former CEO Travis Kalanick's participatory role in Donald Trump's advisory council, his resignation after criticism, and his awareness but passiveness about internal sexual harassment issues and toxic company culture [37, 38]. In 2017, the company faced a lawsuit by an Indian rape victim who sued the company for gaining access to her medical records whereas Uber executives suggested that her claim may be an attempt to sabotage its business in India [39]. In January 2017, the hashtag "#deleteuber" gained popularity on Twitter after Uber was accused of having used surge pricing in order to profit off of a taxi driver strike at the JFK airport. The strike was organized in solidarity as a response to Donald Trump's executive order, banning travelers from seven Muslim-majority countries from entering the U.S. [40]. In February 2019, the former Uber employee Susan Joy Fowler published a blog post on her website in which she outlined details of hostile work culture for female employees, reported cases of sexual harassment that she experienced herself, and described a passive

human resources department which was unwilling to punish offenders [41]. Shortly after its publication, Uber's then senior vice president of engineering was asked by Uber CEO Kalanick to resign after it became public that he did not disclose to the company that he left Google in 2016 after an internal investigation had found sexual harassment allegations against him as credible [42]. On June 20, 2017, Travis Kalanick stepped down as CEO in the face of pressure from Uber's shareholders. A week earlier he had taken an indefinite leave of absence after the publication of a report on the company's workplace culture that ultimately leads to 20 employees fired over harassment and discrimination [43, 44]. In August 2017, Dara Khosrowshahi became the new CEO with Kalanick remaining a board member [45, 46]. In early 2019, Khosrowshahi stated that the company's "moral compass was off" under Kalanick's lead [45].

2.2 Uber IPO

After the IPO of Uber's main competitor Lyft and the 20%-fall of Lyft's IPO price, discussions evolved around the valuation goal of Uber and dampened the enthusiasm for Uber's public offering. In April, it was rumored that the company aimed for a valuation of $100 billion with a price per share $48 and $55 [47]. Gross bookings increased by 11% in 2018, but because the company still lost a total of $3 billion and the core revenue growth declined by %1 [48], its own long-term target of being worth $12 trillion became hard to believe [48]. Ahead of its Initial Public Offering (IPO), Uber announced to pay up to $40,000 per driver, a total of $300 million to more than 1.1 million drivers, as a way to celebrate going public and reserved more than 5.4 million shares for drivers to purchase at the IPO price [49]. On May 10, 2019, Uber finally went public on the New York Stock exchange with a share price of $42 and a market cap of $69.7 billion, marking a 7.6% drop on the first day and well below the company's target range of $44 to $50 per share. Uber's IPO took place during difficult times for the U.S. stock market, as trade talks between the U.S. and China took place on the same day after Washington raised tariffs on Chinese goods. In June 2019, Uber traded at its initial IPO price of $45 per share for the first time since May 10 [50].

2.3 Unions, Strikes and Protests

Because of the classification of Uber drivers as independent contractors, they are unable to form labor unions [51]. Several groups under different legal forms have therefore been founded in the last few years, advocating for better conditions for drivers, such as "Rideshare Drivers United" in Los Angeles [52], the "Boston Independent Drivers Guild" [53], the "Chicago Rideshare Advocates" [54] or "Drive United" in Washington D.C., a collaboration between rideshare drivers and Metro DC Democratic Socialists [55]. Some organizations also have committed themselves to the rights of gig workers in general, such as the "Gig Workers Collective" [56] and "Gig Workers Rising" [56, 57].

Demands vary between the different organizations. Rideshare Drivers United advocate for a 10% cap on the commission for both Uber and Lyft. They demand that the companies

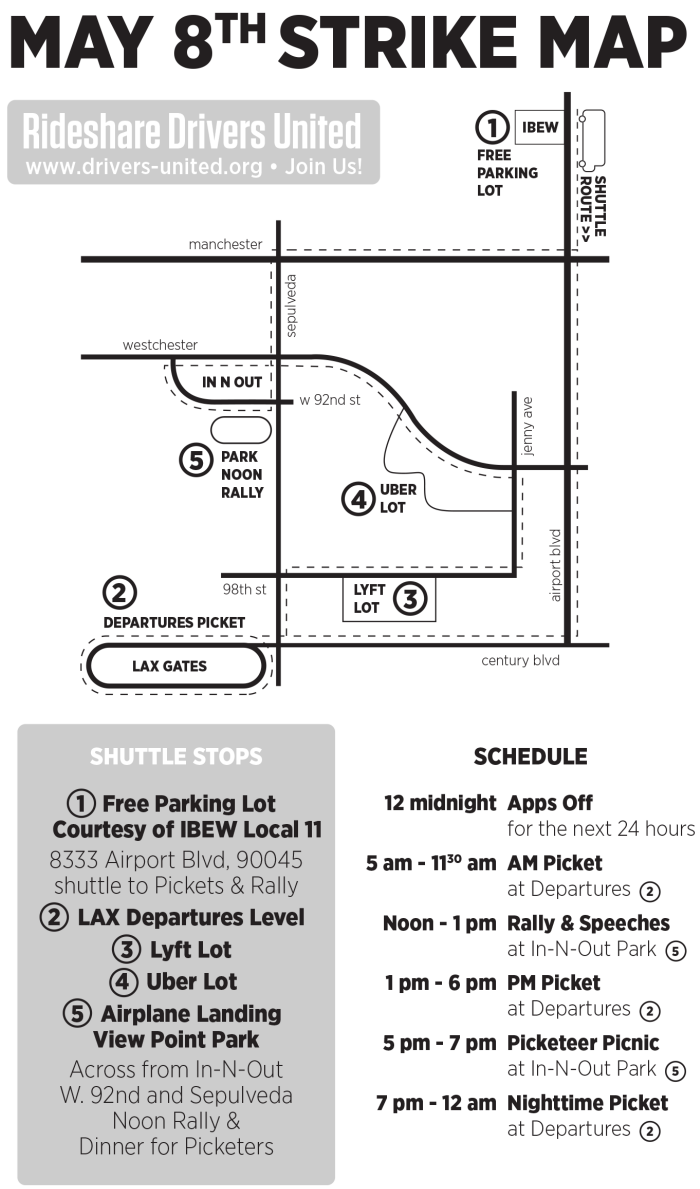


Figure 2.1: The "Strike Map" as posted by Rideshare Drivers United. Source: `https://drivers-united.org/strike`

should pay the drivers an hourly rate during their way to the passenger, suggesting an hourly minimum wage matching New York City's $27.86 before expenses. They also demand a transparent deactivation process, more transparency regarding the fare pricing and trip destination as well as a complete fare breakdown, a board of elected individuals representing drivers which should be appointed to Uber and Lyft's boards of directors, as well as emission standards for all new vehicles added to the platform [52]. Chicago Rideshare Advocates demand a 20% to 25% commission cap, transparent billing, vehicle inspections done by the city, a city oversight of driver account deactivations, and commercial umbrella insurance for all drivers provided by Uber and Lyft [54].

Before the date of Uber's IPO, these groups, among others, organized protests and strikes in various cities worldwide [51]. In New York, about 50 drivers protested the low wages together with supporters chanting "Driver power! Union power!" and saying that they are treated unfairly by ridesharing companies [58]. However, separate rallies seemed to compete with each other, as the Independent Drivers Guild (IDG) and the New York Taxi Drivers Alliance (NYTWA) took turns to protest in front of an Uber hub in Long Island City [59]. The IDG is co-founded by Uber and is the only union-associated group recognized by the ridesharing platform and is limited by a no-strike clause [59]. Uber denied the strike's alleged consequences, reporting a drop of only 500 drivers in New York. In Los Angeles, Rideshare Drivers United, the most militant group, announced the plan of a 24-hour strike with the participation of its estimated 4200 members [51, 59]. Figure 2.1 shows the strike plan published on the organization's website. In Chicago, the Chicago Rideshare Advocates has presented legislative demands including pay-rate increases, a limit on the number of ride-hailing vehicles, and an independent possibility for drivers to appeal suspensions and deactivation of accounts [58]. The streets in front of Uber's headquarters in San Francisco were blocked by hundreds of protesters carrying signs that stated that ridesharing companies "launch IPOs on the backs of their drivers." [58] In East London, protesters released smoke flares and shouted "Uber, Uber, you can't hide, we can see your greedy side." [58] Several politicians expressed their solidarity with the strikes and protests, including Bernie Sanders and Alexandria Ocasio-Cortez [60, 61].

Chapter 3

Related Work

Many studies have investigated aspects of Twitter usage, combining aspects of social research with computational methods. Research has focused on community detection among Twitter users, analyzing usage patterns of Twitter accounts, exploring how information is diffused within communities as well as on ways to analyze usage of language and how sentiments or opinions expressed in tweets. In this chapter, I present a selection of work related to this thesis, starting with research that focuses on topic and community detection. For topic and community detection, a distinction can be made between research that uses network measurements and graph theory to analyze social links between Twitter users, and work that uses the language information available in tweets, linking linguistic topics to user communities. Secondly, I present a selection of research that focuses more on discourse analysis, analyzing language patterns and sentiment, as found predominantly in social research.

Topic and Community Detection

Several methods have been proposed for the task of topic and community detection. Approaches include social network analysis, through applying graph theory or by applying topic detection and expecting communities to share common topics and talking points, the approach is also used by me in this thesis.

As an example for the former approach, Theocharis [62] analyzed the impact of the deactivation of a single account on an activist communication network on Twitter, formed by students in the United Kingdom who protested against the government's deficit reduction policies and used Twitter to spread information about events and to organize. Theocharis focused on classic social network analysis (SNA) of the university occupation network by collecting a list of followers of several university Twitter accounts and constructing their connections. He generated a symmetric adjacency matrix that he entered into social network analysis software which calculated several SNA measures like degree, betweenness, centrality, and closeness. His results show that the network is very robust and that the account of the occupation at University College London (UCL) is the most central one of the entire occupations network. Most of the accounts could be directly reached by one tweet. The accounts could directly reach others among themselves, yet the network centralization turned out to be very low, indicating a high resiliency against node deletion.

Figure 3.1 shows a graph visualization of the university occupation network. Similarly, Hachaj and Ogiela [63] present a community graph generation method that detects Twitter communities based on hashtag usage.

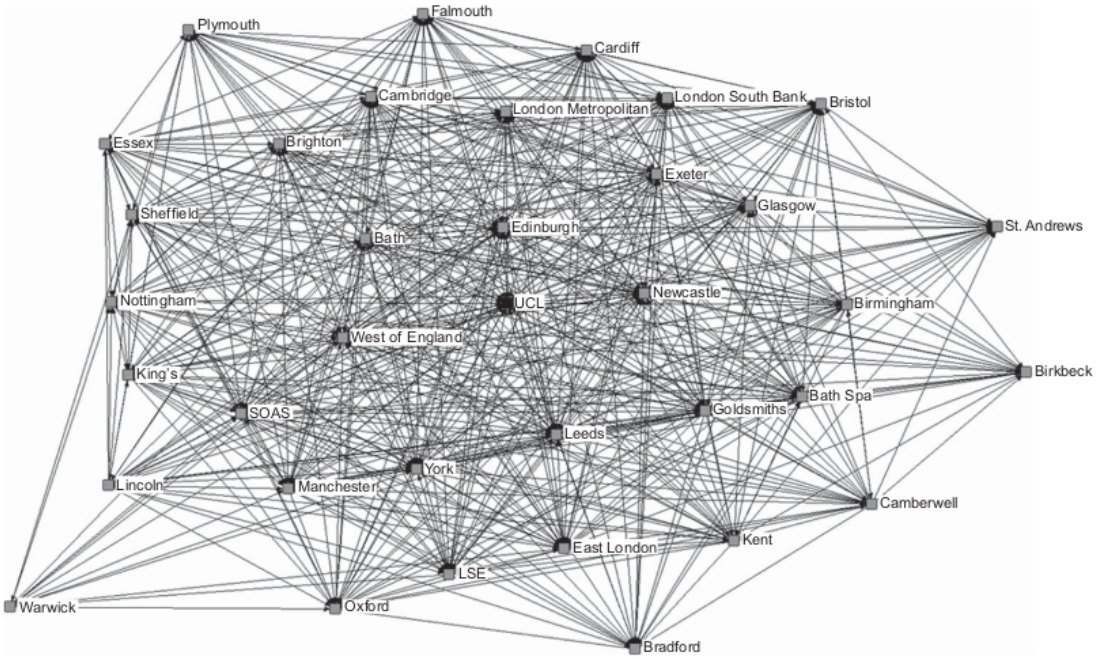


Figure 3.1: Network of the university occupation accounts, as created by Theocharis [62].

Vargas-Calderon et al. [64] used a method falling into the second category of approaches. They propose a way to automatically extract Twitter communities using a Word2Vec model [65], natural language processing techniques, and clustering algorithms, presenting a basis for unsupervised community detection. After collecting 2,634,176 tweets in six months, they discarded short texts (although on a tweet level, not on a user level as in the approach of thesis) and represented the set of tweets corresponding to each user as a vector in a vector space of 150 dimensions by applying the Word2Vec model [66]. They then used the gap statistic [67] to estimate the number of clusters that can be formed. Like in this thesis's approach, they applied k-means clustering on the vector representation of the documents. In contrast to my usage of PCA in this thesis, Vargas-Calderon et al. used PCA after clustering to identify the fifteen most representative documents. Finally, they built a frequency distribution of words for each cluster by taking the most representative documents, identifying each cluster by its most frequent words, characterizing a topic. In this thesis, I apply a similar approach, using the "Shifterator" library to illustrate differences in word frequency between the identified clusters.

Benny et al. [65] propose a method that collects tweets that use a specific keyword and summarizes them to find related topics. Similar to the method I used in this thesis, they used a pattern-based approach but propose a novel variation that uses Associated Gravity Force (AGF) and "Concept for the Imitation of the Mental Ability of Word Association" (CIMAWA) [68] after constructing a tf-idf matrix.

Rafea et al. [69] use an approach very similar to the method I used in this thesis to extract Arabic trending topics on Twitter, using a bisecting k-means clustering algorithm on word

unigrams.

In research that analyzes public health, social media data has previously mainly used for surveys and registries or for data mining applications such as influenza surveillance or predicting mental health disorders through language analysis. Surian et al. [70] used a combination of community detection and topic modeling methods, with their motivation coming from public health education. They explored the spread of healthcare-related information in online communities to help understand the geographical variation in decision making related to health outcomes. Using a data set of 285,417 tweets posted by 101,519 users between October 2013 and October 2015, they evaluated the opinions about human papillomavirus (HPV) vaccines. They used a combined approach of Latent Dirichlet Allocation (LDA), Dirichlet Multinomial Mixture (DMM) models for topic detection, the Louvain algorithm for community agglomeration, and Infomap for the encoding of random walks to detect community structure. Similar to the identification of the three clusters of users in this thesis, Surian et al. selected three representative topics to illustrate the alignment between community structure and language. The first topic contained mainly clinical and scientific evidence, the second topic contained users that post about their personal experiences, and the third topic contained links to anti-vaccine websites. They concluded that the use of community detection in combination with topic modeling appears to be a useful way to characterize Twitter communities for analyzing opinions and that it might present a possibility to help find communities that are influenced by negative opinions about vaccines.

Adewole et al. [71] focused on spam account detection on Twitter. They propose a method combining PCA and k-means clustering to identify the clusters of spammers in over 200,000 Twitter accounts randomly selected from over 200,000 accounts. As in this thesis, "PCA is proposed to improve the clustering process of the tuned k-means algorithm [...]" [71] The use of PCA and k-means shows promising results with evidence of malicious interactions in the identified clusters of spammers. Based on the labeled collection of over 25,000 accounts, they trained three classification algorithms: Multilayer Perceptron, Support Vector Machine (SVA), and Random Forest, with the Random Forest classifier outperforming the others with an accuracy of 96.20/

Bokányi et al. [72] explored how demography and community are represented in the language used on Twitter, using an approach based on word frequencies, similar to the language-based analysis I used. They collected a data set of 335 million geotagged tweets from the United States posted between February 2012 and June 2013 by using the "Filtered Stream" API by Twitter (see Subsubsection 4.1.2). They used a hierarchical triangular mesh scheme to geographically index the data and assign a US county to each tweet. They then modeled the data set as a vector space where documents correspond to the tweets located in a county. Any further analysis was limited to a bag-of-words approach based on word frequencies, without taking inter-word and sentence-based relations into account. This is similar to the per-user document approach I used in this thesis (see Section 4.2. As in my approach, they computed a term-document matrix and applied filtering by only including counties that contain at least 10,000 occurrences of at least 500 individual words and remove stop words in several languages. The matrix was normalized by the inverse document frequency. For further data analysis they used a combination of a Robust Principal Component Analysis (PCA) (see Subsection 4.3.2 for an explanation

of PCA), similar to this thesis's approach, and Latent Semantic Analysis to identify the main sources of variations and thus the most significant topics and linguistic features in the data. Bokányi et al. were able to find significant spatial correlations between geographic information and language usage patterns such as the use of slang or words related to different lifestyle activities, travel, religion, and ethnicity. They stated that "Geographical proximity is thus a main driving force in the similarity of language patterns in Twitter-space." They also used various sources of demographic data sources to correlate with their data model. Bokányi et al. found that patterns in language do not only relate to geographic proximity, but also socioeconomic and cultural factors, such as religion or ethnicity, for example, that the use of slang is higher in regions with a presence of Afro-American communities.

Language and Sentiment Analysis

In addition to research focusing on community and topic detection, many studies, predominantly from the fields of social and political sciences, have been focusing on specifically analyzing the use of language on Twitter. They often combine sentiment analysis with a qualitative analysis including coding and reading of selected tweets, to understand the formation and expression of collective group identity as well as how sentiments and political opinions are presented, often in the context of elections, protests, or other real-life events, similarly to my approach.

For example, Lommel et al. [73] focused on the impact of two political events, similar to my focus in this thesis on the two events of the Uber IPO and the strikes on the previous day: The inauguration of Donald Trump as the 45th president of the United States on January 21, 2017, and the worldwide Women's March on the following day, in which over 100,000 people participated in New York alone [74]. Organizers of the Women's March called for a "Day Without Woman" strike of women in both paid and unpaid work. Lommel et al. looked at the group of protesters intending to analyze how feminist identity expressed itself within the online communities. They put their research in the context of contemporary feminism, as opposed to feminism defined by either one of the "three waves" [75].

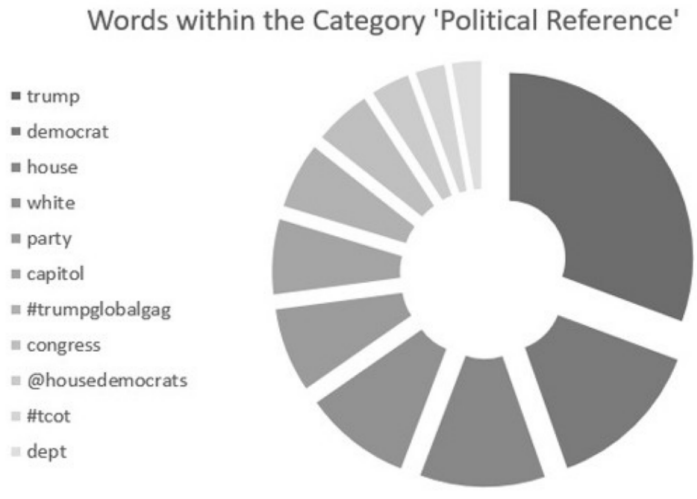


Figure 3.2: The words Lommel et al. coded into the "Political Reference" category [73].

Similar to my approach, Lommel et al. combined a text mining approach with an in-depth qualitative analysis of filtered and selected tweets. They used a web scraper tool called "Netlytic" to sample tweets from Washington DC, New York City, and Los Angeles, using a method very similar to the approach I used (as outlined in Section 4.1) and by Akerlund [76]: Screening Twitter manually to investigate which hashtags come up frequently and creating a list of search terms. They then decided on using four hashtags: "#WomensMarch", "#DayWithoutAWoman", "#WomensStrike" and "#IStrikeFor". Using a basic lexicometric method, they identified 148 words that occurred in at least seven different tweets. Lommel et al. then coded the identified words into categories of meaning and analyzed their frequency as shown in Figure 3.3. Figure 3.2 shows the words coded into the category "Political Reference". Following this data exploration approach, they explored the tweets using critical discourse analysis.

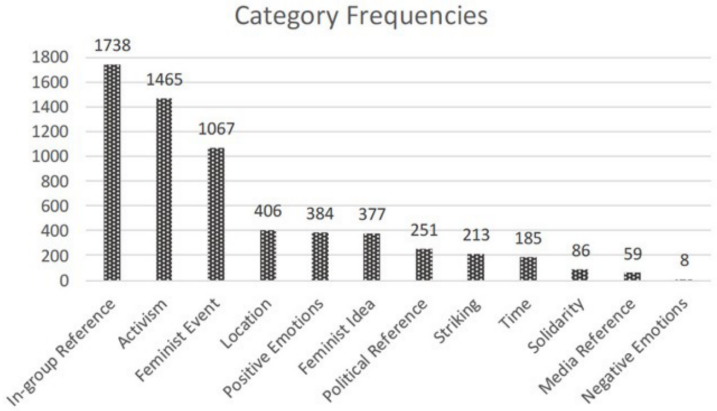


Figure 3.3: The categories identified by Lommel et al. and the freqencies with which words fall into them [73].

Their findings show that there existed a politically defined personalized group boundary, as well as perspectives of in-group identity. They found calls to equality in tweets that for example read "An equal society is a society for all". One Twitter user reported wearing a "Black Lives Matter" shirt together with a T-Shirt saying "Feminist AF", emphasizing an intersectional approach. tweets connected modern feminist claims to older feminist movements, for example saying that feminism is important "today as every day". Donald Trump and his politics were presented as being opposed to feminist demands, underlining the separation of the feminist movement and Trump rather than feminism and non-feminism. The data set showed users merely showing solidarity with the feminist movement as well as users identifying with it. However, as Lommel et al. stated, the line cannot be drawn exactly. There seemed to exist a space between "us" and "them" where users can be placed that did not show a complete group affiliation.

Fernández-Rovira et al. [77] conducted a content analysis within the framework of a comparative study that analyzed feminist positioning by both a male and a female candidate of each of the four main Spanish political parties. They collected tweets published in September and October 2018, with a total sample of 3,489 tweets posted by the eight politicians. In the case of the male representatives, they chose the general secretary or the president of the parties. For the female representatives, Fernández-Rovira et al. chose women according to their position within the part and with a similar media representation to the men. They used several quantitative and qualitative measures, similar to the ones

used by me, such as the number of tweets related to gender issues, the type of tweet, the use of hashtags, aspects related to feminism such as topic, use of feminist concepts, and use of inclusive language. They also analyzed whether the tweet's content is related to facts or news as well as the communicate intentionality of the tweet, meaning whether the tweet is taken as a statement, a request, or an offer for information and whether it states emotions or moral attitudes. They found that most of the tweets of all the parties do not discuss the issue of feminism. The hashtags most used by Podemos, a party associated with democratic socialism and left-wing populism [78], were related to gender violence such as "#violenciamachista" ("sexist violence") and "#MeToo". This was similar to the PSOE (Spanish Socialist Workers' Party), associated with Social Democracy [79], which mostly used "#violenciadegenero" ("gender violence") or "#NiUnaMenos" ("not one (woman) less"). Most of PSEO's tweets refer to feminist concepts, as opposed to other parties. The tweets of conservative politicians were usually related to the media agenda and tend to avoid feminist concepts. Their results also show that women posted more tweets related to feminism than men, however, the difference is marginal: 8% of tweets posted by women are about feminism compared to 5% of the men's tweets. Women seem to use tweets more often to offer information and to post moral attitudes. In contrast, men's tweets often contain intellectual and emotional attitudes.

Akerlund [76] analyzed the importance of Twitter as a platform for far-right discourse, aiming to understand its role in mainstreaming far-right sentiments. She researched in what ways influential users use Twitter's functions, how they used Twitter in contrast to other users, and how these patterns contributed to (re)producing far-right discourse. She focused on the period between September 2, 2018, until November 22, 2018, when the Swedish political system faced the difficulties of a hung parliament due to inconclusive election results. 131 days passed before a new government was formed. Akerlund used a combination of descriptive statistics, sentiment analysis, and close readings, similar to my approach. By using 23 hashtags, she collected 74,336 tweets, retweets, replies, and "quote tweets" posted by 6,809 users. The hashtags were sampled using the same method as I used as described in Section 4.1 of Chapter 4 and the one used by Lommel et al. [73]: Through a process of "snowball" sampling hashtags were found first via manual search using the Twitter website, in Akerlund's case starting with the hashtag "#SD" standing for the far-right party Sweden Democrats (SD). Additional hashtags were then selected via co-occurrence. To identify influential users in her data set, Akerlund used a measure in relation to the retweet function on Twitter, arguing that it shows the capability of a user to reach outside of his community. She used the "Valence Aware Dictionary and Sentiment Reasoner" (VADER) for sentiment analysis, as I did in my sentiment analysis. 2,598 tweets by influential users were also read and coded to find representation patterns of far-right discourse, in relation to the results found by the sentiment analysis. Akerlund found that the 14 remaining accounts of influential users all had a very narrow focus on far-right politics in terms of the content they post and their profile description. She concluded that influential users defined the conversation, rather than engaged in it. Her findings show that issues about immigration and crime were rated most negative by the sentiment analysis. Influential users described immigrant men as violent and criminals, rapists, and terrorists. Most surprisingly, the compound sentiment score showed that influential users tweeted in a very emotionally neutral way, which goes against the expectation of finding extreme language among these users. These users often used factual statements and direct

quotes of media articles together with open questions, provocations, and suggestions, with followers taking the implication further in the comments and posting hateful sentiments. This neutrality can be seen as a strategy to avoid getting banned by Twitter or as a way to appeal to a larger crowd, which further enables the mainstreaming of far-right narratives.

A similar kind of statistical and content analysis of discourse on Twitter was conducted by Hannak et al. [80], who analyzed misinformation in online communities. They examined the contexts and consequences of fact-checking interventions of Twitter and the social relationship between individual users who issue the fact-check and those whose statements are challenged. Using the term “to snope“ to describe the act of posting fact-checking replies to tweets, relating to the fact-checking website “Snopes.com“ [81], they explored the research questions “Who snopes whom?“, “Do snopes matter?“ and “Where do snopes happen?“. Using a corpus of tweets posted between January 2012 and August 2013, they collected a subset of 3,969 tweets that met the criteria of being a reply from a snoper and include an URL linking to either “Snopes.com“ or to the other fact-checking platforms “PolitiFact.com“ and “FactCheck.org“. They based their analysis on four distinct contexts of where snoping can happen: The snopee and the snoper mutually follow each other, the snopee was snoped by a follower, the snopee was snoped by a followee (as in the snoper follows the snoper), and finally, the snopee was snoped by a stranger. Hannak et al. also examined the structural positions of users by analyzing their social connections on Twitter and temporal relationships between snopees and snopers. Their findings show that the most typical form of snoping is the least effective: Fact-checking interventions are most often used by strangers acting as snopers, but they draw more user attention and responses when coming from friends of the snopee. Additionally, they found that corrections from strangers are better received by snopees if they happen within ongoing discussions rather than being posted as “out-of-the-blue“ interventions. Status and popularity also seem to play a role: Individuals with dense networks and highly-connected followers snope popular individuals with sparse social networks.

Chapter 4

Methods

In this chapter, I cover the methods that were chosen for data collection and analysis, leading to the results described in Chapter 5. First, I discuss the different approaches and decisions regarding data collection in Section 4.1. In Section 4.2, I cover the different data filtering methods I used in the work of this thesis. Finally, in Section 4.3, I cover the data analysis methods that I used, namely the construction of a tf-idf matrix as described in Subsection 4.3.1, the process of conducting a principal component analysis explained in Subsection 5.4.1, as well as performing k-means clustering as shown in Subsection 4.3.3. Finally, in Subsection 4.3.4, I cover the Shifterator library, while in Subsection 4.3.5 I describe sentiment analysis using VADER (Valence Aware Dictionary and sEntiment Reasoner).

4.1 Data Collection

One data collection method that I explored during the work of this thesis was collecting data by using the official Twitter API. As mentioned in Subsubsection 4.1.2, using the Twitter API comes with various limitations on download rate. I used the "Filtered Stream v1" API mentioned in Subsubsection 4.1.2 with a search rule consisting of a combination of keywords combined through an "OR" operator, meaning that any tweet containing one or more of the keywords or hashtags will get covered by the rule and included in the result returned by the Twitter API. I used parentheses to summarize multiple words that are separated by spaces as one search term. With an additional search rule, I ensured that only tweets written in English are included. I used the following keywords in the search rule:

- uber
- #uber
- uberdriver
- #uberdriver

- (uber driver)
- uberx
- #uberx
- uberride
- #uberride
- (uber ride)

I used geographical search rules in combination with a script from the "GeoSearch-Tweepy" repository available on Github [82], using it to download a collection of tweets that were posted within a bounding box roughly outlining the United States. In another approach, I collected tweets concerning a specific Uber-related event in one particular location and within a very short time frame, for example January 2017 when Uber was accused of profiting off the lower number of taxis operating due to a strike near JFK airport, leading to the rise of the "#DeleteUber" movement in which people on Twitter announced that they would delete the Uber mobile application from their phones.

I discarded both this approach and the one using a pre-collected data set because they either returned a low number of tweets with geospatial information or a low number of Uber-related tweets in general. I did not use the official Twitter API because of the limitations regarding download rates and the incompleteness of the data. Neither of the methods offered access to historical data. The fact that the pre-collected data set was collected in 2020 did also not only limit the research opportunities to that time frame, but it also consisted of many tweets that deal with how ridesharing platforms respond to the new problems posed by the COVID-19 pandemic, which would have posed a great challenge for data filtering.

I, therefore, collected the final data set using the Twint library described in Listing 1, focusing on the Uber initial public offering (IPO) on May 10, 2019, and the related strikes and protests that happened on the day before the IPO. On one hand, my intention was to narrow down the time frame and thus reduce the amount of data to a size that is feasible to analyze. On the other hand, I expected the data in this time frame to contain tweets by a variety of users including activists, Uber drivers, or users talking about investing in the Uber stock. This offered the possibility to analyze how these different user groups differ from each other in terms of engagement, speech, or sentiment. I decided to collect tweets that were posted between April 1, 2019, and May 31, 2019. This also enabled me to compare engagement to the day of the IPO, May 10, to engagement prior and post to the event. I included the terms in the search query by using a "snowball" approach. For example, by searching Twitter for tweets containing the words "uber" and "strike", I found many tweets that use the hashtag "#uberlyftstrike". In those tweets, I found tweets using "#driversunite" or "#uberstrike". I applied this process until no widely used hashtag could be detected anymore. I intentionally did not take hashtags such as "#workersunite" or "#generalstrike" into account for querying as they are too general on their own and using them would likely include tweets that are irrelevant to Uber in specific. I decided to not use the keyword "uber" as a search term, but rather as a hashtag, because it would have

included tweets in the result that use “uber“ as an adjective rather than to describe the company, posing a challenge in term of data filtering. Thus, the final search terms were:

- #uber
- #uberlyftstrike
- #uberstrike
- #ubershutdown
- #lyft
- #lyftstrike
- #uberipo
- #driversunite
- #strikeuberlyft

I used these hashtags non-exclusively, meaning that any tweet containing at least one of these hashtags were included in the data collection. This final approach yielded a total set of 51,853 tweets. The Python script that I used for data collection can be found in the appendix. The data set is stored as a text file in the “.txt“ format. Every line in the file consists of a single tweet stored in the JavaScript Object Notation format (JSON). An example can be seen in Listing 1.


```
{
    "id":1126905717477257216,
    "conversation_id":"1126905717477257216",
    "created_at":1557510234000,
    "date":"2019-05-10",
    "time":"19:43:54",
    "timezone":"+0200",
    "user_id":1653081355,
    "username":"pauljdub",
    "name":"Paul Warren",
    "place":"",
    "tweet":"I'm a couple days late to this, but I 100% respect the
    ↪  #UberLyftStrike \n\nI stopped driving about 6 months ago,
    ↪  because I had an undisputable report placed on my driving
    ↪  profile for refusing to take a minor to a liquor store to buy
    ↪  alcohol for them.",
    "language":"en",
    "mentions":[

    ],
    "urls":[
```

```
    ],
    "photos":[

    ],
    "replies_count":0,
    "retweets_count":0,
    "likes_count":9,
    "hashtags":[
        "#uberlyftstrike"
    ],
    "cashtags":[

    ],
    "link":"https://twitter.com/pauljdub/status/1126905717477257216",
    "retweet":false,
    "quote_url":"",
    "video":0,
    "near":"",
    "geo":"",
    "source":"",
    "user_rt_id":"",
    "user_rt":"",
    "retweet_id":"",
    "reply_to":[
        {
            "user_id":"1653081355",
            "username":"pauljdub"
        }
    ],
    "retweet_date":"",
    "translate":"",
    "trans_src":"",
    "trans_dest":""
}
```


Listing 1: A JSON file containing a tweet downloaded with the Twint library.

4.1.1 Geographical Restriction Rules

Both the Twitter API as well as the Twint Library support a restriction of their search by using geographical criteria such as bounding boxes, geographical codes, or names of places. According to Twitter documentation, there are three different ways to extract geographic metadata from a tweet:

1. tweet location:
   (a) Longitude and latitude pairs (-85.7629, 38.2267)
   (b) Twitter places with a name (‘Louisville Central‘)
   (c) Four pairs of latitude and longitude pairs defining a ‘bounding box‘
2. Mentioned location: parsing the tweet message for location information.
3. Profile location: parsing the account-level location for locations information

Twitter allows users to “geotag“ their tweets at the time of posting. This results in a location either defined by an exact point location, called a “Twitter Place“ with a bounding box of coordinates defining a larger area [83]. Listing 2 shows how the geographical information is structured.

```
{
    "coordinates":{

    },
    "place":{

    }
}
```

Listing 2: The structure of geographical information in a tweet.

In a geo-tagged tweet, the place object is always included. The coordinates only has a value assigned to it when an exact location was assigned to the tweet [83]. Twitter Place objects are named locations with corresponding coordinates. They are assigned, among other fields, an identification number (ID), a type of place such as “city“, a country code, and a bounding box consisting of coordinates. Tweets that contain a coordinates object with a “Point“ coordinate come from GPS enabled devices. Only one to two percent of all tweets are geotagged, drastically decreasing the number of tweets returned when filtering for tweets that are geotagged within a certain area.

Geospatial metadata can also be detected by analyzing the tweet content, looking for places mentioned by the user, such as in the sentence “Just landed at JFK.“ This method may involve natural language processing, including named entity recognition and detecting whether the word “JFK“ is used to describe the former president John F. Kennedy or the airport located in New York.

As a third way, location can be extracted from the user’s profile. Compared to geo-tagging and parsing the tweet content this method provides the largest source of geospatial metadata. It’s not an exact source of location information, as users are free to set their location to a text string of their choice, including “My parents’ place“ or “The most beautiful city in the world“.

4.1.2 Data Collection Methods

Twint Library

"TWINT - Twitter Intelligence Tool" is an open-source intelligence (OSINT) scraping tool written in Python that can be used for collecting data from Twitter without using the Twitter API and without any authentication or Twitter account by using Twitter's search operators [84]. It can fetch almost all tweets and can be used without any rate limitations, in contrast to the non-complete "Standard" plan of the Twitter API.

Twint offers several configuration options for constructing a query to use with its ".Search" function, including the possibility to define the desired time frame, filtering for tweets with geographical information, or for tweets containing email or phone number information. Besides the ".Search" functionality, Twint also offers functionality to scrape a user's timeline or favorite tweets. For these additional search functionalities, Twint can only retrieve about 3,200 tweets because Twitter limits the scrolls on a user's timeline. This thesis uses the ".Search" functionality together with limitations on the time frame, language and hashtags [85].

Pre-collected Data Sets

In addition to limiting the usage of its API, Twitter also prohibits the public distribution of collected tweets, which complicated the online search for a pre-collected data set. One such data set was made available to me by the Social Computing Group of the Department of Informatics at UZH. It consisted of two separate lists, one with a total of 32,220 tweets containing keywords relevant to Uber, another consisting of 2,148 tweets that use Uber-related hashtags, both collected in early 2020 by using the "Filtered Stream" API.

Twitter API

The official Twitter API was introduced in September 2006, with the interest to interact with the data rising immediately after the platform's launch earlier that year [86]. The API saw steady growth in usage and popularity between its launch in 2006 and 2010, with developers implementing several different projects on top of it, for example, a URL shortening service. In 2010, among other decisions that were received negatively by the developer community, Twitter introduced the requirement of using OAuth (an open standard for access delegation authentication) to access the API, leaving developers with only nine weeks of migration time [87]. Additionally it launched its own URL shortening service [88]. In 2012, Twitter continued to restrict the usage limits of its API, introducing a user token limit of 100'000, majorly affecting the usability of many applications using the API [89, 90].

Twitter now offers a wide range of API endpoints serving a variety of data, including accounts and users, tweets, direct messages, media, trends, and geographical information. It offers a "Search" API that enables developers to query tweets using a combination of

criteria and search keywords. Additionally, endpoints exist that offer subscription to a live stream of tweets with support for filtering. As part of a "Twitter Developer Labs" program that motivates developers to help shape new APIs, Twitter also offers a new version of its "Filtered Stream" endpoint [91]. The API in general is divided into three different levels of access. As can be seen in Figure 4.1, Twitter restricts access to historic data to its "Premium" and "Enterprise" plans. The "Standard" plan is limited to access to a sample of tweets that were posted roughly within the last seven days [92]). The "Standard" plan is designed to return tweets prioritized in relevance, access to a complete data set is limited to "Premium" and "Enterprise" plans [92].

v1.1 subscription levels

Tweets	Standard	Premium	Enterprise
Publish and engage	✓		
Search Tweets: 7-days	✓		
Search Tweets: 30-days		✓	✓
Search Tweets: Full-archive		✓	✓
Filter Tweets	✓		✓
Sample Tweets	✓		✓
Batch Tweets			✓
Direct Messages	✓		
Account and users	✓	✓	✓
Metrics			✓

Figure 4.1: Twitter Subscription levels [93]

Additionally, Twitter offers a free version access level (called "Sandbox") to the "30-day" and "Full-Archive" level of the "Search", both normally limited to "Premium" users, as can be seen in Figure 4.2. As Figure 4.2 shows, the "Sandbox" level access is limited to thirty requests per minute ("RPM") and ten requests per second ("RPS"), while the "Premium" level access allows to send 60 requests per minute. The "Standard" access level of the Twitter API is rate limited based on 15-minute windows [94].

Feature type	Sandbox	Premium
Timeframe	Last 30 days *or* full-archive	Last 30 days *or* full-archive
Tweets per data request	100	500
Tweet Counts endpoint	No	Yes
Query length	30-Day - 256 chars Full Archive - 128 chars	1024 chars
Operator availability	Sandbox	Premium
Rate limit	30 RPM, 10 RPS	60 RPM, 10 RPS
Enrichments	n/a	Expanded URLs, Profile Geo, Polls

Figure 4.2: Comparison of the Sandbox environment [95]

4.2 Data Filtering and Pre-Processing

In order to be able to use the collected data in later steps, I pre-processed the data using several approaches, including methods typical for natural language processing. The steps I undertook for text pre-processing were:

1. Converting the tweet to lowercase
2. Removing numbers
3. Removing URLs
4. Removing punctuation
5. Removing emojis
6. Tokenizing
7. Removing English stop words
8. Lemmatizing
9. Removing words containing non-alpha characters
10. Filtering out lemmas shorter than two characters

The code I used for pre-processing can be found in the appendix. I performed the tasks of tokenization, the removal of stop words, and lemmatizing using the Natural Language Toolkit library (NLTK) [96]. The "TweetTokenizer" is a class from the NLTK library that is specially designed to correctly tokenize tweets [97].

4.2.1 Document Length Distribution

After pre-processing using the mentioned steps, I further processed the data by extracting the unique user names in the data set and creating a dictionary that maps the usernames to their tweets. I concatenated the tweets to form a list of tokens collected from users' posts in April and May 2019. This list of tokens will also be referred to as a "document" in the rest of this thesis. There were 25,558 unique users in the data set. Using this dictionary, I was able to analyze the data set in regards to document length distribution, shown in Figure 4.3. It can be seen that the majority of the documents consist of fewer than seventy-five tokens. Assuming a correlation between the level of account activity and the length of the document, I concluded that the data set consists of many accounts that show a general low activity during the two months. For all further steps, I filtered the data set for users that have a minimum amount of 120 tokens. The filtering reduced the data set from 25,558 documents to 557 documents, corresponding to the top 2.17% in terms of document length.

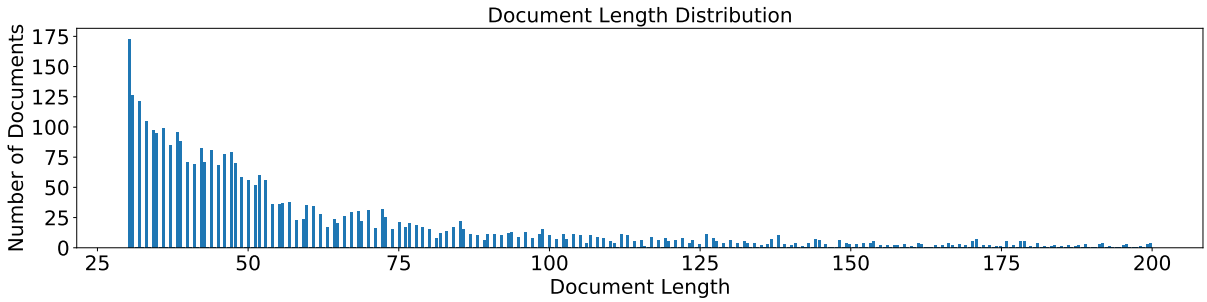


Figure 4.3: Distribution of document length in the data set.

4.2.2 Lexical Diversity

After analyzing and filtering for document length, I analyzed the remaining documents regarding their lexical diversity. Lexical diversity is a quantitative measure of the number of "different" words in a text, the key idea being the notion of "non-repetition". In the context of analyzing tweets, I assumed that documents with a low lexical diversity might generally be "spam" accounts that, for example, post promotional codes for the Uber mobile app, as can be seen in Figure 4.5, which shows a tweet by the account "uber4london". The account posts a link pointing to the Uber sign up page together with a promo code every day and has a very low lexical diversity of 0.075%, but one of the highest document length with 17,280 total tokens. This indicates that lexical diversity is a simple yet crucial measurement for filtering. Figure 4.4 shows the distribution of lexical diversity in the data. It should be noted that this figure only takes documents longer than 120 tokens into account. Filtering for lexical diversity resulted in a final set of 353 users/documents.

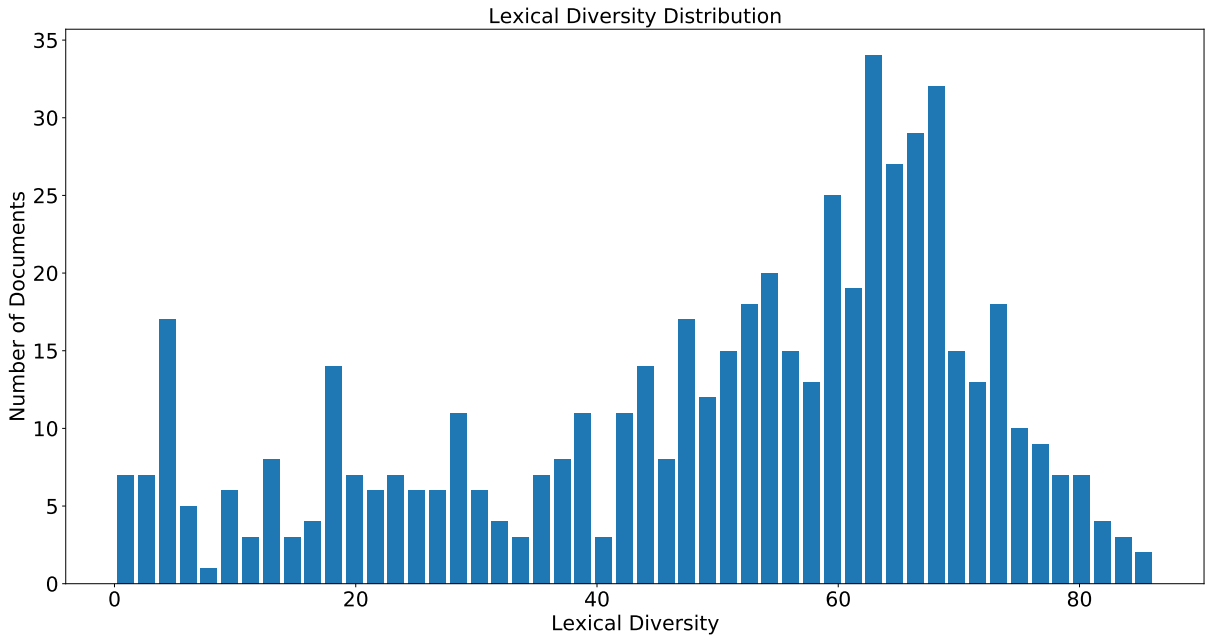


Figure 4.4: Distribution of lexical diversity in the filtered data set.

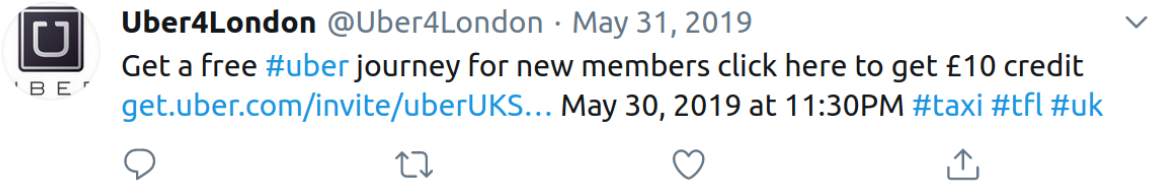


Figure 4.5: A sample ‘spam‘ tweet posted by the account “uber4london“.

4.3 Data Analysis

4.3.1 Tf-idf Matrix

In order to analyze the data sets, I clustered users to form communities that share a common vocabulary, with the intention of defining a cluster as a group of users with similar interests. The degree of cluster membership thus represents the extent to which users are integrated into a common topic.

Within this approach, a topic can be defined as a decomposition of a user’s tweets respectively a linear combination of the words occurring in the user’s text. This approach is generally used in topic detection and document clustering. In this approach, users are treated as documents in terms of the fact that their tweets are concatenated and treated as one text/document, as described in Section 4.2.

After performing the filtering and pre-processing steps, I created a term-document matrix using the term frequency-inverse document frequency metric (tf-idf). Tf-idf is a metric with the goal of reflecting the importance of a word in relation to a document in a corpus. It is defined as the frequency with which a term occurs in a document multiplied by the

	stock	strike	take	taxi	think	time	today	try	uberipo	uberlyftstrike
0	0.000000	0.276414	0.000000	0.143618	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	0.117880	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	0.000000	0.000000	0.227768	0.098295	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	0.031203	0.302998	0.027022	0.034985	0.000000	0.029394	0.000000	0.000000	0.000000	0.000000
4	0.000000	0.166965	0.000000	0.057834	0.000000	0.000000	0.054968	0.000000	0.000000	0.591819
5	0.167379	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.653958	0.000000
6	0.506347	0.156090	0.125283	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.092212
7	0.000000	0.000000	0.000000	0.000000	0.111235	0.000000	0.000000	0.000000	0.000000	0.000000
8	0.000000	0.000000	0.058835	0.076172	0.000000	0.000000	0.000000	0.175988	0.000000	0.000000
9	0.208506	0.000000	0.090282	0.000000	0.000000	0.000000	0.111093	0.000000	0.000000	0.000000

Table 4.1: A sample of the tf-idf matrix.

inverse logarithmically scaled frequency with which the term occurs in all the documents in the corpus. Formula 4.6 shows the basic formula, where $tf\,(t,d)$ is simply the number of times each word occurred in a document d. Formula 4.7 explains the idf part.

$$tfidf\,(t,d,D) = tf\,(t,d) \times idf\,(t,D)$$

t	terms
d	document
D	collection of documents

Figure 4.6: The formula for tf-idf.

Using this measure, I constructed a tf-idf matrix using the "scikit-learn" library, specifically the class called "TfidfVectorizer" [98]. The class offers several parameters, including parameters to define the minimum and maximum document frequency, meaning that features (in this case the tokens) that appear in a lower percentage of all documents than the given "min_df" parameter will be discarded. Accordingly, terms that have a document frequency of more than the "max_df" parameter will not be included in the resulting tf-df matrix. For this use case, I chose a minimum document frequency of 0.2 and a maximum document frequency, corresponding to 20% resp. 80% of all documents [98]. Table 4.1 shows a sample from the resulting tf-idf matrix.

$$idf\,(t,D) = \log \frac{|D|}{1 + |\{d \epsilon D : t \epsilon d\}|}$$

D	document space
m	mass
$\lvert\{d \epsilon D : t \epsilon d\}\rvert$	total number of times in which term t occurred in all documents

Figure 4.7: The formula for idf.

4.3.2 Principal Component Analysis

A principal component analysis (PCA) is a method used to reduce the dimensionality of a matrix. The main goal of performing a PCA is condensing the information in a large number of variables into a smaller and lower set of new composite dimensions, revealing relationships in the data defined by a low-dimension set of axes that summarize the data. PCA performs a linear transformation on the data set, moving the original data to a new space that is composed of so-called principal components. In tasks involving classification and segmentation analysis, PCA is often performed before a clustering method is used. As mentioned before, PCA can reveal confining composite features. Clustering methods such k-means, further described in Subsection 4.3.3, mathematically group samples into a number of clusters, based on how closely samples are distributed in relation to each cluster's center. Transforming the data set from its original dimensional space to a new one defined by the principal components can facilitate clustering, yielding more useful and insightful results. Using the PCA class from the "scikit-learn" library with the optional "n_components" parameter [99], I transformed the tf-idf matrix described in Subsection 4.3.1 to a new space defined by thirty components.

4.3.3 K-means Clustering

K-means clustering is a clustering method with the goal of partitioning n data points into k clusters. K-means will put the data points into a given number of clusters with each point belonging to the nearest mean, the so-called cluster center, or cluster centroid. It minimizes the squared Euclidean distance within clusters. I performed k-means clustering using the "KMeans" class from the "scikit-learn" library [100], clustering the output of the principal component analysis into three different clusters.

4.3.4 Shifterator

"Shifterator" is a Python package that "provides functionality for constructing word shift graphs, vertical bar charts that quantify which words contribute to a pairwise difference between two texts and how they contribute." [101] It was developed in the context of a publication called "Generalized Word Shift Graphs: A Method for Visualizing and Explaining Pairwise Comparisons Between Texts" by Gallagher et al. [102]. It offers various graph variants that are based on different textual measurements, including entropy-based graphs such as the ones I used in this thesis to analyze the linguistic differences between the identified clusters, as can be seen in Chapter 5.

4.3.5 VADER (Valence Aware Dictionary and sEntiment Reasoner)

"VADER" (Valence Aware Dictionary and sEntiment Reasoner) is a "lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media."

[103]. It contains several different text corpora and sentiment ratings from 10 trained human raters, having rated over 9000 token features on a sentiment scale of -4 (extremely negative) to 4 (extremely positive). VADER uses all tokens that were rated a non-zero rating and which had a standard deviation of less than 2.5 measured by the aggregate of the ten raters, which left the library with 7500 tokens. VADER is also able to rate the sentiment of emoticons in text, making it especially appropriate to analyze tweets and social media language.

The output of the library for a sentence contains four different scores: A score for positive sentiment, a score for neutral sentiment, a score for negative sentiment, and a compound core. The compound score is the overall sentiment score for a sentence calculated by the valence (the intensity of sentiment regardless of negativity or positiveness) for all the words in the sentence, with -1 meaning that the sentence is rated as extremely negative in sentiment, whereas a score of 1 means an extremely positive score was given. The positive, neutral, and negative scores are the ratios of the proportions of the text that fall into the respective category of sentiment and therefore add up to 1 respectively 100% [103]. An example for a sentence is "VADER is smart, handsome, and funny", which returns a score of `{'pos': 0.746, 'compound': 0.8316, 'neu': 0.254, 'neg': 0.0}`, meaning that VADER rates this sentence as having no part of it falling into the negative sentiment, 25.4% into a neutral sentiment and 74.6% into positive sentiment. A compound of 83.16% means that this sentence has a high valence respectively intensity in expressed sentiment. For the analysis in this thesis, I used VADER as the built-in version of the NLTK library [104]. I performed the sentiment analysis on the original, unprocessed tweets in order to take emojis and other sentiment-defining parts of language into account.

Chapter 5

Results

As mentioned in Chapter 1, my goal in this thesis is to detect communities using ridesharing platforms such as Uber or Lyft, activists talking about workers rights, politicians talking about regulation and policies targeting either the companies or their employees, or investors posting about the stock market price of these companies and understand how Twitter is used as a communication tool by either of those various groups. In Chapter 4, I explained the methods I used for data pre-processing and analysis, including dimensionality reduction using PCA and community detection via k-means clustering. In this chapter, I will first show general insights I gained about the nature of the collected data set, such as the distribution of tweets throughout April and May 2019 in Section 5.1, the usage of hashtags, and how they appear in combination in 5.2. In Section 5.4, I will explain the results of the PCA and k-means clustering approach, analyzing the identified communities in terms of common topics and how they differ compared to other user groups. In Section 5.5, I will focus on illustrating linguistic differences between the communities before I present the results of the sentiment analysis in Section 5.6, showing correlation patterns between expressed sentiment and amount of retweets, as well as discussions of Uber-related cultural, political, and legal conflicts in tweets with high valence scores. Please note that I performed the data analysis that is not specific to clusters on the entire data set of 51,853 tweets, while the analysis in Section 5.4 takes the only the final selection of 353 documents/users into account.

5.1 Tweets over Time

To gain some first insight into the data set, I analyzed the general user activity within the observed time frame, shown in Figure 5.1. It can be seen that the general activity shows a small spike around April 12, 2019, which coincides with the day where the filing for a public offering by Uber became public. Two other clear spikes are visible: One on May 8, 2019, the day when drivers came together globally to protest and strike for better working conditions. The other spike occurs on May 10, 2019, the day of Uber's IPO. Both events are marked in the figure with red vertical lines.

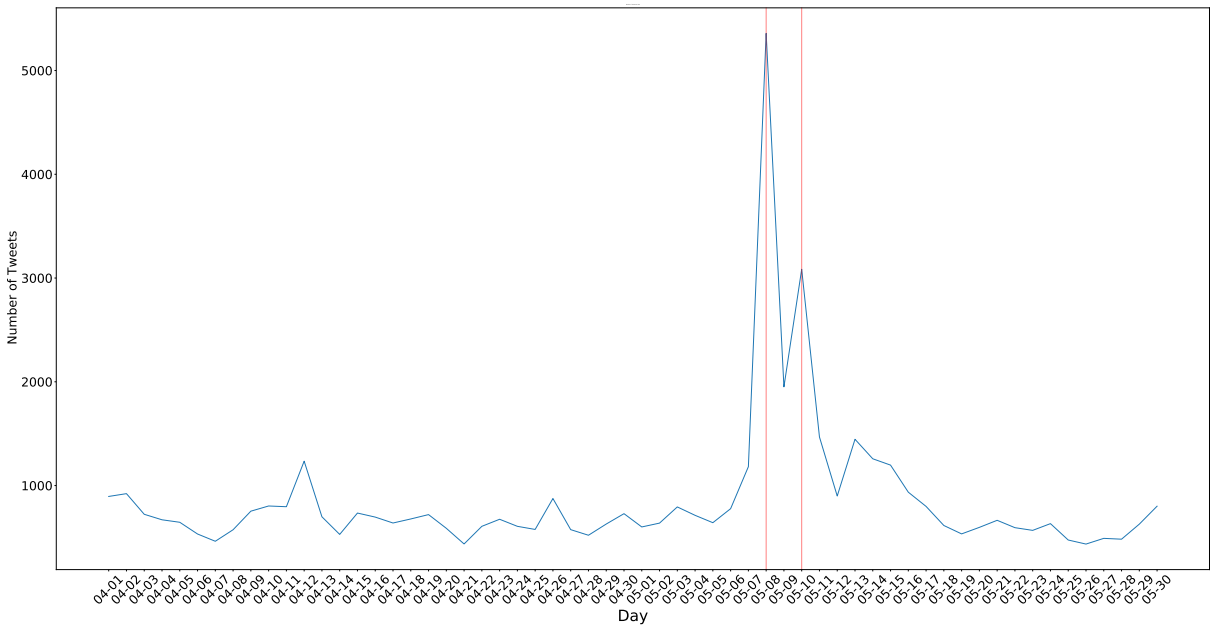


Figure 5.1: Number of tweets per day in the data set in April and May 2019, showing clear spikes on the day of the strikes, May 8, and Uber's IPO, May 10. Both events are marked with red vertical lines.

5.2 Hashtag Occurrences

Analyzing which hashtags appear how often gives a first broad overview of the topic of the data set. Figure 5.2 shows that 72.9% (37820 tweets) of all tweets in the data set contain the hashtag '#uber', followed by '#lyft' with 28.7% (14880 tweets) and '#uberlyftstrike' with 7.2% (3739 tweets). This shows that a great majority of the tweets contain hashtags indicating either the general topic of Uber or its competitor Lyft. A smaller but significant percentage of tweets contains '#uberlyftstrike', indicating the importance of the topic of the driver strikes ahead of the company's IPO. Note that tweets may contain multiple hashtags and thus these groups do not add up to 100%. The appearance of the hashtag '#taxi' among the ten most used hashtags in the data set can be explained with the term being generally related to ridesharing services, as Uber and Lyft represent alternative forms of personal transportation. The usage of '#uk' can be explained as Twitter users explicitly labeling their tweets about Uber to be from the United Kingdom, as the largest amount of users are located in the United States [105]. '#tfl' stands for 'Transport for London', a government organization responsible for transportation services in Greater London, including London Underground, London Overground, London's trams, buses, and taxis [106]. As can be seen in Section 5.3, '#tfl' occurs often in combination with '#uk' and '#taxi', indicating that the hashtag occurs often in discussions about the position of Uber within London's transport system and the political and cultural conflict surrounding it.

5.2.1 Hashtag Occurrences over Time

Analyzing hashtag can detect potential broad communities defined by hashtag use and get a first impression of their behavior and activity over the time frame reflected in the

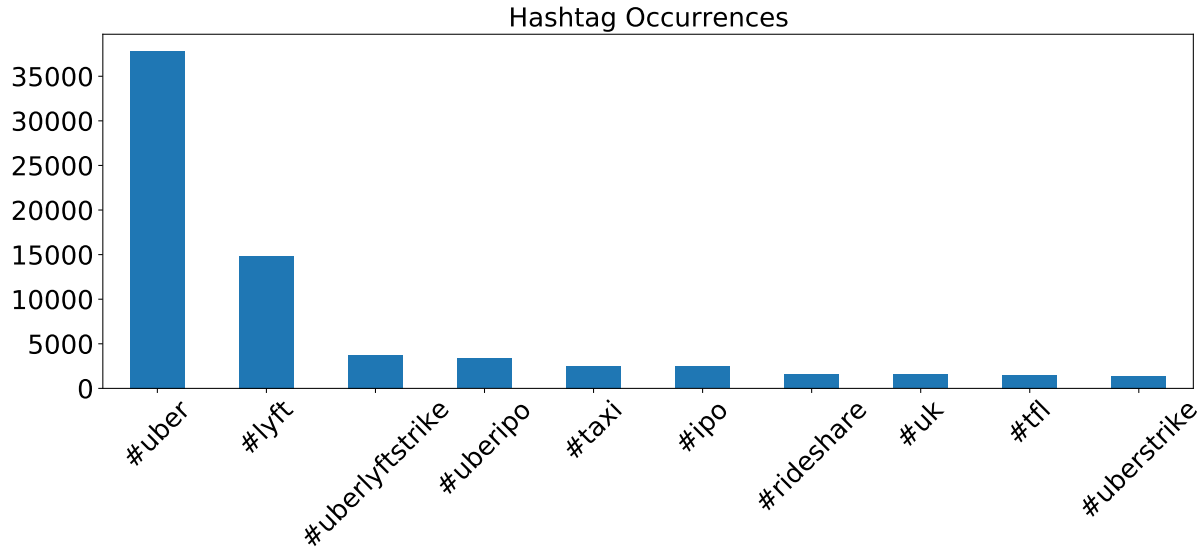


Figure 5.2: Occurrences of hashtags in the collected data set, giving a general overview of the topic distribution in the data set.

data set. Figure 5.3 shows the ten most common hashtags in the data set and the number of occurrences over time, with a vertical line marking the day of the IPO, May 10, 2019. The horizontal axis reflects the time frame of the two months of April and May 2019. It can be seen that there exists a somewhat consistent usage of ‘#uber‘ throughout the two months of April and May 2019 but a clear rise in usage is visible around May 10, 2019, the day of the company’s IPO, as well as a decline in usage after that day. The same spike can be seen with ‘#ipo‘, although there exists a second spike around April 11 and April 12, the day on which the filing for an IPO became public [27]. It should be noted that the search terms for the data set did not contain ‘#ipo‘, meaning that no tweets were included in the result that merely contained ‘#ipo‘. The same spike of usage around the day of the IPO can be seen with ‘#uberipo‘. The hashtags ‘#uberlyftstrike‘, ‘#rideshare‘, and ‘#uberstrike‘ share a spike in usage on May 8, the day where many Uber and Lyft drivers participated protests to demand better pay and working conditions [107].

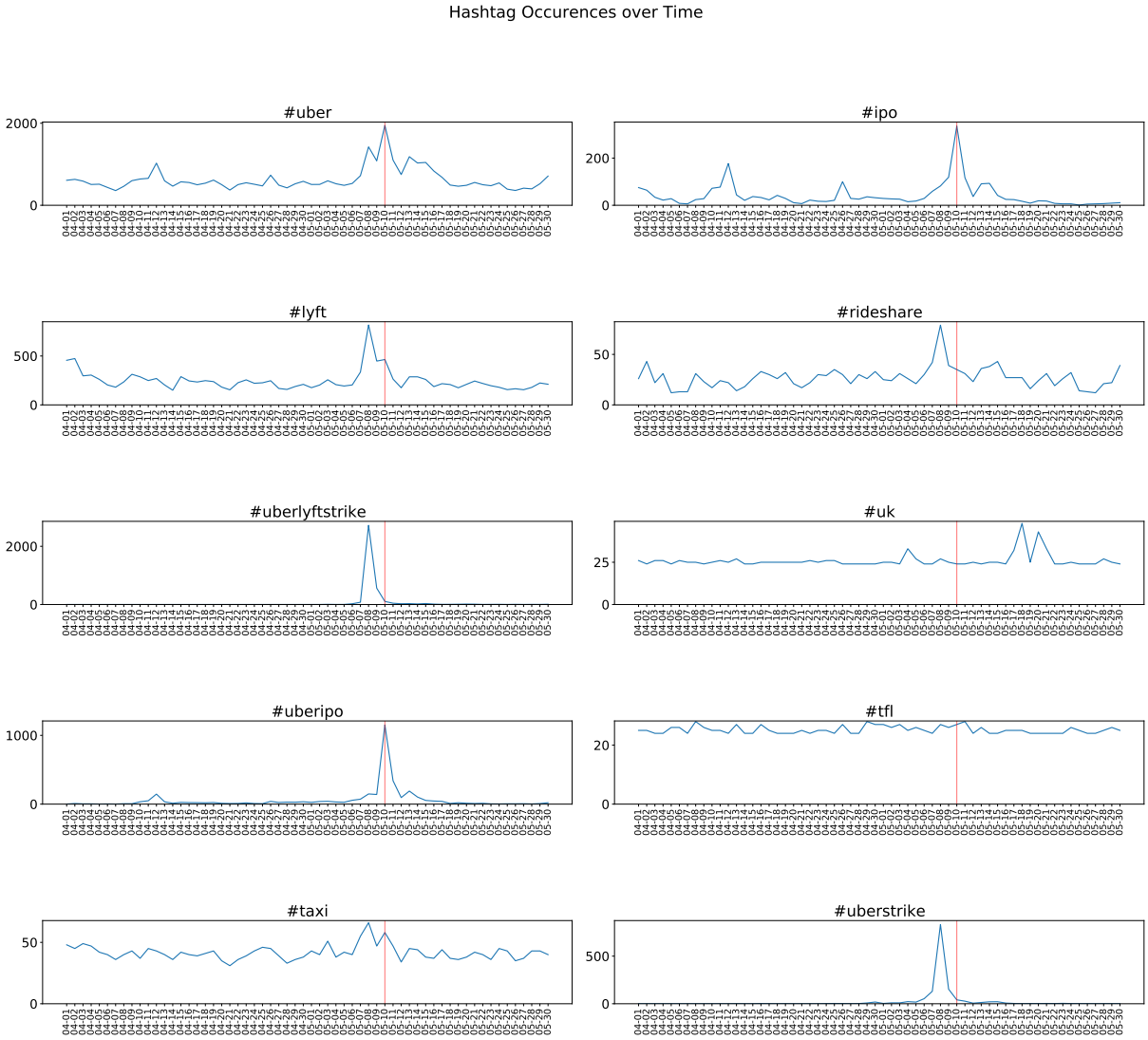


Figure 5.3: Occurrences of hashtags over time in the collected data set, showing a general topic distribution and topic-related spikes in activity. The day of the IPO is marked with a red vertical line.

5.3 Hashtag Co-Occurrences

I analyzed the collected data regarding the co-occurrence of hashtags to detect related topics in the data set. Figure 5.4 shows hashtags as nodes, with the size of the node proportional to the respective hashtags degree within the co-occurrence network. “#uber“ is the hashtags resp. node with the highest degree, indicating that is the hashtag most often occurring in pairs, which is expected as it is the most used hashtag in the data set in general. It is often used in combination with “#lyft“ indicating discussions about the relation of Uber towards its competitor Lyft. A network of related hashtags can be seen in the group of “#taxi“, “#uk‘, and “#tfl“, indicating discussions about the relation of Uber towards the Taxi transport system in London. “#tradewar“ relates to the ‘trade war‘ between the United States and China [108], indicating a high relation between Uber and financial topics in discussions about the United States stock markets during the time of Uber’s IPO. Uber Eats is a food delivery service offered by Uber that was launched in 2014.

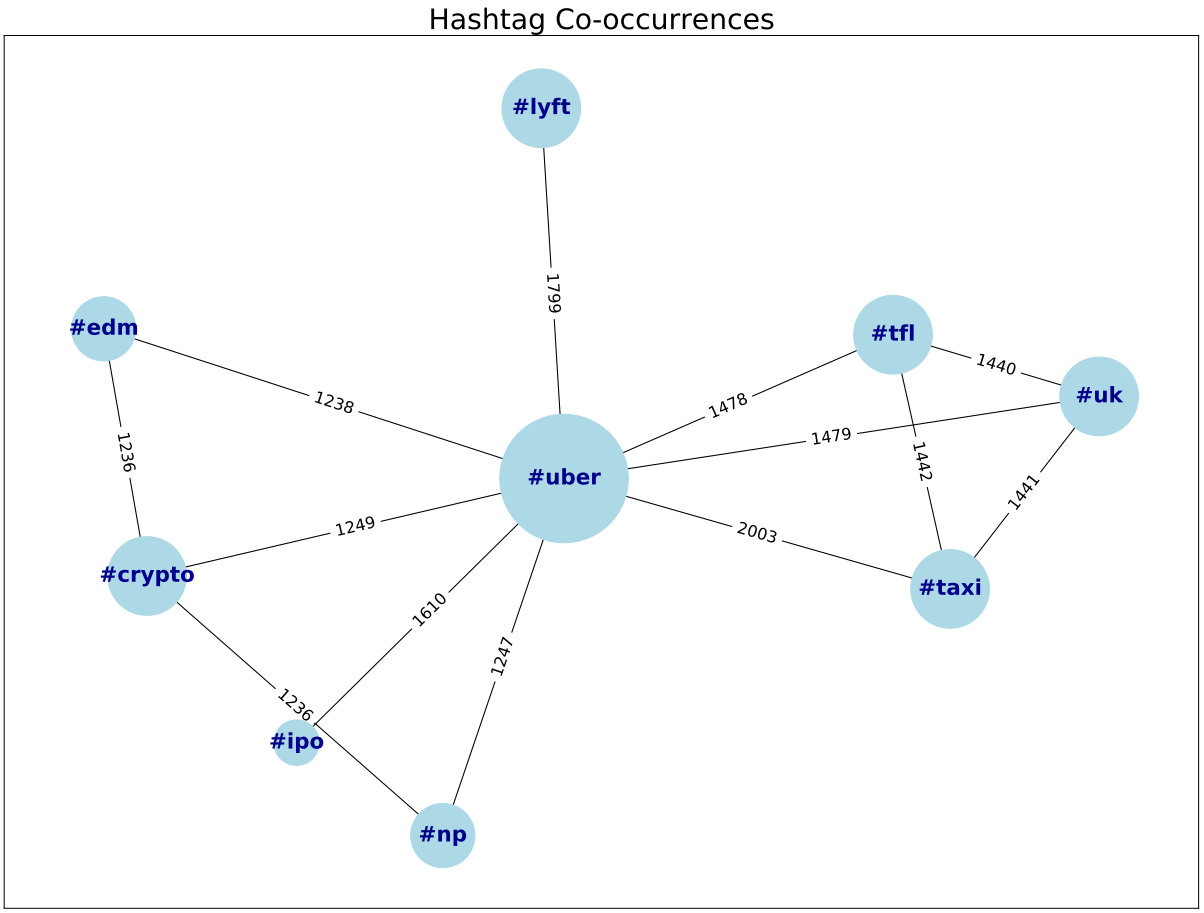


Figure 5.4: Co-Occurrences of hashtags in the collected data set, showing clusters or related hashtags.

5.4 Community Detection

I pre-processed the collected data set of tweets posted in April and May 2019 (see Section 4.1 of Chapter 4) using the steps outlined in Section 4.2. After collecting all tweets for every unique user in the data set and concatenating them, forming one document per user, I further prepared the data was by constructing a tf-idf matrix (see Subsection 4.3.1 of Chapter 4). As a next step, I transformed the matrix into a lower-dimensional space using PCA (see Subsection 5.4.1), after which I performed k-means clustering (see Subsection 4.3.3 of Chapter 4) intending to be able to identify user communities that share a common topic. Within this approach, a topic can be seen as a linear combination of words or decomposition of users' texts. A cluster of users can be defined as users with similar interests or the extent to which users are integrated into a topic. Following, I will discuss the results of this approach.

5.4.1 PCA

As described in Chapter 4, I performed the principal component analysis with thirty components. The scikit-learn library offers a way to see how much variance in the data each principal component explains:

The amount of variance explained by each of the selected components.
Equal to n_components largest eigenvalues of the covariance matrix of X.
New in version 0.18.

Table 5.1 shows the explained variance per principal component for the first ten principal components after performing the PCA on the data set. It can be seen that the second roughly explains half the variance in the data as the first component does. From the second component onwards, the added explained variance is low. Summed up, the thirty principal components have a total explained variance of 74%. Figure 5.5 shows the explained variance in relation to the number of principal components used.

PCA	Explained Variance
1	0.125383
2	0.077134
3	0.054076
4	0.049704
5	0.038319
6	0.033597
7	0.028052
8	0.024165
9	0.023803
10	0.021019

Table 5.1: The explained variance for each of the first ten principal components.

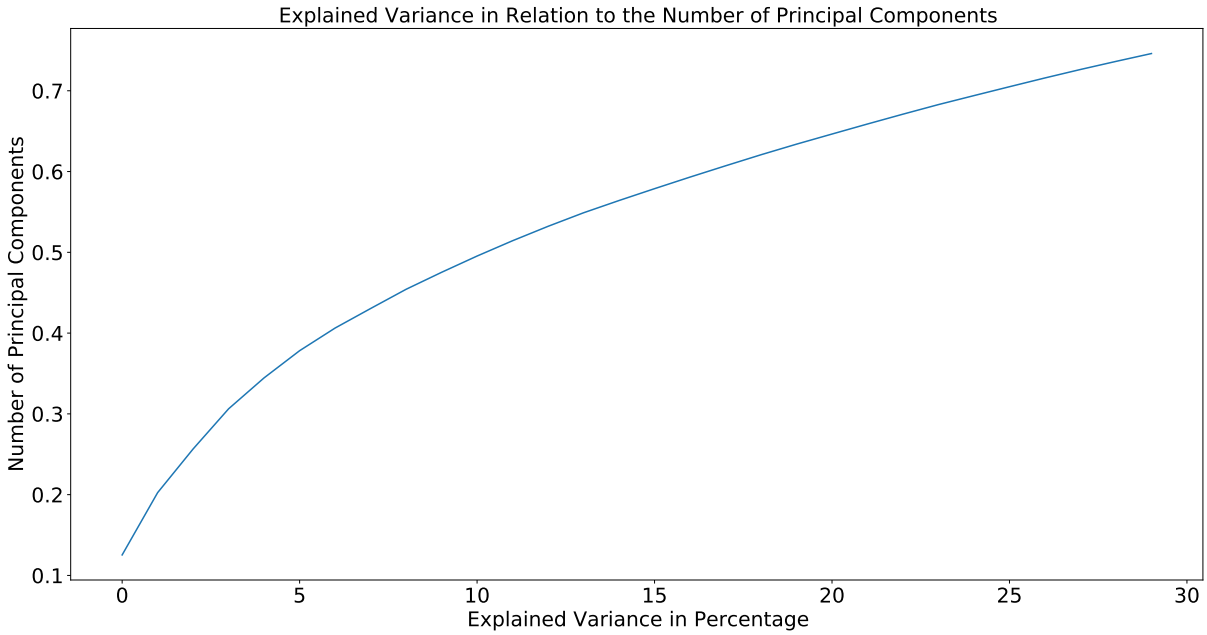


Figure 5.5: Explained variance in relation to the number of principal components, showing that the second principal component explains roughly half the variance compared to the first one.

Figure 5.6 shows a two-dimensional scatter plot displaying the data in relation to the first two principal components. It can be seen that groups exist after performing the

transformation. A clearly distinctive group exists with lower values for the first principal component (visible in the lower-left corner in Figure 5.6). Another distinction can be made along the axis of the second principal component. Figure 5.7 shows the first three principal components. It shows that there is also a distinction that can be made along the axis of the third principal component.

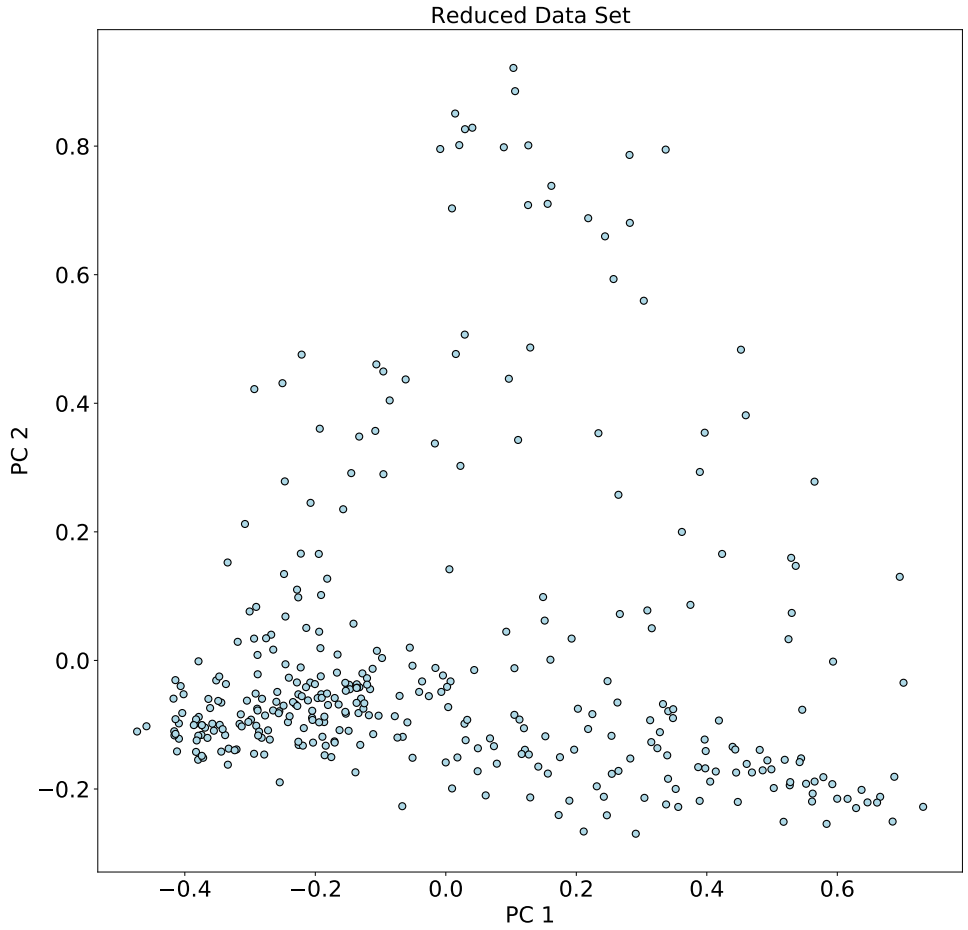


Figure 5.6: The first two principal components of the reduced data set. It can be seen that groups exist in the data set after performing the PCA.

Table 5.2 shows the ten features that explain the most variance (the ten most ‘important‘ features) for the first three principal components. It can be seen that for the first principal component words that are expected to occur in the context of finance have a high positive value, such as ‘ipo‘, ‘stock‘, ‘price‘, ‘billion‘, ‘share‘, ‘market‘ and ‘investor‘, while ‘driver‘ and ‘uberlyftstrike‘ have negative values indicating that this principal component is best explained by a lack of usage of those words. The second principal component shows a very high value for the word ‘uberipo‘ while showing negative values for eight of the other ten words. The third principal component shows a high value for ‘uberlyftstrike‘ and ‘driver‘, as well as ‘strike‘ and ‘ipo‘. Based on these words and their explained variance, one can already come to the assumption that there is a community of users associated with each of those principal components. The first principal component seems to construct a compound of characteristics associated with users talking about Uber and its IPO in the context of investing and finances. The second principal component seems to compound words associated with users talking about the IPO in general, while the third principal component

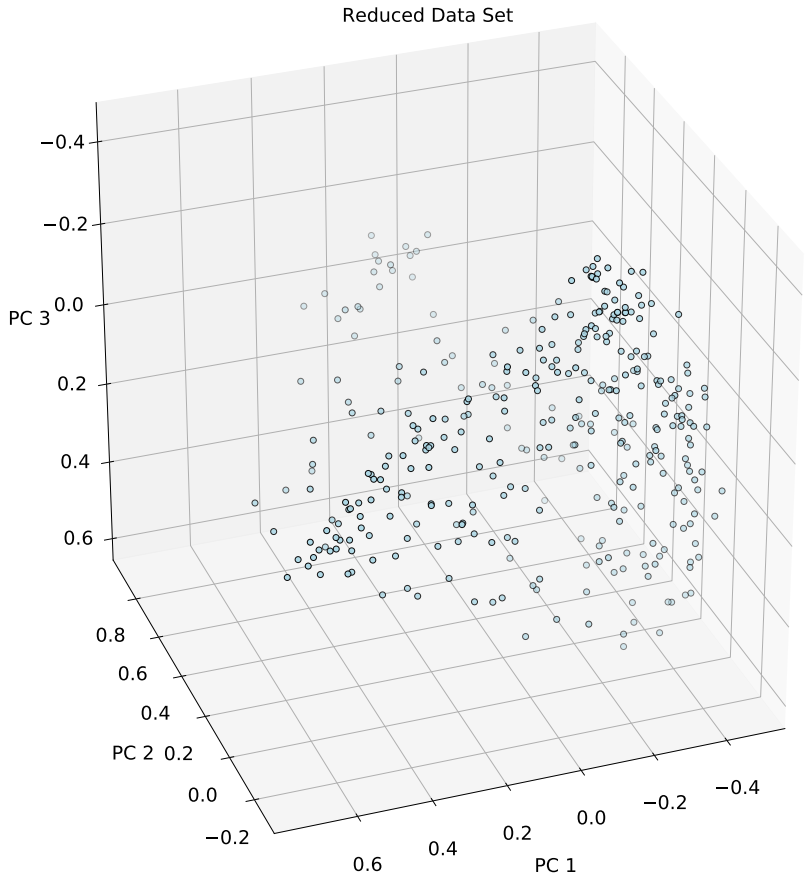


Figure 5.7: The first three principal components of the reduced data set. It can be seen that there exists a grouping in three dimensions as well.

features words that seem characteristic of users talking about Uber in a political context, mentioning drivers and striking.

Feature	Importance
ipo	0.680243
driver	-0.389888
stock	0.210885
uberlyftstrike	-0.186251
billion	0.182166
price	0.143981
share	0.143515
market	0.134462
public	0.131463
investor	0.124639

(a) The features that explain the most variance in the first principal component.

Feature	Importance
uberipo	0.934133
ipo	-0.236044
driver	-0.101842
car	-0.096105
billion	-0.072423
strike	-0.070060
uberlyftstrike	0.066637
rideshare	-0.064721
ride	-0.062816
app	-0.060335

(b) The features that explain the most variance in the second principal component

Feature	Importance
uberlyftstrike	0.552496
driver	0.519282
strike	0.297654
ipo	0.264821
taxi	-0.213645
know	-0.152077
pay	0.148680
car	-0.134816
ride	-0.129205
get	-0.126066

(c) The features that explain the most variance in the third principal component.

Table 5.2: The features that explain the most variance in the first three principal components, showing some confining features related to finances in the first component, the Uber IPO in the second component and the strikes and strikes in the third.

5.4.2 K-means

As explained in Subsection 4.3.3 of Chapter 4, I chose the k-means clustering approach to identify clusters of users respectively their documents consisting of all a user's tweets for the two observed months, defining those clusters as topics. I ran the k-means algorithm with the goal of finding three clusters, as the first three principal components of the PCA suggested three distinct groups of users, as is also supported by clusters already apparent in the plotted data as can be seen in Figure 5.6 and Figure 5.7.

Figure 5.8 shows the result of the k-means algorithm in two dimensions. The cluster centroids are displayed by the larger red circles, while the data points are colored according to their cluster membership. It can be seen that the cluster found in the bottom left corner of the plot (described as "cluster 1" as from now) has the most data points belonging to it in general and the most points showing a very short distance to the cluster centroid, while the other two identified clusters ("cluster 3" depicted by yellow points and "cluster 2" depicted by green points) do not seem to have data points as distinctly close to their respective centroids. The clusters differ in size as well: cluster 1 has 202 data points resp.

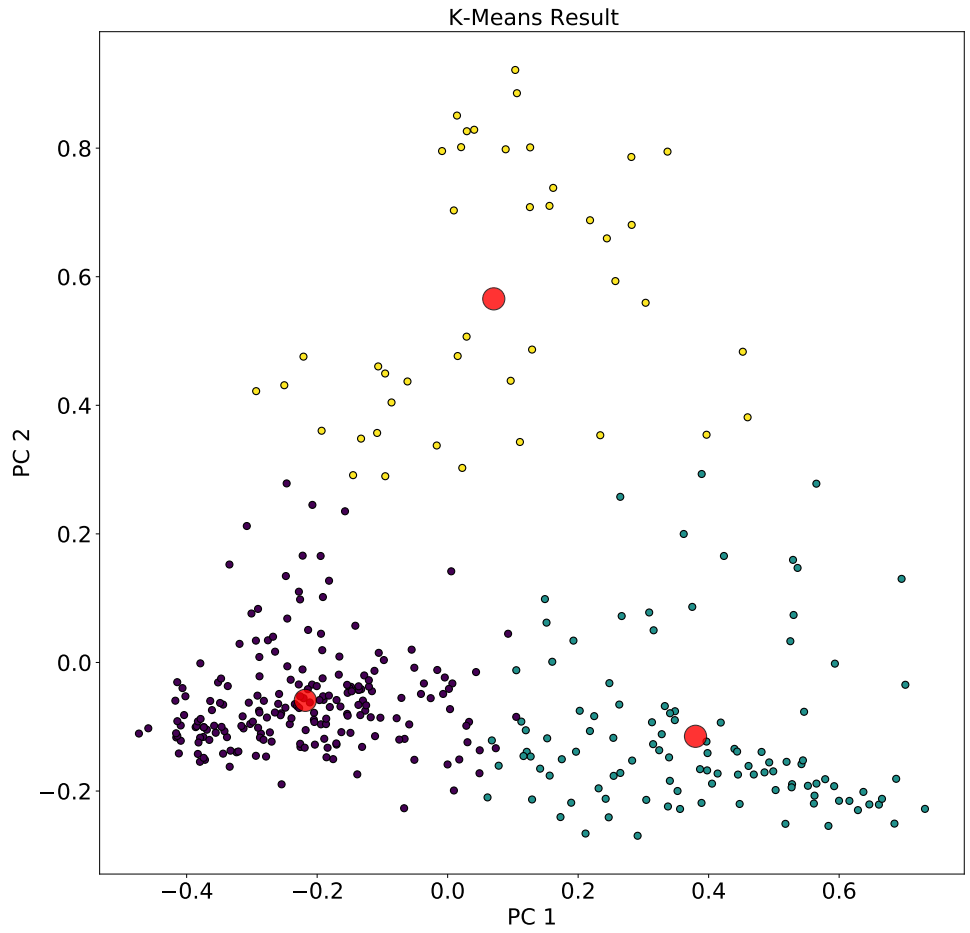


Figure 5.8: The clusters found by the k-means algorithm plotted in two dimensions.

users belonging to it. Cluster 2 is only half the size with 108 users, and 43 users belong to cluster 3. Some data points also seem to not specifically belong to a cluster appear to lie between clusters. Cluster 3 has the highest average distance of points to the cluster centroid if considering the first two dimensions. Figure 5.9 shows the clusters identified by the k-means algorithm in three dimensions, again showing the difference in the size of the clusters and the closeness of the respective points belonging to them.

Figure 5.10 shows the ten most used hashtags for the tweets of each of the identified cluster. It can be seen that in cluster 1 the hashtags ‘#uberlyftstrike‘, ‘#strikeuberlyft‘, ‘#uberstrike‘ as well as ‘#lyfstrike‘ are among the most used. Together with the usage ‘#drivers‘, this suggests that this cluster is composed of users talking about the drivers’ role in the strikes that took place before the IPO. The second cluster suggests a general topic of finance with hashtags like ‘#stocks‘ and ‘#investing‘, as well as technology and news. Cluster 3 seems to show an overlap with cluster 1 as can be seen in the use of both ‘#uberlyftstrike‘ and ‘#strikeuberlyft‘. The usage of ‘#ubercrash‘ and ‘#safetyneverstops‘ suggests topics surrounding safety concerns. ‘#deleteuber‘ and ‘#ubered‘, which Urban Dictionary defines as “To be massively conned into thinking you booked a taxi, but actually get a clueless gimp charging you silly money.“ [109], suggests a negative sentiment towards Uber. The usage of ‘#ipo‘ in cluster 3 also indicates an overlap with the topic of finances most prominently shown in cluster 2.

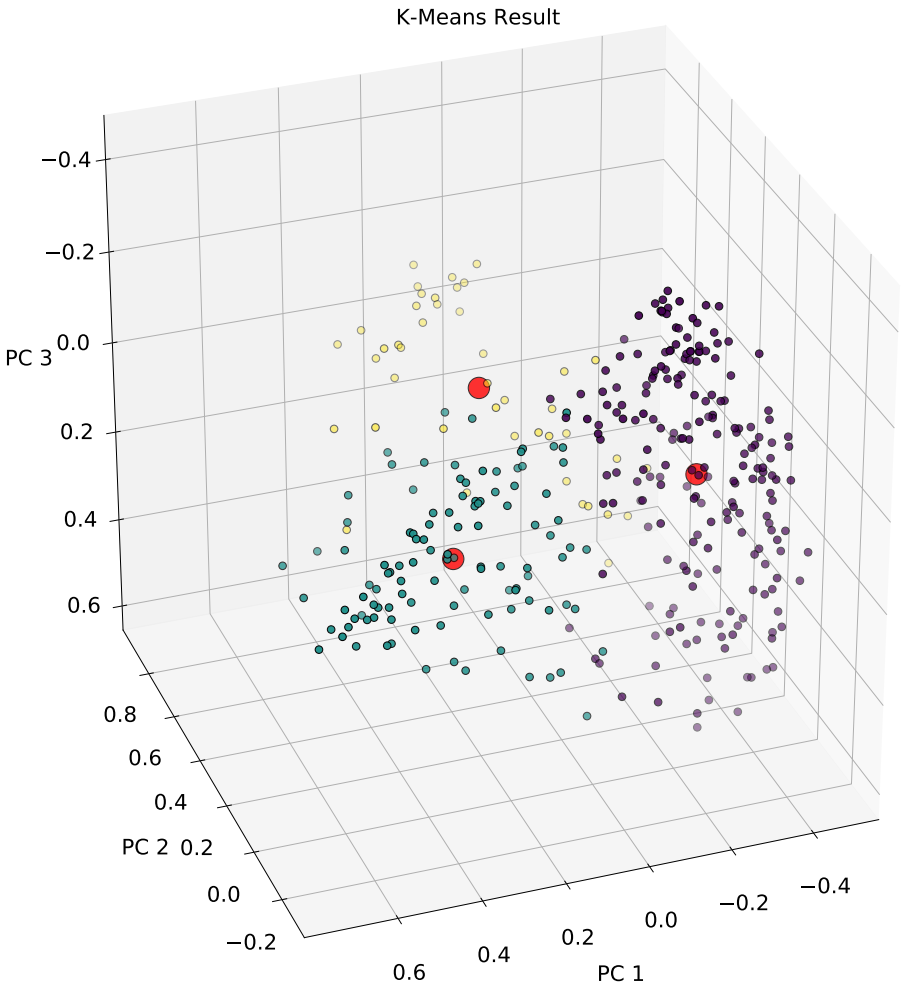


Figure 5.9: The clusters found by the k-means algorithm plotted in three dimensions.

Figure 5.11 shows the number of tweets for every cluster over the two months that were observed in the data set. Cluster 1 shows a relatively low number of tweets over the two months, with a significant spike on the day of the driver strikes on May 8, 2019. This additionally supports the assumption that this cluster is mainly consisting of users resp. documents covering Uber from a political standpoint, talking about worker's rights. Cluster 2 can be seen to also show clear spikes in the distribution of tweets. A first spike in activity is seen on April 12, a day after Uber filed its S-1 form in order to go public [27, 110]. Another spike is visible on April 26, the day on which Uber filed an amendment to its S-1 form [111].

Figure 5.12 shows the four most-liked tweets from users put in the first cluster. It can be seen that the tweets, except for the tweet of user 'FitzTheReporter' that reports the strike happening, are talking about the driver strikes in a supporting and sympathetic manner, prompting followers to show solidarity with striking workers and criticizing both Uber and Lyft for underpaying workers while providing corporate executives with millions by profiting off the IPO.

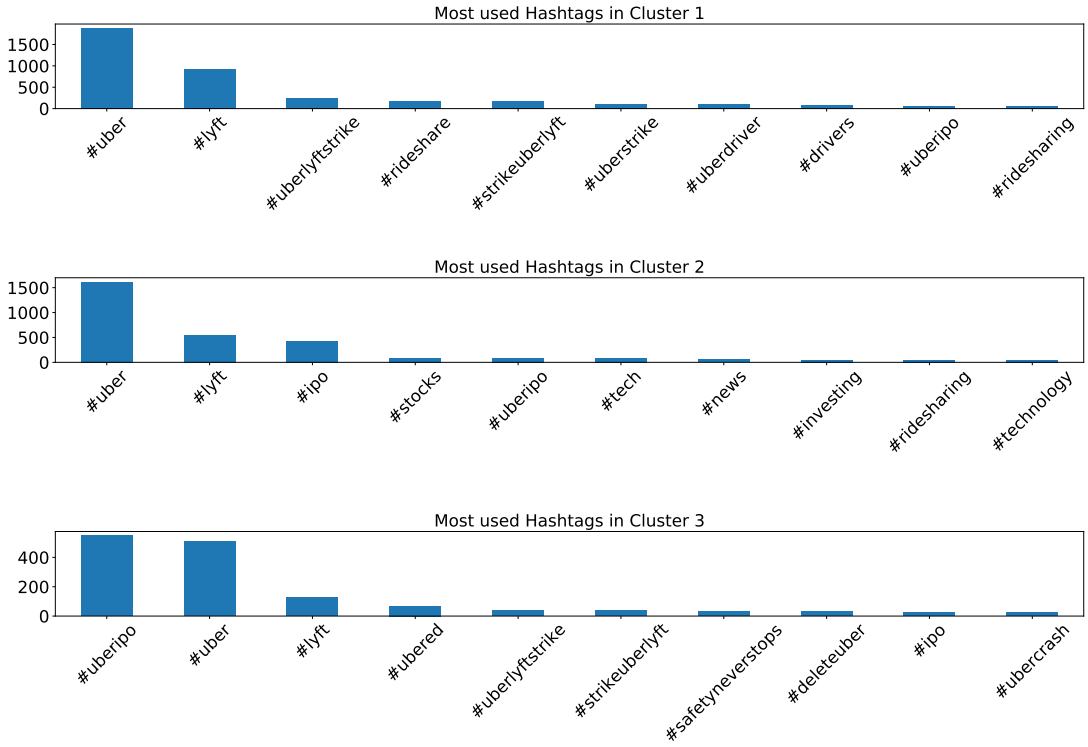


Figure 5.10: The ten most used hashtags of the tweets in each respective cluster, showing distinct topics for each of the identified clusters.

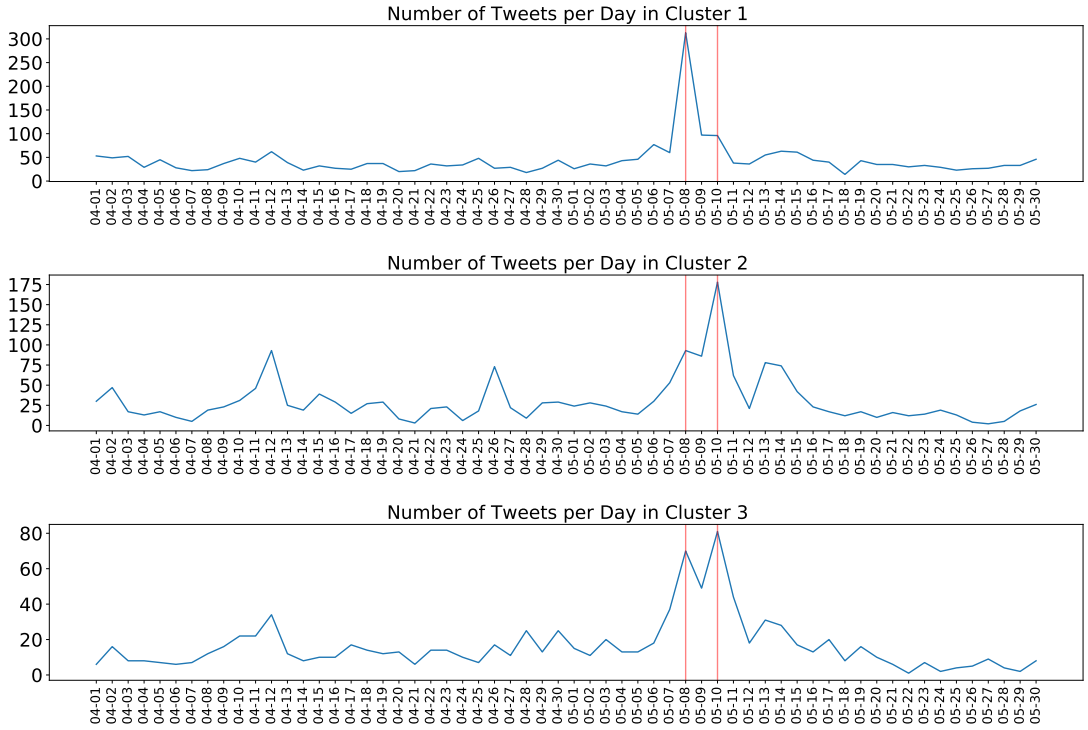


Figure 5.11: The distribution of tweets over the observed two months for each cluster, showing a spike on the day of the strikes for the first cluster, a spike on the day of the IPO for the second cluster, and both spikes in the third cluster, indicating the mentioned overlap. Both events are marked with red vertical lines.

Figure 5.13 shows the four tweets from the second cluster with the most likes. It can

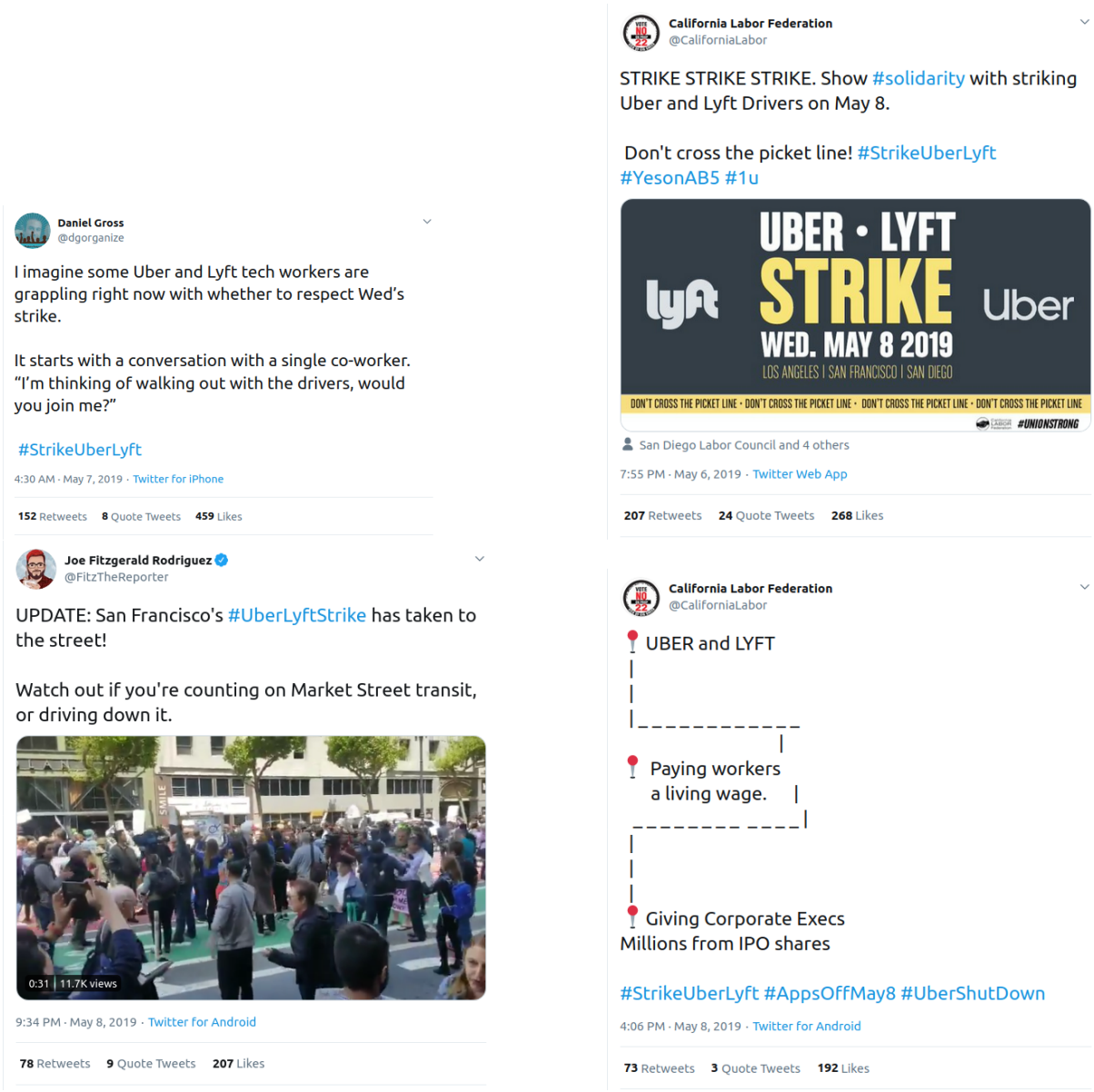


Figure 5.12: The four most liked messages of the first cluster.

be seen that the users talk about Uber and Lyft in the context of stock prices, finances, and technology, mentioning the user base of each of the company's respective apps and the risk of the stock plunging after the initial public offering, as mentioned by "eToro", a social trading platform [112]. Uber's stock price did indeed after the IPO, as mentioned in the tweet by "Schuldensuehner", the account of Holger Zschaeptitz, editor in chief for economics at "Die Welt", a German national daily newspaper [113].

The four most-liked tweets in cluster 3 can be seen in Figure 5.14. It can be seen that the tweets support the appearance that the third identified cluster is overlapping with the first cluster in regards to the topic, with the exception of the tweet by the user "NagleKieron", which emphasized road safety and shows a video of a car crash involving an Uber drivers car. This supports the classification of this cluster's document talking about safety concerns.

eToro
@eToro

The U.S.' 2 biggest transportation network services are often compared even as both companies hope to promote shared & sustainable transpo services.

#Lyft began trading on 29.03 but soon after plunged by 12%. How will #Uber perform with the planned IPO expected in April?

2:04 PM · Apr 4, 2019 · Twitter Web Client

67 Retweets **2** Quote Tweets **384** Likes

eToro
@eToro

Uber currently has over 90 million users, whereas rival Lyft only has around 20 million. Uber believes that the future of mobility is increasingly shared, sustainable and most importantly, automated. As such, the company is placing a strong focus on self-driving cars.

#Uber #IPO

12:47 PM · May 2, 2019 · Twitter Web Client

12 Retweets **5** Quote Tweets **164** Likes

Holger Zschaepitz
@Schuldensuehner

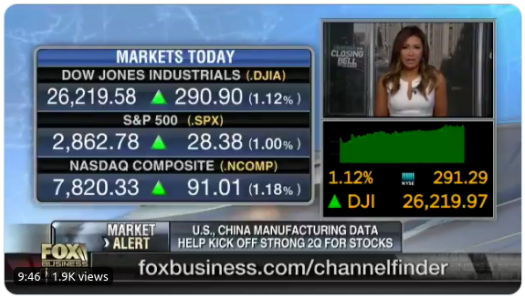
#Uber plunges again, trades now about 18% below IPO price.

7:46 PM · May 13, 2019 · Twitter Web Client

88 Retweets **20** Quote Tweets **156** Likes

Susan Li
@SusanLiTV

Americans feeling richer with #stockmarket hitting 2019 highs this week! Covered on Closing Bell #China #Trade #Brexit #border safety #2020election hopefuls #technology #ElonMusk #Lyft $lyft $fb $amzn $dji $spy $ndq



4:47 AM · Apr 3, 2019 · Twitter Media Studio

16 Retweets **102** Likes

Figure 5.13: The four most liked messages of the second cluster.

Rideshare Drivers United
@_drivers_united

WE'RE OFFICIALLY CALLING FOR A 24 HR STRIKE AT LAX ON MAY 8TH. APPS OFF FROM MIDNIGHT TO MIDNIGHT.

This will be in concert with protests in other cities in solidarity with @_drivers_united LA strike.

#StrikeUberLyft #May8thLAXStrike #NationalDayOfAction

6:14 PM · Apr 20, 2019 · TweetDeck

192 Retweets **28** Quote Tweets **323** Likes

Nag the Cab
@NagleKieron

Wouldn't be a normal day/night in #london without the abnormal #uber accident
#SafetyNeverStarted
#Safetyneverstops
#TFLers
#UberCRASH
@TfL
@LondonAssembly
@metpoliceuk

5:21 PM · May 27, 2019 · Twitter for Android

183 Retweets **22** Quote Tweets **148** Likes

Rideshare Drivers United
@_drivers_united

Its nearing 12am PST and we're about ready to hang it up. Today went better than we could've ever dreamt a month ago when we started to plan. THANK YOU to all who made it possible, to the organizers in our sister cities and drivers who protested today. RDU LA OUT.
#UberLyftStrike

8:44 AM · May 9, 2019 · Twitter for Android

29 Retweets **87** Likes

La Acera
@carbonederigau

We're ending the strike at LAX with street heat on #UberLyftStrike breakers. One driver pulled over to talk, said he was sorry for scabbing but needs to buy groceries. Hes on food stamps and section 8. Been driving 12 hrs, $142 earned and spent 1/2 that in gas. Heartbreaking.

7:08 AM · May 9, 2019 · Twitter for Android

29 Retweets **8** Quote Tweets **74** Likes

Figure 5.14: The four most liked messages of the third cluster.

5.5 Shifterator

As I explained in Subsection 4.3.4 of Chapter 4, I analyzed the clusters regarding their linguistic difference using the 'Shifterator' library providing 'entropy word shift graphs', which quantify the words contributing to the difference between two texts [101]. The graphs show the difference between Shannon entropies of the two clusters' documents. The Shannon entropy calculation over a single word can be interpreted as the "surprisal" of a word. When looking at an entire text it can be seen as the average "surprisal" of a text [114]. The difference between the texts' entropies illustrates the difference that each word contributes. A positive contribution score for a word means that the word has a higher score in the second text, a negative score means its score is higher in the first text [114].

Figure 5.15 shows an entropy shift graph illustrating the difference between the texts formed by the documents included in cluster 1 and cluster 2. By the title of the graph showing the average score and the sum $\sum$, we can see that the text of cluster 1 is slightly more unpredictable. By looking at the type ranks, we can see that words like "driver" and "uberlyftstrike" contribute heavily to the surprisal of the text of the first cluster, indicating that the cluster's users are generally talking about politics and drivers' rights. The words contributing most to the surprisal of the second cluster's text are "ipo", "stock", "billion", "price" and "share" as well as "trading" and "investment". This indicates a heavy leaning of the cluster's text towards financial topics. Words such as "technology", "tech" and "news" also indicate a topic surrounding technological news. The results shown in Section 5.4 support that interpretation of the clusters' general topic: The four most liked messages of the first cluster show users tweeting about political support for the drivers' strikes, while for the second cluster they show a mixture of technological and financial news surrounding Uber as a company as well as the price of the public stock. The cumulative contribution plot in the lower-left corner of the graph shows how much the summed contribution changes in relation to the number of words included in the calculation. The horizontal line in the cumulative contribution plot shows the cutoff of the words shown in the graph to those that are omitted. We can see that the 50 words in the graph explain about 20% of the text difference.

Cluster 3 has previously in this section identified to be the smallest cluster as well as having the highest average distance of its data points to its cluster centroid in the first two dimensions. In Subsection 5.4.2 in Section 5.4 it was shown that cluster 3 seems to show an overlap in hashtag usage with cluster 1, as shown by the occurrence of hashtags relevant to the strikes in both clusters. The usage of "#ipo" in cluster 3 as well as in cluster 2 also suggests an overlap, although a smaller one. Figure 5.16 shows the entropy shift graph for cluster 2 and 3. Similar to Figure 5.15, cluster 1 is generally slightly more unpredictable. Similar to the comparison between cluster 1 and cluster 2, the 50 plotted words contribute to about 20% of the difference. The graph also shows that for cluster 1, words like "uberlyftstrike", "people", "worker" and "driver" contribute to the surprisal of the text, further supporting the interpretation of this cluster as the "political" one. For cluster 2, we can see that words like "london", "safetyneverstops", "ubercrash" contribute to the difference. This supports the interpretation of this cluster to be generally negative towards Uber, but that the clusters text is focused on safety concerns and problems with Uber separate from drivers rights, as the contribution of the word 'uberrape' suggests,

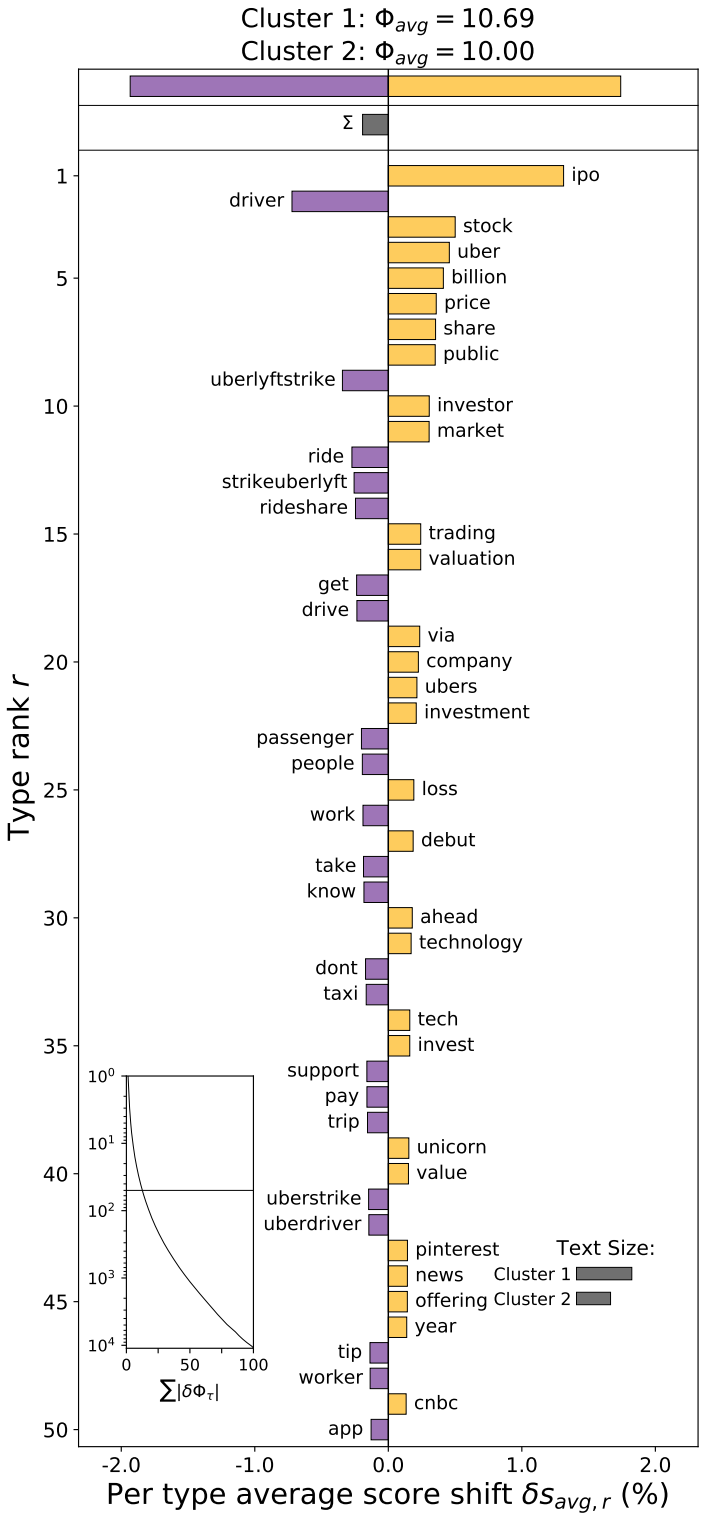


Figure 5.15: Entropy word shift graph between the first and the second cluster, showing the difference in used language by illustrating the most "surprising" words for each cluster.

which is a hashtag used to refer to the sexual harassment accusations against Uber drivers in multiple cities as well as accusations against the political executives of these cities to not take the necessary measures.

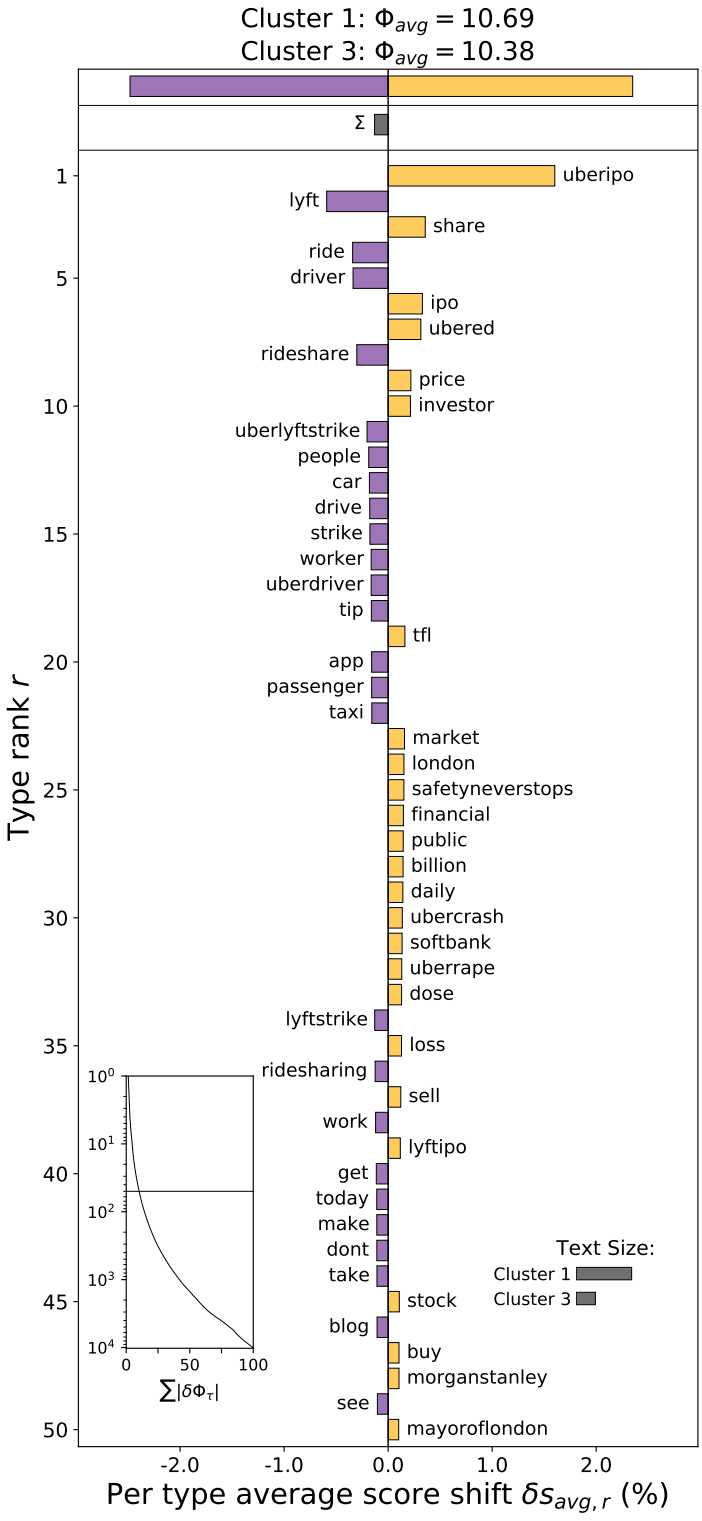


Figure 5.16: Entropy word shift graph between the first and the third cluster.

5.6 Sentiment Analysis

In addition to the language-based analysis in the previous section, I also analyzed the clusters regarding their expressed sentiment. For each of the three clusters, the ten tweets with the highest (positive valence) and lowest (negative valence) compound score of VADER's output were listed. I explain the scoring in Subsection 4.3.5 of Chapter 4.

In cluster 1, the cluster labeled as having a political topic, four out of then of the tweets having the highest compound score was authored by the user "nicomoe", including the three tweets with the highest compound score. "nicomoe" is the account of Nicole Moore, a driver and strike organizer who describes herself on her profile as "Proud Civil Service Worker" and "Justice Seeker" [115]. Figure 5.17 shows the tweet with the highest compound score in cluster 1. It can be seen that Nicole Moore shows her support with an opinion piece written by an organizer affiliated with Gig Workers Rise. Nicole Moore is also among the ten most retweeted users in her cluster with a total amount of 273 retweets during the two months of April and May 2019. Other tweets authored by Moore among the tweets with the highest compound scores show her expressing pride towards the strike, emphasizing the alleged success of the global protests, as shown by Figure 5.18. Figure 5.19 shows a tweet by Moore mentioning the account of Rideshare Drivers, the main organizing group behind the protests in Los Angeles United, in reply to a tweet by Gig Workers Rising, another organization described in Chapter 2. This indicates a high level of connection among the organizations. The tweet frames bonuses given to drivers by Uber as a way to keep drivers from protesting and includes a call of action, telling them to strike regardless.

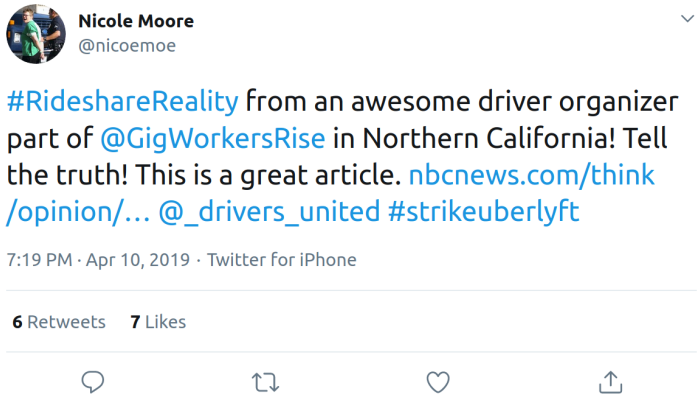


Figure 5.17: Tweet by user "nicomoe" with the highest compound score of 0.9018. Source: `https://twitter.com/nicoemoe/status/1116027987906813952`

Many of the tweets with the most negative compound scores in cluster 1 are posted by the user "AdryanCroydon" with the screen name "plasticscouser". In Figure 5.20 we can see a tweet with a compound score of -0.9251. Figure 5.21 shows another tweet by the user, in which the user posts pictures of an article by the British conservative newspaper "The Daily Telegraph" [116], describing the man depicted in the article as a "convicted serial rapist" and saying that "Transport for London" would license him as an Uber driver as he claims that they do not check criminal history. Another tweet by the user shown in Figure 5.22 states that only "6% of #Uber drivers identify as British" and says that Uber is unsafe, implying that foreign drivers pose a greater danger than non-foreigners. In the

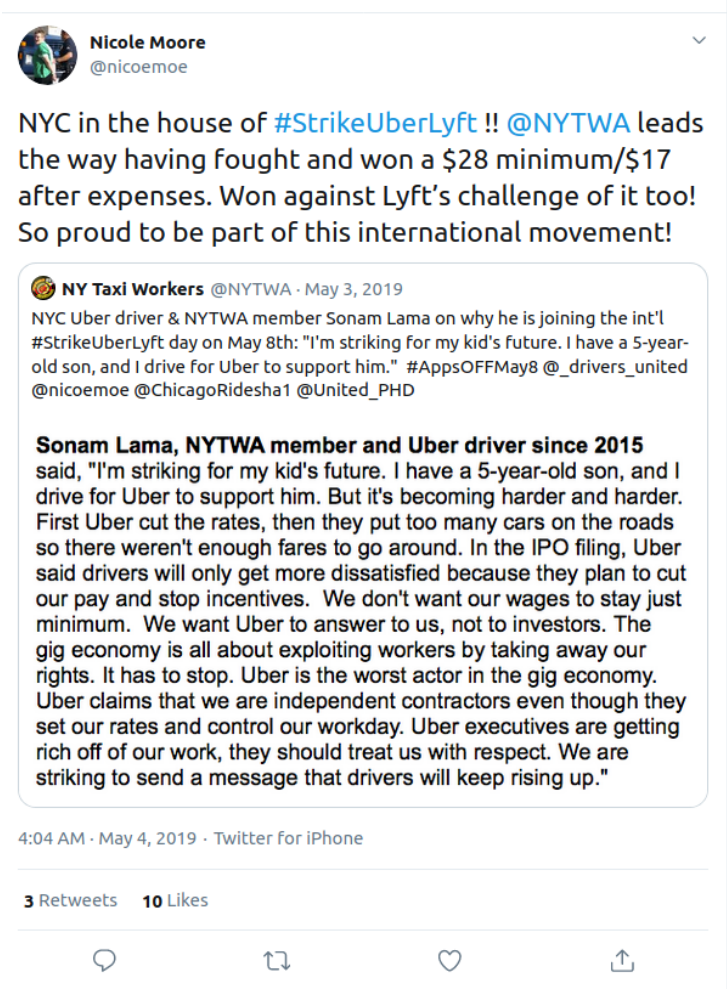


Figure 5.18: Tweet by user "nicomoe" with the highest compound score of 0.8992. Source: `https://twitter.com/nicoemoe/status/1116027987906813952`

entire data set, 17 tweets by this user can be found, with 5 of them being among the 10 tweets with the lowest compound sentiment score. Ten of the user's 17 tweets contain the word 'foreign'. Other tweets by the account contain reactions against racism allegations against his person and negatively highlight the economic impact of refugees in the United Kingdom claim. In a reply to the latter tweet, the user states that asylum seekers make no contribution to society [117, 118].

Figure 5.23 shows a tweet by the user "DDDaughters' linking to an article by CNN reporting 103 cases of Uber drivers accused of sexual assault or abuse. The is account described as "'#DDD Dads Defending Daughters: Highlighting the prolific Rape, Sexual and Physical Assaults in London Minicabs and Ubers which #TfL cover up. #UberRape is real." [119] The account is appearing three times in the ten tweets in cluster 1 having the lowest compound scores.

The username "plasticscouser" respectively "scouser" (a word describing Liverpool natives or inhabitants) indicates that this account as well "DDDaughters" are run by people located in the UK. This seems to confirm what has been described as "deep cultural roots" of the discussion around Uber's business in the UK and London specifically and as special compared to the sentiments that the company has seen directed against itself in other

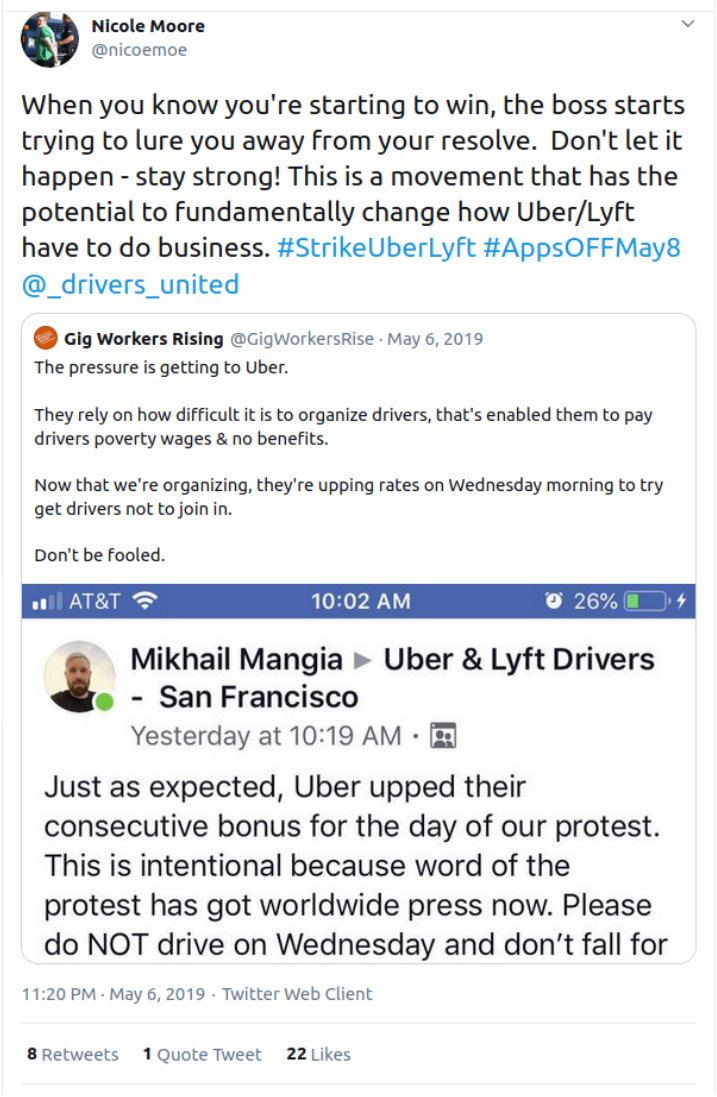


Figure 5.19: Tweet by user "nicomoe' with a compound score of 0.8748. Source: `https://twitter.com/nicoemoe/status/1125510693636136960`

cities around the globe. In the UK, the conflict between Uber drivers and traditional Taxi drivers can be seen to mirror the "cultural wars" taking place before the vote of the United Kingdom to leave the European Union and can be described as "immigrant versus native, old versus new, global versus national.", as several reports of racist abuse against Uber drivers by other cab drivers indicate. One explanation of this conflict could be demographic: A 2017 statistic by Transport for London reports that 68% of London's Taxi drivers as white and British compared to only 6% of drivers working for private ride-hailing services such as Uber [120]. Cluster 3 shows a similar distribution of topics in the most negative scored tweets, showing tweets that discuss sexual abuse, but also including tweets negatively highlighting Uber's IPO and finances, further indicating a slight overlap of this cluster with the first two.

If these tweets and their sentiment scores are compared with accounts and their received retweets among the two months, a general relation between the two measurements can be made. All of the authors of the ten tweets with the highest compound scores in cluster 1

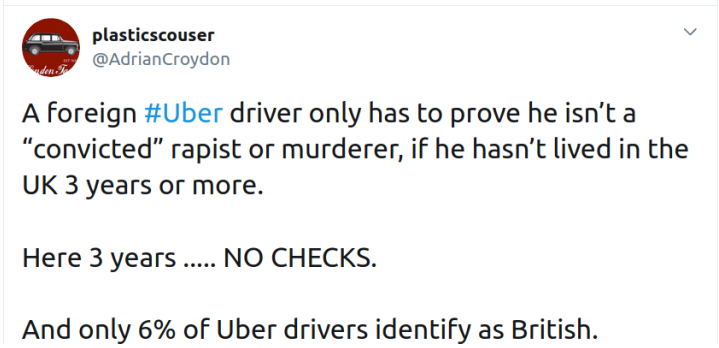


Figure 5.20: Tweet by user "plasticscouser'. The rest of the tweet shows the "Living and Working Abroad Form" by Transport for London, supporting the user's statement. Source: `https://twitter.com/AdrianCroydon/status/1116611237561847808`

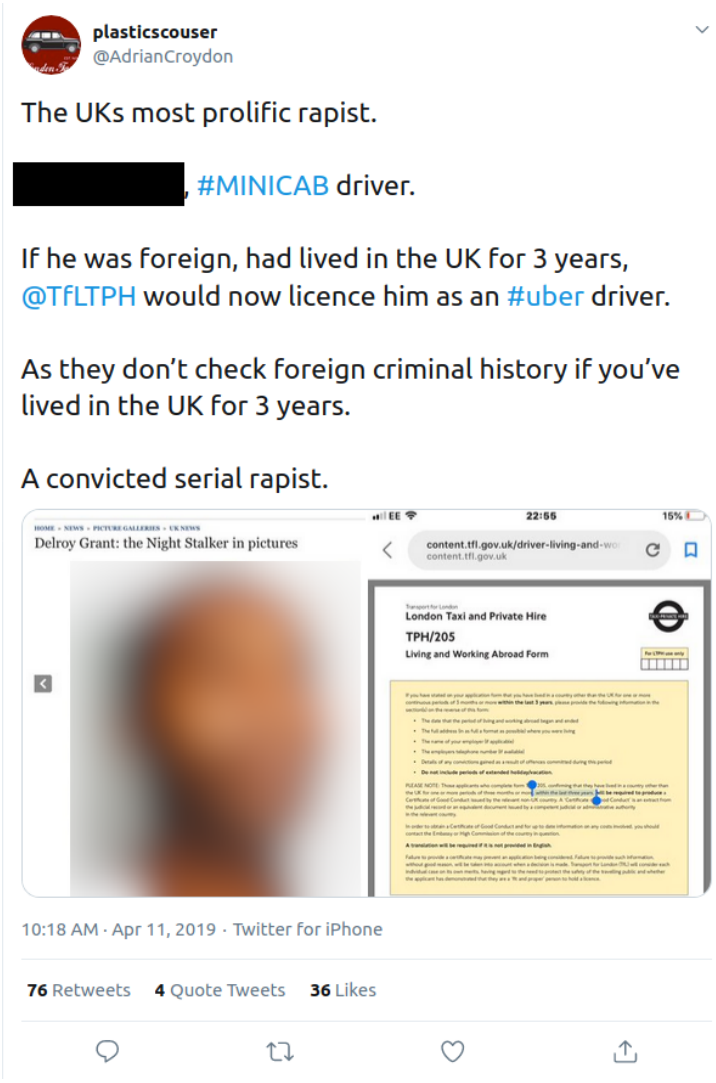


Figure 5.21: Tweet by user "plasticscouser" (Name and face intentionally made anonymous by author). Source: `https://twitter.com/AdrianCroydon/status/1116254126520897537`

can be found among the ten most retweeted users in cluster 1, while all the ten tweets with the lowest compound scores have been posted by users among the eight most retweeted users in the cluster. This suggests that accounts and tweets which express strong emotions or highlight emotional topics, such as sexual harassment and rape are retweeted more

Figure 5.22: Tweet by user "plasticscouser". Source: `https://twitter.com/AdrianCroydon/status/1116587635185573888`

Figure 5.23: Tweet by user "DDDaughters". Source: `https://twitter.com/DDDaughters/status/1127344598148362241`

often. The same pattern can be seen for cluster 2 with a financial topic, as all of the

accounts which posted the ten most negative tweets are among the ten most retweeted, and the users of the ten most positive tweets among the twelve most retweeted accounts. In cluster 2 however, as the only cluster, the most retweeted account "schuldensuehner" with 344 retweets, belonging to Holger Zschaeptitz, editor in chief for economics at "Die Welt" [113], can not be found among either the ten most negative or positive tweets, indicating high neutrality in his language. This is illustrated by Figure 5.25, showing a tweet by Zschaeptitz which was rated by the sentiment analysis with a neutral score of 1.0, while the positive and the negative scores were 0.0, as well as the compound score. The same pattern can be found for cluster 3, with the authors of the ten most positive tweets being among the ten most retweeted and the authors of the ten most negative tweets among the twelve most retweeted users respectively.

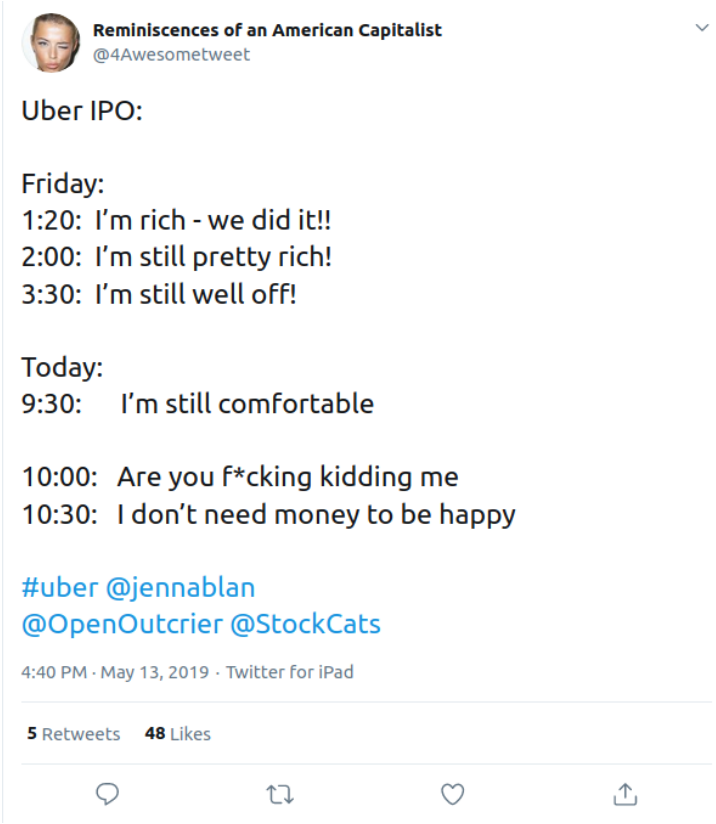


Figure 5.24: Tweet by the account "4awesomeTweets". Source: `https://twitter.com/4AwesomeTweet/status/1127946603883483137`

Figure 5.24 shows a tweet by the account "4awesomeTweets", an account in cluster 1, who describes their profile as "Trading - Investing - Humor - 20+ Years Professional Trader / PM" [121]. The tweet has the highest compound score with 0.9685 in cluster 1 and is rated by the sentiment analysis as having a positive score of 0.394, a neutral score of 0.606 a negative score of 0.0. This rating shows the limits of the sentiment analysis: As this user is talking about the day of Uber's IPO, it is clear that they show disappointment about the falling stock price. In relation to the excitement expressed at the beginning of the tweet, this disappointment should be rated more negatively.

Figure 5.25: Tweet by the account "schuldensuehner" with a neutral score of 1.0. Source: `https://twitter.com/Schuldensuehner/status/1127993557237760000`

Chapter 6

Conclusion

In this thesis, I collected a data set of 51,853 tweets published between April 1 and May 31, 2019, using hashtags related to Uber's IPO on May 10 and the strikes and protests on the days before, organized by various drivers' organizations. Using a combination of filtering and pre-processing methods and an approach in which users and their corresponding tweets over the two months are treated as documents, I reduced the data set consisting of 557 users/documents, which I further analyzed using feature selection via tf-idf followed by PCA dimensionality reduction and k-means clustering.

Results show a clear increase in tweet activity surrounding the date of the IPO and the global strikes and protests preceding it. In an analysis of co-occurring hashtags, I identify the hashtag "#uber" as the central hashtag in terms of occurrence in the data set, with groups of related hashtags such as "#taxi", "#tfl" (Transport for London), and "#uk". In the following community detection approach, I identified three clusters. By analyzing the most used hashtags per cluster, comparing their activity timelines, and reading the most liked tweets in the clusters, I identified differences in their hashtag usage and topic they discuss. I showed that correlation patterns exist between the activity of cluster-related hashtags like "#uberlyftstrike" and "#ipo" and different days of the observed time frame on Twitter. The three clusters can be classified as the first cluster linked to political topics, discussing the Uber and Lyft strikes, a second cluster talking about Uber's success in the stock market, and the company's financial position, as well as a third cluster containing discussions about safety concerns. In a comparative language analysis of the three clusters using word shift graphs, I identified the most distinct differences in word usage, further highlighting the clusters'/communities' different topics. In a valence aware sentiment analysis I showed a correlation between strongly expressed sentiment, both positively and negatively, and the amount of retweets users receive, with statements about sexual abuse allegations against Uber drivers dominating negatively scored tweets. Expressions of Uber-related cultural, political, and legal conflicts, such as the one between traditional London taxi employees and Uber drivers, safety concerns such as an allegation of sexual abuse against drivers, and discussions around fairer wages for gig workers as well as around other Uber-related controversies can be observed in high-valence tweets.

The results of this thesis further indicate that the document-per-user approach followed by dimensionality reduction with PCA as a pre-processing step in combination with

clustering methods can help identify meaningful clusters of topics and corresponding user communities. Correlations between community activity and real-life events emphasize the increasing importance of Twitter as a tool for political communication, expression of opinions, and the formation of a collective identity. In addition to observed online discussions of real-life cultural and political conflicts, these findings indicate an increasingly interdependent relationship between real life and online social platforms.

Future work could also include a focus on the application of SNA methods to further explore the social connections, network topologies, and communication patterns between drivers' organizations, unions, activists, drivers, and users of ridesharing platforms. A combination with other well-proven methods like Latent Dirichlet Allocation (LDA), Dirichlet Multinomial Mixture (DMM), and Latent Semantic Analysis could be of interest as well and could lead to promising insights in terms of topic detection within Uber-related Twitter data. More recent cutting-edge methods such as the approach for political ideology detection using "Multi-task Multi-relational Embeddings" proposed by Xiao et al. [122] have the potential of aiding further research related to online discussions of social, cultural, political, and legal conflicts surrounding not only ridesharing platforms specifically, but the gig and platform economy in general.

Bibliography

[1] J. Woodcock and M. Graham, *The Gig Economy: A Critical Introduction*. Cambridge: Polity, Dec. 2019. [Online]. Available: http://oro.open.ac.uk/68716/

[2] Shane McFeely and Ryan Pendell, "What Workplace Leaders Can Learn From the Real Gig Economy," Aug. 2018. [Online]. Available: https://www.gallup.com/workplace/240929/workplace-leaders-learn-real-gig-economy.aspx

[3] Greg Iacurci, "The gig economy has ballooned by 6 million people since 2010. Financial worries may follow," Feb. 2020. [Online]. Available: https://www.cnbc.com/2020/02/04/gig-economy-grows-15percent-over-past-decade-adp-report.html

[4] "Uber | Sign Up." [Online]. Available: https://www.uber.com/a/join-new

[5] Harry Redhead, "With its foray into facial recognition, Uber is normalising mass surveillance," Jan. 2020. [Online]. Available: https://www.cityam.com/with-its-foray-into-facial-recognition-uber-is-normalising-mass-surveillance/

[6] U. Bajwa, D. Gastaldo, E. Ruggiero, and L. Knorr, "The health of workers in the global gig economy," *Globalization and Health*, vol. 14, Dec. 2018.

[7] S. Gross, G. Musgrave, and L. Janciute, *Well-Being and Mental Health in the Gig Economy*, Aug. 2018.

[8] Sam Levin, "Uber drivers often make below minimum wage, report finds," Mar. 2018. [Online]. Available: https://www.theguardian.com/technology/2018/mar/01/uber-lyft-driver-wages-median-report

[9] Catherine New, "How Much Are People Making From the Sharing Economy?" Mar. 2020. [Online]. Available: https://www.earnest.com/blog/sharing-economy-income-data/

[10] Michael Sainato, "'They treat us like crap': Uber drivers feel poor and powerless on eve of IPO," May 2019. [Online]. Available: https://www.theguardian.com/technology/2019/may/07/uber-drivers-feel-poor-powerless-ipo-looms

[11] Hazel Sheffield, "An Interview with the Uber Driver Who Had Enough," Apr. 2017. [Online]. Available: https://www.vice.com/en_au/article/wn9444/an-interview-with-the-uber-driver-whos-had-enough

[12] "Uber Doesn't Care About Its Drivers," Aug. 2020. [Online]. Available: https://www.nytimes.com/2020/08/19/opinion/letters/uber-workers.html

[13] Annie Nova, "Uber drivers block traffic in Manhattan, protesting low pay and poor working conditions," Sep. 2019. [Online]. Available: https://www.cnbc.com/2019/09/17/uber-drivers-are-protesting-again-heres-what-the-job-is-really-like.html

[14] Liane Yvkoff, "Uber lives on despite SFMTA cease-and-desist," Dec. 2010. [Online]. Available: https://www.cnet.com/roadshow/news/uber-lives-on-despite-sfmta-cease-and-desist/

[15] Malia Spencer, "City sends Uber cease-and-desist order, Uber fires up a petition," Dec. 2014. [Online]. Available: https://www.bizjournals.com/portland/blog/techflash/2014/12/city-sends-uber-cease-and-desist-order-uber-fires.html

[16] Josh Lowensohn, "Uber hit with another cease and desist order, this time in South Carolina," Jan. 2015. [Online]. Available: https://www.theverge.com/2015/1/15/7554561/uber-hit-cease-and-desist-south-carolina

[17] Reuters Staff, "Factbox: Uber's legal challenges around the world," Nov. 2019. [Online]. Available: https://www.reuters.com/article/us-uber-britain-factbox-idUSKBN1XZ25F

[18] Chris Fox, "Uber promises changes to avoid Germany ban," Dec. 2019. [Online]. Available: https://www.bbc.com/news/technology-50865154

[19] Sarah Butler, "Uber loses appeal over driver employment rights," Dec. 2018. [Online]. Available: https://www.theguardian.com/technology/2018/dec/19/uber-loses-appeal-over-driver-employment-rights

[20] Cyrus Farivar, "Uber to pay $20M settlement to drivers, which one expert sees as 'mostly a win for Uber'," Mar. 2019. [Online]. Available: https://www.nbcnews.com/news/us-news/uber-pay-20m-settlement-drivers-which-one-expert-sees-mostly-n982561

[21] Andrew J. Hawkins, "Uber goes public: everything you need to know about the biggest tech IPO in years," May 2019. [Online]. Available: https://www.theverge.com/2019/5/10/18564197/uber-ipo-stock-valuation-pricing-fares-drivers-public-market

[22] Maya Yang, "Uber, Lyft drivers strike in cities worldwide ahead of Uber IPO," May 2019. [Online]. Available: https://www.aljazeera.com/news/2019/05/uber-lyft-drivers-strike-cities-worldwide-uber-ipo-190508120954420.html

[23] Elisabeth Schnulze, "Uber drivers' strike takes off in front of company headquarters ahead of $90 billion IPO," May 2019. [Online]. Available: https://www.cnbc.com/2019/05/08/uber-drivers-strike-over-low-wages-benefits-ahead-of-ipo.html

[24] Megan Rose Dickey, "Uber and Lyft drivers are striking ahead of Uber's IPO," May 2019. [Online]. Available: https://techcrunch.com/2019/05/06/uber-and-lyft-drivers-are-striking-ahead-of-ubers-ipo/

[25] Jonathan Cousar, "Ridester's 2018 Independent Driver Earnings Survey," 2018. [Online]. Available: https://www.ridester.com/2018-survey/

[26] Carmel DeAmicis, "Despite Uber's Arguments, Flexibility for Employees Is a Company's Choice," Aug. 2015. [Online]. Available: https://www.vox.com/2015/8/11/11615468/despite-ubers-arguments-flexibility-for-employees-is-a-companys-choice

[27] United States Securities And Exchange Commission, "Form S-1 Registration Statement Uber Technologies, Inc." Apr. 2019. [Online]. Available: https://www.sec.gov/Archives/edgar/data/1543151/000119312519103850/d647752ds1.htm

[28] Dara Khosrowshahi, "I Am the C.E.O. of Uber. Gig Workers Deserve Better." Aug. 2020. [Online]. Available: https://www.nytimes.com/2020/08/10/opinion/uber-ceo-dara-khosrowshahi-gig-workers-deserve-better.html

[29] The Local, "Uber cancels low-cost UberPop service in Zurich," Aug. 2017. [Online]. Available: https://www.thelocal.ch/20170810/uber-cancels-low-cost-uberpop-service-in-zurich

[30] NZZ, "Uber verzichtet in der Schweiz auf umstrittenen Pop-Service," Dec. 2017. [Online]. Available: https://www.nzz.ch/wirtschaft/uber-verzichtet-in-der-schweiz-auf-umstrittenen-pop-service-ld.1338911?reduced=true

[31] Romain Dillet, "Uber Suspends UberPOP In France Following Turmoils And Arrests," Jul. 2015. [Online]. Available: https://techcrunch.com/2015/07/03/uber-stops-uberpop-in-france-following-turmoils-and-arrests/

[32] "Tear Gas and Burnt Tires in Paris as Taxis Protest Uber," Jan. 2016. [Online]. Available: https://www.telesurenglish.net/news/Tear-Gas-and-Burnt-Tires-in-Paris-as-Taxis-Protest-Uber-20160126-0025.html

[33] Julie Ruvolo, "The Fight Against Uber Is Getting Violent In Brazil," Oct. 2015. [Online]. Available: https://techcrunch.com/2015/10/01/the-fight-against-uber-is-getting-violent-in-brazil/

[34] Christopher Woody, "A protest against Uber in Mexico 'paralyzed the roadways' before turning into a violent street riot," Mar. 2016. [Online]. Available: https://www.businessinsider.com/uber-protest-in-mexico-turns-violent-and-causes-street-riot-2016-3?r=US&IR=T

[35] Liberty Vittert, "Uber's data revealed nearly 6,000 sexual assaults. Does that mean it's not safe?" Dec. 2019. [Online]. Available: https://theconversation.com/ubers-data-revealed-nearly-6-000-sexual-assaults-does-that-mean-its-not-safe-128689

[36] Adriennce Lafrance and Rose Eveleth, "Are Taxis Safer Than Uber?" Mar. 2015. [Online]. Available: https://www.theatlantic.com/technology/archive/2015/03/are-taxis-safer-than-uber/386207/

[37] Dave Lee, "Uber's mess reaches beyond sexism - and Silicon Valley," Feb. 2017. [Online]. Available: https://www.bbc.com/news/technology-39065526

[38] Sam Levin, "Uber launches 'urgent investigation' into sexual harassment claims," Feb. 2017. [Online]. Available: https://www.theguardian.com/technology/2017/feb/20/uber-urgent-investigation-sexual-harassment-claims-susan-fowler

[39] Jackie Wattles, "Uber moves to settle rape victim's lawsuit," Dec. 2017. [Online]. Available: https://money.cnn.com/2017/12/09/technology/uber-settlement-india-rape-case/index.html

[40] German Lopez, "Why people are deleting Uber from their phones after Trump's executive order," Jan. 2017. [Online]. Available: https://www.vox.com/policy-and-politics/2017/1/29/14431246/uber-trump-muslim-ban

[41] Susan J. Fowler, "Reflecting on one very, very strange year at Uber," Feb. 2017. [Online]. Available: https://www.susanjfowler.com/blog/2017/2/19/reflecting-on-one-very-strange-year-at-uber

[42] Kara Swisher, "Uber's SVP of engineering is out after he did not disclose he left Google in a dispute over a sexual harassment allegation," Feb. 2017. [Online]. Available: https://www.vox.com/2017/2/27/14745360/amit-singhal-google-uber

[43] Julia Carrie Wong, "Uber CEO steps down from Trump advisory council after users boycott," Feb. 2017. [Online]. Available: https://www.theguardian.com/technology/2017/feb/02/travis-kalanick-delete-uber-leaves-trump-council

[44] ——, "Embattled Uber CEO Travis Kalanick takes indefinite leave of absence," Jun. 2017. [Online]. Available: https://www.theguardian.com/technology/2017/jun/13/uber-ceo-travis-kalanick-leave-absence-scandal

[45] Matthew J. Belvedere, "'Moral compass' was off at Uber under co-founder Kalanick, says new CEO Dara Khosrowshahi," Jan. 2018. [Online]. Available: https://www.cnbc.com/2018/01/23/uber-moral-compass-under-co-founder-kalanick-was-off-new-ceo-says.html

[46] Uber Team, "Uber's New CEO," Aug. 2017. [Online]. Available: https://www.uber.com/newsroom/ubers-new-ceo-3/

[47] Maureen Farrell, "Uber Aims for Public Valuation of as Much as $100 Billion, Below Expectations," Apr. 2019. [Online]. Available: https://www.wsj.com/articles/uber-aims-for-public-valuation-of-as-much-as-100-billion-below-expectations-11554915215?mod=article_inline

[48] David Trainer and Great Speculations, "Uber's IPO Valuation Makes No Sense," Apr. 2019. [Online]. Available: https://www.forbes.com/sites/greatspeculations/2019/04/22/ubers-ipo-valuation-makes-no-sense/#406e1f1b540d

[49] Theron Mohamed, "Uber is paying drivers up to $40,000 each to celebrate its IPO," Apr. 2019. [Online]. Available: https://markets.businessinsider.com/news/stocks/initial-public-offering-uber-rewarding-drivers-2019-4-1028143191

[50] Kate Clark, "Uber is finally trading above its IPO price," Jun. 2019. [Online]. Available: https://techcrunch.com/2019/06/05/uber-is-finally-trading-above-its-ipo-price/

[51] Alexia Fernández Campbell, "The worldwide Uber strike is a key test for the gig economy," May 2019. [Online]. Available: https://www.vox.com/2019/5/8/18535367/uber-drivers-strike-2019-cities

[52] Rideshare Drivers United, "Rideshare Drivers United Homepage." [Online]. Available: https://act.drivers-united.org/

[53] Boston Independent Drivers Guild, "Boston Independent Drivers Guild Homepage." [Online]. Available: https://bidg.org/

[54] Chicago Rideshare Advocates, "Chicago Rideshare Advocates Homepage." [Online]. Available: https://chicagorideshareadvocates.org/

[55] Drive United, "Drive United Homepage." [Online]. Available: https://driveunited.org

[56] Gig Workers Collective, "Gig Workers Collective Homepage." [Online]. Available: https://www.gigworkerscollective.org/

[57] Gig Workers Rising, "Gig Workers Rising Homepage." [Online]. Available: https://gigworkersrising.org/

[58] Kate Conger, Vicky Xiuzhong Xu, and Zach Wichter, "Uber Drivers' Day of Strikes Circles the Globe Before the Company's I.P.O." May 2019. [Online]. Available: https://www.nytimes.com/2019/05/08/technology/uber-strike.html

[59] Bryan Menegus, "New York's Rideshare Organizers Clash Amid Unprecedented Uber Strike," May 2019. [Online]. Available: https://gizmodo.com/new-yorks-rideshare-organizers-clash-amid-unprecedented-1834623838

[60] John Haltiwanger, "Bernie Sanders says he stands in 'solidarity' with Uber and Lyft drivers going on strike," May 2017. [Online]. Available: https://www.businessinsider.com/bernie-sanders-applauds-uber-lyft-driver-strike-new-york-los-angeles-2019-5?r=US&IR=T

[61] Emily Birnbaum, "Ocasio-Cortez throws weight behind Uber, Lyft strike," May 2017. [Online]. Available: https://thehill.com/policy/technology/442731-ocasio-cortez-throws-weight-behind-uber-lyft-strike

[62] Y. Theocharis, "The Wealth of (Occupation) Networks? Communication Patterns and Information Distribution in a Twitter Protest Network," *Journal of Information Technology & Politics*, vol. 10, pp. 35–56, 2013.

[63] T. Hachaj and M. R. Ogiela, "Clustering of trending topics in microblogging posts: A graph-based approach," *Future Generation Computer Systems*, vol. 67, pp. 297 – 304, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X16300863

[64] V. Vargas-Calderón and J. E. Camargo, "Characterization of citizens using word2vec and latent topic analysis in a large set of tweets," *Cities*, vol. 92, pp. 187 – 196, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0264275119300599

[65] A. Benny and M. Philip, "Keyword Based Tweet Extraction and Detection of Related Topics," *Procedia Computer Science*, vol. 46, pp. 364 – 371, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050915000964

[66] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient Estimation of Word Representations in Vector Space*, 2013, _eprint: 1301.3781.

[67] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001, publisher: Wiley Online Library.

[68] A. Klahold, P. Uhr, M. Fathi, and F. Ansari, "A Framework to utilize the Human Ability of Word Association for detecting Multi Topic Structures in Text Documents," *Intelligent Systems, IEEE*, vol. 29, 2014.

[69] A. Rafea and N. Gaballah, "Topic extraction in social media," May 2013, pp. 94–98.

[70] D. Surian, D. Q. Nguyen, G. Kennedy, M. Johnson, E. Coiera, and A. G. Dunn, "Characterizing Twitter Discussions About HPV Vaccines Using Topic Modeling and Community Detection," *J Med Internet Res*, vol. 18, no. 8, p. e232, Aug. 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27573910

[71] K. S. Adewole, T. Han, W. Wu, H. Song, and A. K. Sangaiah, "Twitter spam account detection based on clustering and classification methods," *The Journal of Supercomputing*, vol. 76, no. 7, pp. 4802–4837, 2020, publisher: Springer.

[72] E. Bokányi, D. Kondor, L. Dobos, T. Sebok, J. Stéger, I. Csabai, and G. Vattay, "Race, Religion and the City: Twitter Word Frequency Patterns Reveal Dominant Demographic Dimensions in the United States," *ArXiv*, vol. abs/1605.02951, 2016.

[73] L. Lommel, M. Schreier, and J. Fruchtmann, "We Strike, Therefore We Are? A Twitter Analysis of Feminist Identity in the Context of #DayWithoutAWoman," *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, vol. 20, no. 2, 2019. [Online]. Available: https://www.qualitative-research.net/index.php/fqs/article/view/3229

[74] Erica Chenoweth and Jeremy Pressman, "This is what we learned by counting the women's marches," Feb. 2017. [Online]. Available: https://www.washingtonpost.com/news/monkey-cage/wp/2017/02/07/this-is-what-we-learned-by-counting-the-womens-marches/

[75] Sally Ann Drucker, "Betty Friedan: The Three Waves of Feminism - Ohio Humanities," Apr. 2018. [Online]. Available: http://www.ohiohumanities.org/betty-friedan-the-three-waves-of-feminism/

[76] M. Åkerlund, "The importance of influential users in (re)producing Swedish far-right discourse on Twitter," *European Journal of Communication*, vol. 0, no. 0, p. 0267323120940909, 2020, _eprint: https://doi.org/10.1177/0267323120940909. [Online]. Available: https://doi.org/10.1177/0267323120940909

[77] C. Fernández-Rovira and I. Villegas-Simón, "Comparative study of feminist positioning on Twitter by Spanish politicians," 2019.

[78] M. Cini and N. P.-S. Borragán, *European union politics*. Oxford University Press, 2016.

[79] P. N. Diamandouros, R. Gunther, and others, *Parties, politics, and democracy in the New Southern Europe*. JHU Press, 2001.

[80] A. Hannak, D. Margolin, B. Keegan, and I. Weber, "Get Back! You don't know me like that: The social mediation of fact checking interventions in twitter conversations," *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, pp. 187–196, 2014.

[81] Snopes Media Group Inc., "Snopes.com - The definitive fact-checking site and reference source for urban legends, folklore, myths, rumors, and misinformation." [Online]. Available: https://www.snopes.com/

[82] Chris Cantey, "GeoSearch-Tweepy." [Online]. Available: https://github.com/Ccantey/GeoSearch-Tweepy

[83] "Geo Objects." [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview/geo-objects

[84] "TWINT - Twitter Intelligence Tool." [Online]. Available: https://github.com/twintproject/twint

[85] "TWINT - Twitter Intelligence Tool Wiki," Aug. 2019. [Online]. Available: https://github.com/twintproject/twint/wiki

[86] Biz Stone, "Introducing the Twitter API," Sep. 2006. [Online]. Available: https://blog.twitter.com/2006/introducing-the-twitter-api

[87] Ben Parr, "Twitter Launches Countdown to OAuthcalypse," Apr. 2010. [Online]. Available: http://mashable.com/2010/04/24/twitter-oauthcalypse/

[88] Zee, "Twitter to launch its own URL shortener," Apr. 2010. [Online]. Available: https://thenextweb.com/apps/2010/04/15/twitter-launch-link-shortener

[89] Kimber Streams, "Tweetro says it's 'completely crippled' by Twitter's strict 100,000 user token limit," Nov. 2012. [Online]. Available: https://www.theverge.com/2012/11/11/3631108/Tweetro-user-token-limit-api

[90] Anthony Ha, "Twitter Handcuffs Client Apps With New API Changes," Aug. 2012. [Online]. Available: https://techcrunch.com/2012/08/16/twitter-api-client-apps

[91] "Filtered Stream v1." [Online]. Available: https://developer.twitter.com/en/docs/labs/filtered-stream/overview

[92] "Search tweets." [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/overview

[93] "Twitter Subscription Levels." [Online]. Available: https://developer.twitter.com/en/products/twitter-api

[94] "Rate limits." [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/v1/rate-limits

[95] "Premium search | Docs | Twitter Developer." [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/overview/premium

[96] "NLTK 3.5 documentation." [Online]. Available: https://www.nltk.org

[97] "nltk.tokenize package — NLTK 3.5 documentation." [Online]. Available: https://www.nltk.org/api/nltk.tokenize.html#nltk.tokenize.casual.TweetTokenizer

[98] "sklearn.feature_extraction.text.TfidfVectorizer — scikit-learn 0.23.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[99] "sklearn.decomposition.PCA — scikit-learn 0.23.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[100] "sklearn.cluster.KMeans — scikit-learn 0.23.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[101] Ryan Gallagher, Ronald E. Robertson, and Liz McQuillan, "Shifterator." [Online]. Available: https://github.com/ryanjgallagher/shifterator

[102] R. J. Gallagher, M. R. Frank, L. Mitchell, A. J. Schwartz, A. J. Reagan, C. M. Danforth, and P. S. Dodds, *Generalized Word Shift Graphs: A Method for Visualizing and Explaining Pairwise Comparisons Between Texts*, 2020, _eprint: 2008.02250.

[103] C.J. Hutto, "VADER-Sentiment-Analysis." [Online]. Available: https://github.com/cjhutto/vaderSentiment

[104] "nltk.sentiment.vader — NLTK 3.5 documentation." [Online]. Available: https://www.nltk.org/_modules/nltk/sentiment/vader.html

[105] J. Clement, "• Twitter: most users by country | Statista," Jul. 2020. [Online]. Available: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries

[106] "Keeping London moving - Transport for London." [Online]. Available: https://tfl.gov.uk/

[107] Kate Gibson, "Uber and Lyft drivers plan 24-hour strike to protest pay - CBS News," Apr. 2019. [Online]. Available: https://www.cbsnews.com/news/uber-lyft-drivers-plan-24-hour-strike-to-protest-pay

[108] David Lawder and Ben Blanchard, "Worries of longer, costlier U.S.-China trade war hits markets | Reuters," May 2019. [Online]. Available: https://www.reuters.com/article/us-usa-trade-china-idUSKCN1SQ0XQ

[109] Urban Dictionary, "Urban Dictionary: Ubered." [Online]. Available: https://www.urbandictionary.com/define.php?term=Ubered

[110] Graham Rapier, "Uber IPO: Ride-hailing giant files for biggest stock offering in years - Business Insider," Apr. 2019. [Online]. Available: https://www.businessinsider.com/uber-ipo-documents-filed-what-could-be-biggest-stock-offering-years-2019-4

[111] United States Securities And Exchange Commission, "Amendment No. 1 to Form S-1 Registration Statement Uber Technologies, Inc." Apr. 2019. [Online]. Available: https://www.sec.gov/Archives/edgar/data/1543151/000119312519120759/d647752ds1a.htm

[112] "eToro - The World's Leading Social Trading and Investing Platform." [Online]. Available: https://www.etoro.com/

[113] "Holger Zschäpitz: Artikel, Kontakt & Profil - Autorenseite - WELT." [Online]. Available: https://www.welt.de/autor/holger-zschaepitz/

[114] Ryan J. Gallagher, "Frequency-Based Shifts — Shifterator documentation." [Online]. Available: https://shifterator.readthedocs.io/en/latest/cookbook/frequency_shifts.html

[115] "Nicole Moore (@nicoemoe) / Twitter." [Online]. Available: https://twitter.com/nicoemoe

[116] Bryan Curtis, "Strange days at the Daily Telegraph," Oct. 2006. [Online]. Available: http://www.slate.com/articles/news_and_politics/letter_fromlondon/2006/10/paper_tiger.html

[117] "plasticscouser on Twitter: "If you take advantage of underpaid, exploited foreigners to save yourself money, and then report them to their "employer" when they fail to meet your expectations.....YOU are the wonderful racist bastard. https://t.co/fSuCAzcM4Q" / Twitter," Aug. 2020. [Online]. Available: https://twitter.com/AdrianCroydon/status/1291107389895303173

[118] "plasticscouser on Twitter: "@StrongbowsPub @DaveScoff @SovereignSally @ukhomeoffice @Nigel_farage Really !!! 35,000 asylum seekers last year. 84,000 in 2002. Shall we say 50,000 per year ? Cost £10,000 each per year. £500 million per year. Or 10,000 extra nurses and 10,000 extra Police Officers per year. https://t.co/f3JyTpRtvX" / Twitter," Aug. 2020. [Online]. Available: https://twitter.com/AdrianCroydon/status/1295810913082183682

[119] "#DDD (@DDDaughters) / Twitter." [Online]. Available: https://twitter.com/DDDaughters

[120] Yasmin Serhan, "London's Uber Beef Has Deep Cultural Roots," Sep. 2017. [Online]. Available: https://www.theatlantic.com/technology/archive/2017/09/the-tension-between-london-and-uber/540852/

[121] "Reminiscences of an American Capitalist (@4Awesometweet) / Twitter." [Online]. Available: https://twitter.com/4Awesometweet

[122] Z. Xiao, W. Song, H. Xu, Z. Ren, and Y. Sun, "TIMME," *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Aug. 2020, iSBN: 9781450379984 Publisher: ACM. [Online]. Available: http://dx.doi.org/10.1145/3394486.3403275

Abbreviations

IPO	Initial Public Offering
VADER	Valence Aware Dictionary and sEntiment Reasoner
NLP	Natural Language Processing
PCA	Principal Component Analysis
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
CEO	Chief Executive Officer
IDG	Independent Drivers Guild
NYTWA	New York Taxi Drivers Alliance
SNA	Social Network Analysis
UCL	University College London
AGF	Associated Gravity Force
CIMAWA	Concept for the Imitation of the Mental Ability of Word Association
DMM	Dirichlet Multinomial Mixture
SVA	Support Vector Machine
API	Application Programming Interface
SD	Sweden Democrats
JSON	JavaScript Object Notation
GPS	Global Positioning System
OSINT	Open Source Intelligence
NLTK	Natural Language Toolkit
TfL	Transport for London
PSEO	Partido Socialista Obrero Español

List of Figures

List of Tables

Appendix A

Data

Twitter prohibits the public distribution of data sets containing full-text information about Tweets, but the repository (`https://github.com/MethDamon/uber_thesis`) contains a "pickled" (`https://wiki.python.org/moin/UsingPickle`) version of the tf-idf matrix. PCA and k-means can then be directly performed on the pickled version of the matrix. All of the data analysis can also be performed using any other data set which was collected with the Twint library and is in the the same data format as described in Section 4.1 of Chapter 4. Additionally, the script shown in Listing 3 can be used to collect the data set used in the thesis. It is also available via the GitHub repository.

Appendix B

Code

The full code base for this thesis can be found online: `https://github.com/MethDamon/uber_thesis`

```
import twint

c = twint.Config()
c.Since = '2019-04-01'
c.Until = '2019-05-31'
c.Debug = True
c.Lang = 'en'
c.Hide_output = False
c.Store_json = True
c.Search = """
    #uberride OR #uberdriver OR #uber OR #uberlyftstrike
    OR #uberstrike OR #ubershutdown OR #lyft OR #lyftstrike
    OR #uberipo OR #driversunite OR #strikeuberlyft OR
    #driverslivesmattertoo"""
c.Store_object = False
c.Stats = True
c.Count = True
c.Filter_retweets = True
c.User_full = True
c.Output = 'data.txt'

twint.run.Search(c)
```

Listing 3: The Python script used for data collection.

```
import re
import string
```

```
3
4  import nltk
5  from nltk.corpus import stopwords
6  from nltk.tokenize import TweetTokenizer
7  from nltk.stem import WordNetLemmatizer
8  from nltk.corpus import wordnet
9
10 stop_words = stopwords.words('english')
11 tokenizer = TweetTokenizer()
12 lemmatizer = WordNetLemmatizer()
13
14 def word_tokenize(tweet):
15     return tokenizer.tokenize(tweet)
16
17 def get_wordnet_pos(word):
18     tag = nltk.pos_tag([word])[0][1][0].upper()
19     tag_dict = {"J": wordnet.ADJ,
20                 "N": wordnet.NOUN,
21                 "V": wordnet.VERB,
22                 "R": wordnet.ADV}
23
24     return tag_dict.get(tag, wordnet.NOUN)
25
26 def remove_emoji(string):
27     emoji_pattern = re.compile("["
28                             u"\U0001F600-\U0001F64F"  # emoticons
29                             u"\U0001F300-\U0001F5FF"  # symbols &
                                ↪ pictographs
30                             u"\U0001F680-\U0001F6FF"  # transport & map
                                ↪ symbols
31                             u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
32                             u"\U00002702-\U000027B0"
33                             u"\U000024C2-\U0001F251"
34                             "]+", flags=re.UNICODE)
35     return emoji_pattern.sub(r'', string)
36
37 def preprocess_tweet_text(tweet):
38     tweet = tweet.lower()
39     #Remove numbers
40     tweet = re.sub(r"[0-9]", '', tweet, flags=re.MULTILINE)
41     # Remove urls
42     tweet = re.sub(r"http\S+|www\S+|https\S+", '', tweet,
       ↪ flags=re.MULTILINE)
43     # Remove punctuation
44     tweet = tweet.translate(str.maketrans('', '', string.punctuation))
45     # Remove emojis
```

```
    tweet = remove_emoji(tweet)
    # Tokenize
    tweet_tokens = word_tokenize(tweet)
    # Remove stopwords
    filtered_words = [w for w in tweet_tokens if not w in stop_words]
    # Lemmatizing
    lemmatized = [lemmatizer.lemmatize(w, get_wordnet_pos(w)) for w in
    ↪  filtered_words]
    # Remove non-alpha words
    lemmatized_filtered = [w for w in lemmatized if w.isalpha()]
    # Filter out short lemmas
    final_tokens = [w for w in lemmatized_filtered if len(w) > 2]

    return final_tokens
```

Listing 4: The Python script used for pre-processing.