

MASTER'S THESIS

Understanding the Role of Diasporas as Proponents of Disinformation during the Start of the Russo-Ukrainian conflict

de Ridder, C.

Award date:
2022

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

Take down policy

If you believe that this document breaches copyright please contact us at:

pure-support@ou.nl

providing details and we will investigate your claim.

Downloaded from <https://research.ou.nl/> on date: 02. Jul. 2022

Open Universiteit
www.ou.nl



**UNDERSTANDING THE ROLE OF DIASPORAS AS
PROponents OF DISINFORMATION DURING THE
START OF THE RUSSO-UKRAINIAN CONFLICT**
MASTER'S THESIS

Author **Christiaan de Ridder**
Student number
Course code **IM9906**
Thesis committee **Dr. Ir. Clara Maathuis (first supervisor),** Open University
Dr. Ir. Sylvia Stuurman (second supervisor), Open University





Title **Understanding the Role of Diasporas as Proponents of Disinformation during the Start of the Russo-Ukrainian conflict**

Author **Christiaan de Ridder**

Student number

Course code **IM9906**

Thesis presentation **11-02-2022**

Thesis committee **Dr. Ir. Clara Maathuis (first supervisor),** Open University
Dr. Ir. Sylvia Stuurman (second supervisor), Open University

Open University of The Netherlands, Faculty of Science
Master's Programme in Software Engineering

Samenvatting

Door een plotselinge politieke koerswijziging in 2014, sloeg onrust in Oekraïne snel om in groot-schalige protesten welke uiteindelijk resulteerden in de annexatie van het Krim-schiereiland door Rusland, een gewapend conflict tussen Oekraïne en door Rusland gesteunde rebellen in Oost-Oekraïne en het neerschieten van vlucht MH17. Hoewel het Rusland-Oekraïne-conflict in een impasse verkeert, lopen de spanningen momenteel weer op met politieke botsingen en beschuldigingen tussen de westerse landen en Rusland.

Informatie-oorlogsvoering is niet nieuw, maar in het tijdperk van sociale media is het gemakkelijker dan ooit om een publiek te bereiken en hun mening te beïnvloeden. Gedurende de afgelopen acht jaar hebben er veel strategische informatieoperaties plaatsgevonden binnen de context van het Rusland-Oekraïne, waarvan een groot deel zich afspeelde op sociale media.

Onderzoek naar de oorsprong, kenmerken en verspreiding van gerichte desinformatie kan belangrijke inzichten opleveren en de weerbaarheid verbeteren tegen dergelijke strategische informatieoperaties (i.e. de inspanningen van individuen en groepen, zowel overheids- als niet-overheidsacties, om de publieke opinie te manipuleren en de perceptie van gebeurtenissen in de wereld te veranderen door opzettelijke aanpassingen aan de informatieomgeving). Vanwege de grote hoeveelheden gegevens die hiermee gemoeid zijn, is er veel academisch onderzoek naar het oplossen van deze uitdagingen vanuit een computerwetenschappelijk perspectief, vaak met behulp van data scraping en machine learning-technieken.

Dit onderzoek richt zich op het gebied waar strategische informatieoperaties en het aanhoudende Russisch-Oekraïense conflict elkaar ontmoeten. Specifieker gezien, richt het zich op de vraag of leden van de diaspora van beide landen een significante rol spelen in de verspreiding van desinformatiecampagnes. Dit laatste biedt een specifieke en unieke context ten opzichte van bestaand onderzoek naar nepnieuws, desinformatie en/of informatie-oorlogsvoering.

Dit wordt bereikt door eerst een diepgaande context te geven van het Russisch-Oekraïne-conflict en desinformatiecampagnes. Dit wordt vervolgens uitgebreid door een samenvatting te geven van gerelateerd werk op het gebied van desinformatiedetectie en verspreiding als ook het profileren van socialmediagebruikers. Door de beschikbaarheid van gegevens en de beperkingen van verschillende sociale mediaplatforms te vergelijken, is het Engelstalige deel van Twitter geselecteerd als een socialemediaplatform waarop het onderzoek is uitgevoerd. Hieruit is de volgende primaire onderzoeksvraag geformuleerd:

Wat is de interactie tussen gebruikers, die als leden van de diaspora worden beschouwd, en bekende desinformatiecampagnes op sociale media in verband met het Rusland-Oekraïne-conflict?

Deze vraag is verder opgesplitst in vier onderliggende onderzoeksvragen:

RQ1. Welke desinformatiecampagne rondom het Rusland-Oekraïne-conflict is geschikt om te gebruiken voor onderzoek naar interactie door leden van de diaspora?

RQ2. Welke Twitter-gebruikersgemeenschappen hebben interactie met de geselecteerde desinformatiecampagne rondom het Rusland-Oekraïne-conflict?

RQ3. Welke machine learning-methoden zijn geschikt voor het voorspellen van diaspora-lidmaatschap op basis van Twitter-gebruikersprofielen?

RQ4. Hoe verhoudt diaspora-lidmaatschap zich tot gedetecteerde Twitter-gebruikersgemeenschappen?

Elk van de gestelde vragen wordt benaderd vanuit een computerwetenschappelijk perspectief, door gebruik te maken van technieken voor gegevensverzameling en scraping, netwerkanalyse en gebruikersclassificatie met behulp van machine learning-technieken. Deze aanpak resulteert in de volgende belangrijke bevindingen:

- Factcheckers kunnen worden gebruikt om nieuwsartikelen te vinden waar desinformatie voor het eerst werd verspreid. Door het gebruik van URL-verkorters is het echter moeilijk om rechtstreeks direct te zoeken naar Twitterberichten die naar deze artikelen verwijzen. Om aanvullende resultaten te verkrijgen moesten aanvullende zoekopdrachten met trefwoorden op basis van artikelstitels worden uitgevoerd.
- Mention networks bieden een handige manier om gemeenschappen binnen het socialemediadiscours te identificeren en het sentiment in elk van deze gemeenschappen te bestuderen. Deze methode geeft echter geen bruikbare inzichten in kleinere datasets zoals Data set 2.
- Random Forests presteren goed voor het classificeren van diaspora-lidmaatschap door middel van statische gebruikersprofielgegevens en latente eigenschappen op basis van statische gebruikersprofielgegevens van gevolgde accounts.
- Op basis van de in dit onderzoek gebruikte datasets spelen Diasporaleden een duidelijke rol in het algemene discours rond het Rusland-Oekraïne-conflict, maar spelen zij geen actieve rol in de verspreiding van desinformatie op het Engelstalige deel van Twitter.

Op basis van deze bevindingen wordt de algemene conclusie getrokken dat zowel leden van de Russische als Oekraïense diaspora een duidelijke rol spelen in het discours rond het Russisch-Oekraïense conflict als geheel, maar dat een duidelijke rol bij het verspreiden van desinformatie niet kon worden geïdentificeerd met behulp van de gegevens en methoden die in het kader van dit onderzoek zijn gebruikt.

Naast de gepresenteerde bevindingen wordt er aanvullend onderzoek voorgesteld op vier gebie-

den:

- Alternatieve methoden voor het verzamelen en archiveren van gegevens op sociale media zijn nodig om beter onderzoek te kunnen doen naar historische data zonder veel van de door recente data geboden context te verliezen. Dergelijk onderzoek voorkomt ook de huidige trend van verminderde beschikbaarheid van gegevens op veel sociale-mediaplatforms.
- Een gedragspatroon dat wordt gezien in de datasets die in dit onderzoek zijn gebruikt, is dat gebruikers die desinformatie verspreiden, deze meestal niet aan hun eigen netwerk richten, maar aan (westerse) media en organisaties.
- Er is een sterk ethisch kader nodig voor onderzoek met betrekking tot gebruikersprofilering, met name onderzoek naar de etnische groep van een gebruiker.
- Er zijn veel manieren waarop diasporaleden (zowel bewust als onbewust) een rol kunnen spelen tijdens een conflict in hun land van herkomst. Dit onderzoek richt zich alleen op de interactie met desinformatie op sociale media en laat de andere, potentieel interessante, gebieden onontgonnen.

Summary

Due to a sudden change in political direction in 2014, unrest in Ukraine quickly turned into large-scale protests and eventually the annexation of the Crimean peninsula by Russia, an armed conflict between Ukraine and Russia-backed rebels in eastern Ukraine, and the shoot-down of flight MH17. While the Russo-Ukrainian conflict is currently at a stalemate, tensions are on the rise again with political clashes and posturing between the western countries and Russia.

Information warfare is not new, but in the age of social media, it has become easier than ever to reach an audience and influence their opinions. During the past eight years, this conflict has seen many strategic information operations, with a large part of them playing out on social media.

Research into the origin, characteristics, and spread of targeted disinformation can provide important insights and improve resilience against such strategic information operations (i.e. the efforts by individuals and groups, including state and non-state actions, to manipulate public opinion and change how people perceive events in the world by intentionally altering the information environment). Due to the large volumes of data involved, there is a lot of academic research focusing on approaching these challenges from a computer science perspective, often using data scraping and machine learning techniques.

This research focuses on the area where strategic information operations and the ongoing Russo-Ukrainian conflict meet. More specifically it focuses on whether members of the diaspora of either country plays any significant role in the spread of disinformation campaigns. The latter part provides a specific and unique context compared to existing research into fake news, disinformation, or information warfare.

This is achieved by first providing an in-depth context of the Russo-Ukrainian conflict and disinformation campaigns. This is then expanded on by summarizing related work in the area of disinformation detection, spread, and user profiling. By comparing data availability and limitations of different social media platforms, English language Twitter is selected as a social media platform on which to conduct the research. Based on these steps the following primary research question is formulated:

What is the interaction between users, considered to be diaspora members, and known disinformation campaigns on social media related to the Russo-Ukrainian conflict?

This question is then further split up into four underlying research questions:

- RQ1. Which known disinformation campaign surrounding the Russo-Ukrainian conflict is suitable to be analyzed for Diaspora interaction?**
- RQ2. Which Twitter user communities exist interacting with the selected disinformation campaign surrounding the Russo-Ukrainian conflict?**
- RQ3. Which machine learning methods are suitable for predicting diaspora membership from Twitter user profiles?**
- RQ4. How does diaspora membership relate to detected Twitter user communities?**

Each of the raised questions is approached from a computer science perspective, by employing data collection and scraping techniques, network analysis, and user classification using machine learning techniques. This approach results in the following key findings being presented:

- Fact-checkers can be used to find news articles where disinformation was first spread. However, due to the usage of URL shorteners, it is hard to directly gather data from Twitter for any references to these articles. Further keyword searches based on article titles and claims had to be performed.
- Mention networks provide a useful way to identify communities within social media discourse and study the further sentiment in each of these communities. However, this method does not provide very useful insights in smaller data sets like Data set 2.
- Random Forests perform well for classifying diaspora membership based on static user profile data and latent properties based on static user profile data of followed accounts.
- Diaspora members play a clear and distinct role in the general discourse surrounding the Russo-Ukrainian conflict, but do not play an active role in the spread of disinformation on English-language Twitter, given the data sets available for this research.

Based on these findings the overall conclusion is reached that both Russian and Ukrainian of diaspora members have a clear role in the discourse surrounding the Russo-Ukrainian conflict as a whole, but that a distinct role in spreading disinformation could not be identified using the data and methods used in the scope of this research.

In addition to the key findings presented, further research is suggested in four areas:

- Alternative ways of collecting and archiving social media data are needed to be able to better conduct research on historical data without losing much of the context that recent data provides. This also counteracts the trend towards reduced data availability across many social media platforms.
- A behavioral pattern seen within the data sets used in this research, is that users spreading disinformation usually do not target it at their own network by mentioning those users, but at (Western-aligned) news outlets and organizations.

- A strong ethical framework is needed for research related to user profiling, especially research looking at the ethnic group of a user.
- There are many ways in which diaspora members can (both consciously and unconsciously) play a role during a conflict in their country of origin. This research only focuses on the interaction with disinformation on social media and leaves the other, potentially interesting, areas unexplored.

Contents

Dutch Summary	ii
Summary	v
List of Figures	xiii
List of Tables	xv
1 Introduction	1
1.1 Motivation	2
1.2 Goal	2
1.3 Definitions	3
1.3.1 Diaspora	3
1.3.2 Types of information	3
1.3.3 Disinformation campaigns	4
1.4 Document outline	5
2 The Russo-Ukrainian conflict	6
2.1 The ethnic composition of Ukraine	7
2.2 Ukraine as a <i>borderland</i>	7
2.3 A country in conflict	7
2.4 The Russian annexation of Crimea	8
2.5 The war in Donbas	8
3 Disinformation campaigns	9
3.1 A western perspective	10
3.2 Traditional media	10
3.3 Embracing social media	10
4 Related work	12
4.1 Disinformation detection and spread	13
4.1.1 Propagation-based detection methods	15
4.1.2 Style-based detection methods	15

4.1.3	Source-based detection methods	15
4.1.4	Conclusion	16
4.2	Disinformation data sets	16
4.2.1	Conclusion	18
4.3	Social media profiling	18
4.3.1	Conclusion	20
5	Research	21
5.1	Research methodology	22
5.2	Research questions	22
5.2.1	RQ Primary research question	22
5.2.2	RQ1 Disinformation campaigns	23
5.2.3	RQ2 Social media communities	23
5.2.4	RQ3 Diaspora membership	24
5.2.5	RQ4 Diaspora interaction	25
5.3	Limitations	26
5.3.1	Bias in related works	26
5.3.2	Social media platforms	26
5.3.3	Definition of diaspora	26
5.4	Machine learning techniques	27
5.4.1	Machine learning paradigms	27
5.4.1.1	Supervised learning	27
5.4.1.2	Unsupervised learning	28
5.4.1.3	Reinforcement learning	28
5.4.2	Applicability	28
6	Identifying suitable disinformation campaigns	30
6.1	Scope and context	31
6.1.1	Time frame	31
6.1.2	Data sources	31
6.1.3	State-controlled media	31
6.1.4	Fact-checkers	32
6.2	Data collection process	32
6.2.1	Gathering disinformation sources	32
6.2.2	Twitter interactions	33
6.2.3	Source URLs and variants	34
6.2.4	URL-shorteners	34
6.2.5	Secondary sources	35
6.2.6	Searching Twitter	36
6.3	Results	37
6.3.1	Donbas data set (1)	37
6.3.2	MH17 data set (2)	37

6.4	Conclusion	38
7	Applying mention networks	39
7.1	Mention network	40
7.2	Visualization	41
7.3	Noise reduction	44
7.4	Community detection	47
7.4.1	Results	47
7.5	Text analysis	49
7.5.1	Results	50
7.6	URL analysis	53
7.6.1	Results	54
7.7	Community sentiment	58
7.8	Conclusion	58
8	Profiling diaspora membership	60
8.1	Data labeling	61
8.1.1	Dataset balance	61
8.1.2	Results	62
8.2	Feature overview	63
8.2.1	Direct country mentions	63
8.2.2	Flag Emoji	64
8.2.3	Country code	64
8.2.4	Cyrillic Characters	65
8.2.5	Naming conventions	65
8.2.6	User-specified location	66
8.2.7	Government and state-affiliated accounts	66
8.2.7.1	Twitter-annotated accounts	66
8.2.7.2	Non-annotated accounts	66
8.3	Feature selection	69
8.3.1	Profile features	69
8.3.2	Followed account features	69
8.3.3	Feature correlation	70
8.3.4	Results	72
8.4	Model selection	73
8.4.1	Performance metrics	73
8.4.2	Performance results	73
8.4.2.1	Random Forest	74
8.4.2.2	Gradient Boosted Decision Trees	75
8.4.2.3	Support Vector Machines	76
8.4.3	Classification errors	77
8.4.4	Cross-validated results	78

8.5	Model tuning	80
8.5.1	Performance metrics	80
8.5.1.1	Random Forest	80
8.5.1.2	Gradient Boosted Decision Trees	82
8.5.1.3	Support Vector Machines	83
8.5.2	Cross-validated results	86
8.6	Conclusion	87
9	Diaspora interaction	88
9.1	Distribution of predicted diaspora membership	89
9.1.1	Results for data set 1	89
9.1.1.1	Results per community	89
9.1.2	Results for data set 2	91
9.1.2.1	Data cleaning	92
9.2	Disinformation interaction	94
9.2.1	Results for data set 1	94
9.2.2	Results for data set 2	94
9.3	Twitter trolls	95
9.3.1	Results for data set 1	95
9.3.2	Results for data set 2	95
9.4	Conclusion	96
10	Conclusions	97
10.1	Primary research question	98
10.2	Disinformation campaigns	98
10.3	Social media communities	99
10.4	Diaspora membership	100
10.5	Diaspora interaction	100
10.6	Conclusion	101
11	Discussion	102
11.1	Key findings	103
11.2	Limitations	103
11.2.1	Data availability	103
11.2.2	Data collection	103
11.2.3	Mention networks	104
11.2.4	Diaspora classification	104
11.3	Recommendations	104
11.3.1	Limitations to historical data	105
11.3.2	Mention networks and disinformation	105
11.3.3	User profiling and classification	105
11.3.4	The role of diasporas during conflict	105

Appendices	114
A StopFake articles	114
B Unwound URLs	116
C Official accounts	118
D Technical artifacts and tools	125
D.1 Data collection	125
D.1.1 Scraping process	125
D.1.2 Implementation	127
D.2 Data storage	127
D.3 Data analysis and Machine Learning	128
D.4 Other tools	129

List of Figures

1.1	Information classification (Wardle and Derakhshan, 2017)	4
4.1	Classification of detection methods (Zhou and Zafarani, 2020)	14
5.1	DSRM Process Model	22
5.2	Classification of machine learning techniques	27
6.1	Fact-check selection process	33
6.2	Twitter share	35
6.3	Twitter collection process	36
7.1	Simplified mention network	40
7.2	Full mention network for Data set 1	42
7.3	Full mention network for Data set 2	43
7.4	Reduced mention network for Data set 1	45
7.5	Reduced mention network for Data set 2	46
7.6	Communities found using Leiden community detection	48
8.1	Tagged data distribution	62
8.2	Non-annotated and annotated accounts	67
8.3	Pearson coefficient correlation matrix for all features	71
8.4	Confusion matrix for RF	75
8.5	Confusion matrix for XGB	76
8.6	Confusion matrix for SVM	77
8.7	Cross-validated model Accuracy / μF_1 -scores	79
8.8	Confusion matrix for optimized RF	81
8.9	Confusion matrix for optimized XGB	83
8.10	Confusion matrix for optimized SVM	84
8.11	Cross-validated optimized model Accuracy / μF_1 -scores	86
9.1	Predicted data distribution for Data set 1	89
9.2	Predicted data distribution per community	90
9.3	Predicted data distribution for Data set 2	92
9.4	Predicted data distribution for Data set 2 after data cleaning	93

D.1 Scraping with asynchronous request (AJAX) support	126
D.2 Entity relationship diagram	128



List of Tables

6.1	State-controlled media	32
7.1	Top 6 communities found using Leiden community detection	47
7.2	Explicitly excluded tokens	50
7.3	Cluster 0 top 20 bigrams	51
7.4	Cluster 1 top 20 bigrams	51
7.5	Cluster 2 top 20 bigrams	52
7.6	Cluster 3 top 20 bigrams	52
7.7	Cluster 4 top 20 bigrams	53
7.8	Cluster 5 top 20 bigrams	53
7.9	Cluster 0 top 20 URL domains	54
7.10	Cluster 1 top 20 URL domains	54
7.11	Cluster 2 top 20 URL domains	55
7.12	Cluster 3 top 20 URL domains	55
7.13	Cluster 4 top 20 URL domains	56
7.14	Cluster 5 top 20 URL domains	56
7.15	Data set 2 top 20 URL domains	57
7.16	Community sentiment	58
8.1	Country mention features	63
8.2	Country flag emoji features	64
8.3	Country code features	64
8.4	Cyrillic character features	65
8.5	Naming conventions	66
8.6	User specified location	66
8.7	Official accounts	67
8.8	State affiliation labels	68
8.9	Strongly correlating features	72
8.10	Supervised learning algorithm performance	73
8.11	Classification report for RF	74
8.12	Classification report for XGB	76
8.13	Classification report for SVM	77

8.14	Classification report for optimized RF	81
8.15	Parameters used to optimized RF	82
8.16	Classification report for optimized XGB	82
8.17	Parameters used to optimized XGB	83
8.18	Classification report for optimized SVM	84
8.19	Parameters used to optimized SVM	85
A.1	StopFake articles	115
B.1	Source URLs	117
C.1	Official accounts	124

Chapter 1

Introduction



1.1 Motivation

In the digital age, huge amounts of information have become directly available to individuals. Such information streams provide unprecedented opportunities to directly influence these individuals. Social media plays a large role in providing these information streams both in the form of factual information as well as rumors. The large gray area between these types of information streams makes social media a prime candidate for targeted disinformation campaigns, and therefore an area whose properties and uses should be better understood. Researching disinformation campaigns on social media allows for an interesting multidisciplinary approach, by using methods and techniques from software engineering and machine learning (ML) to answer questions related to international relations and politics.

The Russo-Ukrainian conflict represents a recent (and ongoing) armed conflict, that not only saw kinetic, but also strategic information operations being conducted. These strategic information operations include the usage of targeted disinformation campaigns on social media (Khaldarova and Pantti, 2016; Richey, 2018). The overall time frame of the Russo-Ukrainian conflict stretches from early February 2014 up to the current-day frozen conflict. However, to study targeted campaigns the initial phase of the of the conflict was picked, as this period saw a number of major events (Fischer, 2019). Starting with the annexation of the Crimean peninsula (February 2014) and followed by the war in Donbas up until the end of 2015 (Chapter 2).

In addition, both Russia and Ukraine have large diasporas with more than 10 million Russians (Chindea, 2008a; Division, 2020) and more than 5 million Ukrainians (Chindea, 2008b; Ministry of Foreign Affairs of Ukraine, 2019; Division, 2020) respectively. This raises the question whether these groups play any role in the aforementioned disinformation campaigns, either knowingly or unknowingly.

In data science, extensive research has been conducted in the fields of the detection and spread of disinformation campaigns using Machine Learning techniques (Section 4.1). However, less is known about the role of diaspora as a proponent of such disinformation.

To the best of our knowledge, no research has been performed that combines the existing knowledge on the spread of disinformation with the interaction of the diasporas of countries in an ongoing conflict with such disinformation. Combined with a solid foundation of related research, this provides an opportunity to uncover new knowledge.

1.2 Goal

The primary goal of the research is to investigate whether a distinct **role of diaspora in the propagation of disinformation originating from their country of origin on English-language social media** can be identified while using the current Russo-Ukrainian conflict as a frame of reference.

This research achieves this goal by combining existing methods for social media mining, Natural

Language Processing (NLP), and social media profiling (Chapter 4). These methods are used to find social media messages (i.e. Twitter) interacting with known disinformation campaigns (in order to answer RQ1), patterns in these interactions (in order to answer RQ2), the identification of diaspora members (in order to answer RQ3), and whether diaspora members show any distinct interaction patterns with these campaigns (in order to answer RQ4).

1.3 Definitions

This proposal is primarily aimed at computer science researchers but touches on a number of other fields or specializations. For that reason, the most important terms used throughout this document are explicitly defined below.

1.3.1 Diaspora

The term Diaspora originally referred specifically to Jewish communities living outside of Israel. However, it currently holds a broader meaning as *the dispersion or spread of any people from their original homeland*¹. Considering the diverse ethnic make-up of the countries of the former Soviet Union and Ukraine specifically (Section 2.1), this definition can be interpreted in many ways. On top of that, a number of technical limitations require the term to be more explicitly defined within the scope of this research. Section 5.3.3 describes these limitations and the interpretation used for this research.

1.3.2 Types of information

Before going into detail it is important to define the exact framework in which different types of harmful and/or incorrect information can be classified. In the context of this work, the framework presented by Wardle and Derakhshan (2017) is used. The framework is widely used (Shu, Mahudeswaran, et al., 2018) which allows for a structured comparison. Wardle and Derakhshan (2017) identify three categories of information disorder: **misinformation**; i.e. false information created without malign intent, **disinformation**; i.e. false information created with malign intent, and **malinformation**; i.e. genuine information used with malign intent as shown in Figure 1.1. Here, malign intent is defined as aiming to inflict harm on a person, organization, or country. While refraining from the term *fake news* and specifically focusing on disinformation in this proposal, the term is used extensively in existing research, often referring to what Wardle and Derakhshan (2017) call *manipulated* or *fabricated* content. Both of these being classified as disinformation.

¹<https://www.lexico.com/definition/diaspora>

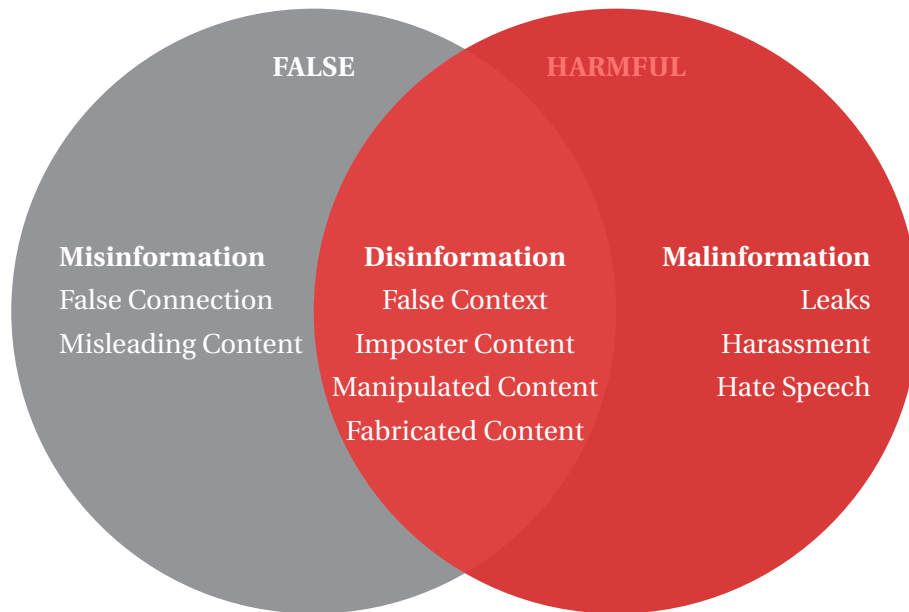


Figure 1.1: Information classification (Wardle and Derakhshan, 2017)

1.3.3 Disinformation campaigns

Building on the definition of disinformation by Wardle and Derakhshan (2017) as false information created with the aim to inflict harm on a person, organization, or country, we can further define what constitutes disinformation campaigns.

The coordinated spread of such false information can be seen as part of the broader area of *strategic information operations*, defined by Starbird, Arif, and Wilson (2019) as "the efforts by individuals and groups, including state and non-state actions, to manipulate public opinion and change how people perceive events in the world by intentionally altering the information environment". Starbird, Arif, and Wilson (2019) then define *disinformation campaigns* as means to conduct the aforementioned efforts using false information.

Within the context of this research, we narrow this definition down to state-actions, and more specifically to actions directly related to the Russo-Ukrainian conflict by state-actors directly or indirectly involved in this conflict. In addition, we specifically focus on online strategic information operations.

This brings us to the following working definition for *disinformation campaigns*:

The efforts by individuals or groups, acting on behalf of a nation-state, that use false information to achieve a strategic goal against another nation-state within a larger conflict by manipulating the public opinion using social media.

1.4 Document outline

This research is structured as follows: Chapters 2 and 3 provide the context in which this research will be conducted. Here, Chapter 2 gives a broad overview of the lead-up to the ongoing Russo-Ukrainian conflict while Chapter 3 gives a short introduction into information warfare and disinformation campaigns. Chapter 4 explores existing academic research related to disinformation campaigns from a computer science perspective. Chapter 5 presents the research questions based on the aforementioned research goal. In addition, it details the research method, validation methods, limitations, and risks. Chapters 6 to 9 explore and answer each of the four research questions, Chapter 10 offers the overall results and conclusions. Finally, Chapter 11 places the key findings into the context of existing research, raises possible limitations, and recommends future research directions.

Chapter 2

The Russo-Ukrainian conflict

This research will use the currently ongoing conflict between Russia and Ukraine as a frame of reference, particularly focusing on the period from February 2014 until the end of 2015. To get a better understanding of this conflict, it is important to know the lead-up and current state of affairs. Ukraine is a country in eastern Europe that encompasses an area with a rich history. The country has had its current form since becoming an independent nation in 1991 as a result of the collapse of the Soviet Union.

This chapter will provide some background to the Russo-Ukrainian conflict by giving an overview of the ethnic composition of Ukraine (in which the conflict started and is taking place), its geopolitical location, and the events that lead up to the eventual armed conflict.

2.1 The ethnic composition of Ukraine

The demographics of Ukraine, more specifically the ethnic and linguistic makeup, play a noteworthy role in the Russo-Ukrainian conflict. Based on the most recent census it becomes clear that Ukraine is home to a number of ethnic groups, with Ukrainians (77.8%) and Russians (17.3%) being the largest ones¹. A very similar distribution can be seen when looking at the mother tongue, which shows Ukrainian (67.5%) being followed by Russian (29.6%) with Russian being the most prevalent in the south and east of the country². According to the 2001 census, four regions have a majority of people with Russian as a native language: The Autonomous Republic of Crimea, The City of Sevastopol (a city with special status on the Crimean peninsula), Donetsk Oblast, and Luhansk Oblast³.

2.2 Ukraine as a *borderland*

Besides this ethnic division running through Ukraine, it also finds itself right on the edge of the spheres of influence of different political and military alliances. More specifically: Hungary, Poland, Romania, and Slovakia (all western neighbors except Moldova) are all members of both the European Union (EU) and the North Atlantic Treaty Organisation (NATO). At the same time Belarus, Moldova, and Russia (the remaining neighbors) are all members of the Commonwealth of Independent States (CIS), an organization promoting cooperation between post-Soviet republics.

2.3 A country in conflict

For a number of years, the country had been working on strengthening its ties to the European Union, of which a Ukraine-EU association agreement was an important step.

However, in a sudden change of direction, in November 2013 the Ukrainian government made the decision to reject the Ukraine-EU association agreement in favor of closer cooperation with Russia. The sudden rejection caused the opposition to call for protest, which eventually resulted in what is now known as the Euromaidan protests. Over time these protests escalated and led to violent and deadly clashes between protesters and the police. In February 2014, this turn to violence resulted in then-president Viktor Yanukovich signing a settlement agreement, promising constitutional reform and new elections. Not long after, Yanukovich fled the country and was formally removed from power by an (at the time unconstitutional) impeachment vote.

¹National composition of population

<http://2001.ukrcensus.gov.ua/eng/results/general/nationality/>

²Linguistic composition of the population

<http://2001.ukrcensus.gov.ua/eng/results/general/language/>

³Distribution of the population of Ukraine's regions by native language

[http://database.ukrcensus.gov.ua/MULT/Dialog/varval.asp?ma=19A050501_02&ti=19A050501_02.%20Distribution%20of%20the%20population%20of%20Ukraine`s%20regions%20by%20native%20language%20\(0,1\)&path=../Database/Census/05/02/01/&lang=2&multilang=en](http://database.ukrcensus.gov.ua/MULT/Dialog/varval.asp?ma=19A050501_02&ti=19A050501_02.%20Distribution%20of%20the%20population%20of%20Ukraine`s%20regions%20by%20native%20language%20(0,1)&path=../Database/Census/05/02/01/&lang=2&multilang=en)

Both the south and east of Ukraine quickly saw a counter-protest movement preferring closer cooperation with Russia, fearing to lose their unique status as a large ethnic and linguistic group within Ukraine.

2.4 The Russian annexation of Crimea

On top of the high number of ethnic Russians, the Crimean peninsula stands out from the other regions because of its historic status. The Crimean Peninsula was formerly part of both the Russian Empire and the Russian Soviet Federative Socialist Republic (RSFSR) but was transferred to the Ukrainian Soviet Socialist Republic (UkSSR) in 1954 (under dubious legal circumstances). This transfer eventually led to it becoming part of modern-day Ukraine. Furthermore, the city of Sevastopol has traditionally been the home of the Russian Black Sea Fleet, after the collapse of the Soviet Union, Russian forces remained stationed there under a lease agreement with Ukraine (Grant, 2015). Ukraine making a turn to closer cooperation with the west could therefore mean Russia would lose this strategic asset (Giles, 2015).

On the 27th of February 2014, armed men wearing unmarked military uniforms gained control over the Crimean peninsula by seizing multiple strategic locations like government buildings and broadcasting facilities. While Russia initially denied any involvement, it later confirmed that these men indeed were Russian military⁴ That same day an emergency closed-door vote was held in the Crimean parliament, which elected a new pro-Russian prime minister. Early March, the newly installed leadership of the Crimean Peninsula formally requested Russian military intervention. Later that month a referendum took place in which the citizens of both the Crimean Autonomous Region and the City of Sevastopol voted to be annexed by Russia with more than 95% being in favor. The results of this referendum have been contested (Grant, 2015).

2.5 The war in Donbas

Just like the Euromaidan protests, the pro-Russian protests in the south and east of Ukraine also became increasingly violent. After the annexation of Crimea, the protests in the Donbas region, primarily consisting of the aforementioned Donetsk and Luhansk oblasts, quickly escalated when government buildings were seized by pro-Russian insurgents and two independent republics were declared, the Donetsk People's Republic (DNR) and Luhansk People's Republic (LNR). Active fighting between the Ukrainian army and these insurgents during 2014 and 2015 eventually led to a stalemate, which remains as of today. Besides armed conflict, the war in Donbas also saw widespread usage of disinformation campaigns (Khaldarova and Pantti, 2016; Richey, 2018), on which this research will focus.

⁴Direct Line with Vladimir Putin
<http://en.kremlin.ru/events/president/news/20796>

Chapter 3

Disinformation campaigns

This chapter will provide a short history of disinformation campaigns in the (former) Soviet Union and how they found their place in the online discourse related to the Russo-Ukrainian conflict.

The targeted and coordinated release of disinformation to discredit an opposing party is not a recent phenomenon. Examples can be found throughout history and evidence of them date back as early as 44 BCE with Octavian's smear campaign against Mark Anthony (Sifuentes, 2019). The campaigns as seen during the Russo-Ukrainian conflict, however, can be lead back to more recent origins.

Because the Russo-Ukrainian conflict takes place right at the edge of both western and Russian spheres of influence it is closely monitored and widely researched from political, historical, and military perspectives. This information, however, is mostly limited to the western perspective.

3.1 A western perspective

From this western perspective, the Russian approach is often referred to as hybrid warfare, using a combination of kinetic (traditional warfare) and non-kinetic actions. Non-kinetic actions encompass alternative forms of warfare including (but not limited to) psychological, electronic, and information warfare (Richey, 2018). Another term commonly used is the *Gerasimov Doctrine*. This term, however, is based on a single opinion piece on an article in a Russian military magazine and has been since been refuted by the author of this opinion piece (Galeotti, 2014; Galeotti, 2019). Giles (2015), argues that hybrid warfare is mostly a popular label applied by western powers, that such a concept does not exist in Russia, and that besides new tools becoming available (e.g. social media) their approach has mostly remained much the same as it was in 1940.

Another term that is used often from the western perspective is cyber warfare, a war paradigm that includes influencing opinion and behavior and disrupting information systems (e.g. by attacking network infrastructure or performing distributed denial-of-service attacks). Russia has been seen as the source of many such attacks in the past years¹. Its actions surrounding the Ukrainian conflict, however, mostly focus on controlling the narrative using strategic information operations. Giles (2015) notes that this is mostly due to Russia already having dominance over many aspects of Ukrainian cyberspace due to its shared Soviet roots.

3.2 Traditional media

Traditional forms of mass media, like newspapers and television, have long been a way to distribute disinformation. The media in the Soviet Union was not free nor pluralistic and to some extent, this is still true for both Russia and Ukraine today. The media in Ukraine can be seen as independent but owned by a small group of influential oligarchs who regularly self-censor, whereas the Russian media can be considered a direct extension of the state (Roman, Wanta, and Buniak, 2017). With the majority of both Russians and Ukrainians still getting their news from traditional forms of media (Roman, Wanta, and Buniak, 2017), this allows for a very direct way of controlling the narrative. In contrast, Roman, Wanta, and Buniak (2017) states that the western media can be considered free and pluralistic. It has to be noted though, that this is not always the case and a decline is seen across western countries (United Nations Educational, Scientific and Cultural Organization, 2018).

3.3 Embracing social media

While the transition towards social media in these countries is still in progress, it has not been overlooked. One thing that does stand out is how the Russian government has embraced the shift from traditional mass media to social media using regular citizens as a means of spreading disinformation, propaganda and countering its role as a platform for dissent (Mejias and Vokuev,

¹Significant Cyber Incidents

<https://www.csis.org/programs/strategic-technologies-program/significant-cyber-incidents>

2017).

Besides making use of regular citizens, there are also more organized efforts with groups of people creating fake persona's online to place incendiary comments, write blog articles questioning trusted sources and provide an overall catalyst for the spread of disinformation. These groups of people are often referred to as internet trolls, or collectively as a troll factory or troll army².

²The Russian troll factory at the heart of the meddling allegations
<https://www.theguardian.com/world/2015/apr/02/putin-kremlin-inside-russian-troll-house>

Chapter 4

Related work

The chapter provides an overview of existing research in the field of disinformation detection and spread, existing disinformation data sets, and social media user profiling.



4.1 Disinformation detection and spread

Research into the detection of disinformation does not only help to slow down or stop its spread. Instead, it leads to a better understanding of how its linguistic properties and propagation differ from genuine information. Zhou and Zafarani (2020) categorize detection methods in four categories: knowledge-based detection, style-based detection, propagation-based detection and source-based detection. Figure 4.1 shows the full classification and sub-classifications. Of these, expert-based manual fact-checking is by far the least scalable, since it relies on experts manually verifying a specific news story or claim. This method is essential though as all the other categories rely on some form of manual fact-checking for verification and training purposes.

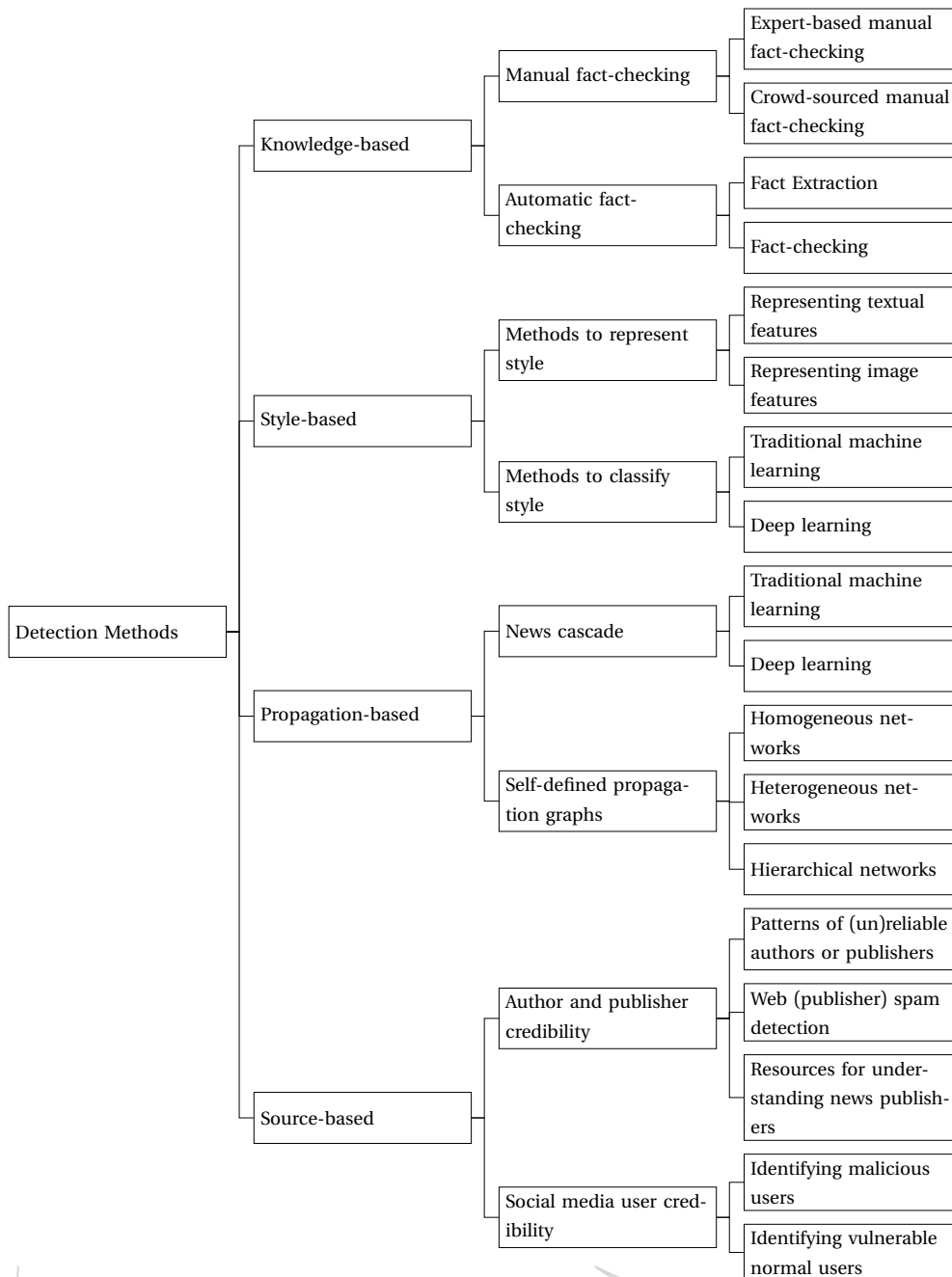


Figure 4.1: Classification of detection methods (Zhou and Zafarani, 2020)

Knowledge-based, style-based and source-based detection methods rely on the original source to detect patterns and properties that distinguish disinformation from genuine information and are therefore less interesting to study user interaction. Propagation- and source-based methods, however, specifically focus on how social media users spread and interact with such disinformation.

4.1.1 Propagation-based detection methods

Propagation-based detection methods use tree-like structures called *news cascades* to visualize the spread of a specific story on social media, with each node being a user that interacts with the story (Zhou and Zafarani, 2020), this not only allows for studying the patterns in the spread itself but also each user involved and their effect on it. Such interaction can be captured in different types of self-defined propagation graphs. Examples of this are *Spreader Net* which attributes users by their susceptibility to disinformation and social media influence (Zhou and Zafarani, 2019) and *Stance Net* which attributes users by their opposing or supporting stance towards a specific piece of information (Jin et al., 2016).

4.1.2 Style-based detection methods

Style-based detection methods try to extract patterns from the content of articles or messages. This is done by performing linguistic analysis of the textual content of these articles. Zhou and Zafarani (2020) reiterate four levels on which such texts can be analyzed: Lexicon-level looks at the text as a collection of words and is often used for extracting features of a text by studying word occurrence. Syntax-level analysis is used to identify the grammatical structure of a text, for example by tagging parts of speech and forming parse trees. Discourse-level analysis captures the rhetorical relations within a text, like elaboration or attribution. Semantic-level analysis categorizes words used in a text using a predefined set of categories, the occurrence of which can be used to detect patterns.

4.1.3 Source-based detection methods

In addition to propagation-based detection methods, source-based detection methods also focus on the role of social media users. Zhou and Zafarani (2020) divide this category into two distinct subcategories: author and publisher credibility, and social media user credibility. Contrary to the categorization used, the research into author and publisher credibility also provides opportunities for identifying the role of specific users (or groups of users). For example, Horne, Nørregaard, and Adali (2019) researched the content-sharing relationship between different news websites by comparing the similarity of articles and clustering them using a community detection algorithm (Leicht and Newman, 2008). The results gathered indicate that there are a number of tightly formed communities. Similar research can be used to find communities amongst the behavior of social media users. Helmus (2018) conducted such a study specifically focused on the Russo-Ukrainian conflict. A similar community detection algorithm (Clauset, Newman, and Moore, 2004) was used to find communities within Russian-language tweets origination from post-Soviet countries. In this case, two overarching meta-communities were found, a general discussion meta-community and a highly politicized meta-community discussing the Russo-Ukrainian conflict in which they could clearly identify the pro-Russian and pro-Ukrainian sides.

4.1.4 Conclusion

It becomes clear that there is very active research into the detection of disinformation which can be categorized into four main categories: knowledge-based detection, style-based detection, propagation-based detection, and source-based detection. These methods are not just suitable for detecting disinformation, but also for studying its propagation and how social media users interact with it.

4.2 Disinformation data sets

Existing research into disinformation detection and spread often makes use of commonly used data sets. This way, a baseline for comparing the results of different methods is provided. A selection of these commonly used data sets is listed below. Each of which consist of English language Twitter messages or news articles published in the United States, usually annotated by expert fact-checkers.

CREDBANK This data set¹ contains 60 million tweets created in 2015 and grouped by event. The dataset was annotated using crowd-sourced manual fact-checking by 30 Amazon Mechanical Turk participants (Mitra and Gilbert, 2015). Since the data is already grouped by topic, it can quickly be checked for any statements relating to the Russo-Ukrainian conflict that have already been verified.

BuzzFeedNews This data set² provides 1627 news articles published on Facebook over the time of one week leading up to the 2016 United States presidential election. For each article, individual claims were verified using expert-based manual fact-checking by a group of five BuzzFeed journalists (Silverman et al., 2016). Considering that this data set was gathered around the 2016 United States presidential election it will most probably not provide any useful data regarding the Russo-Ukrainian conflict.

BuzzFace This data set³ takes the dataset provided by BuzzFeedNews (Silverman et al., 2016) and enhances it by adding metadata detailing user interaction with each of the articles. This metadata was gathered directly from Facebook as well as comment sections on the website of the source news outlets (Santia and Williams, 2018).

LIAR This data set⁴ is based on statements gathered from and annotated by PolitiFact, an expert-based manual fact-checking website (W. Y. Wang, 2017).

BS Detector The BS detector data set is based on data collected using crowd-sourced manual fact-checking by means of a browser extension. While listed by Shu, Mahudeswaran, et al. (2018), both the data set and browser extension are not available anymore. Copies of the original data can still be found⁵.

¹<http://compsocial.github.io/CREDBANK-data/>

²<https://github.com/BuzzFeedNews/2016-10-facebook-factcheck>

³<https://github.com/gsantia/BuzzFace>

⁴https://sites.cs.ucsb.edu/~william/data/liar_dataset.zip

⁵<https://github.com/thiagovas/bs-detector-dataset>

FakeNewsNet This data set⁶ uses the expert-based manual fact-checking websites PolitiFact and GossipCop as a source of news stories and enhances this by adding the original news content and metadata on Twitter interaction (Shu, Mahudeswaran, et al., 2018). Because the current data set relies on fact-checking websites that focus specifically on the United States, it might not contain any useful data regarding the Russo-Ukrainian conflict.

Furthermore, a number of other sources can be used to gather data specific to the scope of this research:

EU vs Disinfo EU vs Disinfo is a fact-checking website with a strong focus on Russian disinformation⁷. It is set up as part of the European External Action Service East StratCom Task Force (EEAS ESCTF) which specifically focuses on promoting activities of the European Union in Eastern Europe⁸. The data set provided⁹ is not officially released in a machine-readable format, but this can easily be extracted using third-party tools¹⁰. Because of its specific focus on Eastern Europe and Russian disinformation, EU vs Disinfo can provide useful insights with regards to the Russo-Ukrainian conflict.

Khaldarova and Pantti (2016) While not directly providing the data set itself, Khaldarova and Pantti (2016) have collected and analyzed around 6000 tweets mentioning specific disinformation campaigns surrounding the Russo-Ukrainian conflict.

Twitter Stream Grab The internet archive is an organization that is building a digital library of Internet sites and other cultural artifacts in digital form¹¹. Besides providing archived versions of web pages, the internet archive also provides a historical archive of tweets retrieved through Twitter streaming APIs¹². These data sets can provide very valuable information but are limited to a sub-selection of 1% of all tweets. However, research by Kergl, Roedler, and Seeber (2014) supports that this selection is representative of the full data set.

Data Requests In some cases it is possible to request a specific data set for academic research directly from social media platforms, but this does require very specific and targeted requests.

In addition to providing the data sets, both Mitra and Gilbert (2015) and Shu, Mahudeswaran, et al. (2018) describe detailed preprocessing steps used to create these data set, information that can be used when existing data sets do not suffice in helping answer the research questions of this research.

⁶<https://github.com/KaiDMML/FakeNewsNet>

⁷<https://euvsdisinfo.eu/about/>

⁸<https://eeas.europa.eu/headquarters/headquarters-homepage/2116/-questions-and-answers-about-the-east-stratcom-en>

⁹<https://euvsdisinfo.eu/disinformation-cases/>

¹⁰<https://www.kaggle.com/corrieaar/disinformation-articles>

¹¹<https://archive.org/about/>

¹²<https://archive.org/details/twitterstream>

4.2.1 Conclusion

Large data sets focusing on disinformation are already available and have been widely used in existing research. However, it has to be noted that there is much less data available concerning disinformation relating to the Russo-Ukrainian conflict. This can mean that additional data gathering is necessary.

4.3 Social media profiling

Profiling social media users is a classification problem, this means that different attributes that a user presents are being used to classify that user into predetermined categories. User profiling sees a lot of attention in the area of disinformation detection (Section 4.1), with a specific mention as *social media user credibility* in the types of disinformation detection methods presented by Zhou and Zafarani (2020).

Pennacchiotti and Popescu (2011) provide a generic model to perform this classification task for Twitter users. This model is based on four classification categories:

The first category, *profile features*, depends on static data found on a user profile, like the name, profile description, and profile picture. Pennacchiotti and Popescu (2011) state that this information by itself does not provide a trustworthy way to classify users. This does, however, not mention the fact that Twitter users often provide their real name, which can be used as a very effective way of classifying ethnicity based on naming conventions (Wong et al., 2020; Bessudnov et al., 2021). Commercial services offering name-based ethnicity classification are already available¹³.

The second category, *tweeting behavior*, relies on temporal (e.g. how often they post or like a message) and numeric properties (e.g. how many followers they have) of the social media user. According to Pennacchiotti and Popescu (2011), these statistics are less effective at classifying users than any of the other methods.

The third category, *linguistic content*, relies on the linguistic properties of messages posted by the social media user, by analyzing these messages using natural language processing. These are the same methods as described in Section 4.1, with the addition that many social media platforms implement the concept of hashtags, which allow for cross-referencing messages discussing the same topic.

The fourth category, *social network*, focuses on the other users that a user interacts with. Examples of such interactions are follower-followee relationships and message-based interactions (e.g. likes, retweets, replies).

Using this framework, Pennacchiotti and Popescu (2011) present a system that extracts a number of attributes for each of these four categories, plugging them into Gradient Boosted Decision Trees (GBDT), and apply this system to an experimental setup. The performance of the proposed system is measured by comparing the precision, recall (sensitivity), and F-score against two baseline

¹³<https://www.namsor.com/>

results. One being all accounts that have explicitly mentioned political affiliation, the other being based on just *profile features* and *tweeting behavior*. The results show that adding *linguistic content* and *social network* features improve both the precision and recall compared to the baseline scores, meaning that using a combination of feature categories will lead to better classification results.

Zamal, W. Liu, and Ruths (2012) show that user classification can be improved by including attributes inferred from on their neighborhood, i.e. the classification of their friends and followers. By combining these new neighborhood-based features with the features from the categories above, Zamal, W. Liu, and Ruths (2012) show that this approach provides improved accuracy for detecting political affiliation. In this case Support Vector Machines were used as they proved to be more performant than Gradient Boosted Decision Trees.

Shu, S. Wang, and H. Liu (2018) take a similar approach to distinguish between *experienced* and *naïve* Twitter users, based on their trust towards false information. This is achieved by making use of both explicit and implicit features. The explicit features consist of *profile features* (verified account, account age) and *tweeting behaviour* (number of posts, number of favorites, follower count, followee count) . The implicit features, user age and personality are derived from the *linguistic content* of the top messages for each user.

One classification problem that more closely aligns with someones stance towards specific sides in a conflict is the prediction American political affiliation (Democrat or Republican) based on Twitter data (Conover et al., 2011; Pennacchiotti and Popescu, 2011; Zamal, W. Liu, and Ruths, 2012). This problem is not only of specific interest because it is similar to classifying diaspora membership but also because it allows for comparing the performance of different methods used in existing research. Here Conover et al. (2011) show that just relying on hashtags instead of a full suite of features can provide similarly effective predictions for political affiliation. However, this bases the classification solely on message content.

Similarly Aparup Khatua, Apalak Khatua, and Cambria (2020) use *linguistic content* to classify a users' political preference in a multi-party system relying the occurrences of a predefined set of keywords.

The topic of social media user profiling also sees activity in the area of text forensics as *author profiling*, which traditionally performed such tasks on longer texts and articles, but has since also touched upon social media. A wide array of approaches is found in the submissions for the PAN¹⁴ events (Rangel and Rosso, 2019; Rangel, Giachanou, et al., 2020; Bevendorff et al., 2021). These approaches mostly rely on *linguistic content*.

Another area of interest is profiling based on multiple social networks, where multiple sources are used to build a profile and fill out any missing information, one such approach is presented by Song et al. (2015) who combine profile information from multiple social networks to predict the user's tendency to volunteer.

¹⁴<https://pan.webis.de/>

4.3.1 Conclusion

Advanced and generic social media profiling methods have been widely researched and can be adopted for this research. The case of predicting American political affiliation is of particular interest, because the problem is similar to detecting diaspora membership.

Linguistic content is widely used for effective predictions for bot user detection, age, gender, political affiliation and topic modeling. For this reason it is suitable for analyzing the sentiment within social media communities (RQ2). However, it is less suitable for diaspora classification because it might skew the results based on how users interact with disinformation (e.g. classifying a user as belonging to the Ukrainian diaspora *because* they refute information from a Russian source). Which in turn can affect the results to be used in RQ4. Instead, *profile features* and *tweeting behavior* can be used to address RQ3, strengthened by the use of latent attributes derived from followed accounts as shown by Zamal, W. Liu, and Ruths (2012).

Chapter 5

Research

In this chapter, the overall scope, methodology, and research questions that this research aims to answer are defined.



5.1 Research methodology

The research will be conducted in a number of stages based on the design science research methodology (DSRM) as defined by Peffers et al. (2007). The DSRM model uses an integrative process that is well suited for research that heavily depends on artifacts. Figure 5.1 shows the DSRM model, consisting of a number of steps and possible entry points.

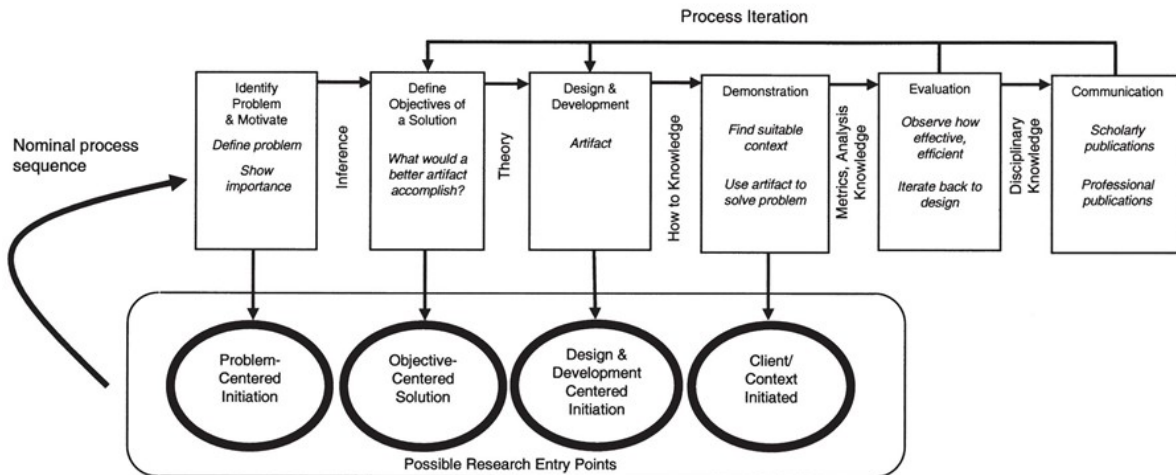


Figure 5.1: DSRM Process Model

Within the DSRM model, the previous chapters serve as the **initial problem definition and motivation**. The research itself will iterate over each following step, from **defining the objectives of a solution to evaluating results**. The results of this iterative process will then be **communicated** by the thesis as a whole. These iterations will happen at least once for each individual research question, but might be repeated multiple times to narrow down the results required for answering a single research question. The following section describes each research question and its DSRM stages.

5.2 Research questions

Based on the research goal presented in Section 1.2 the following primary research question is formed:

5.2.1 RQ Primary research question

RQ. What is the interaction between users, considered to be diaspora members, and known disinformation campaigns on social media related to the Russo-Ukrainian conflict? This question is further split up into the following sub-questions:

5.2.2 RQ1 Disinformation campaigns

RQ1. Which known disinformation campaign surrounding the Russo-Ukrainian conflict is suitable to be analyzed for Diaspora interaction? To properly analyze the role of the Diaspora, a number of well spread and well-known disinformation campaigns are needed, preferably claims or news articles that are already present in existing data sets (Section 4.2).

Solution objectives The primary objective is to find a data set interacting with a specific disinformation campaign. To do this a number of known campaigns have to be analyzed and specific claims, keywords, or links regarding this campaign have to be extracted. Using these attributes both Twitter itself and existing data sets can be searched for matches.

Design and development For this, we write code to analyze and search existing data sets. In addition, tools have to be created or combined to gather new data sets directly from Twitter, either using the Twitter API or using scraping. This data will then be stored in a database for further analysis. Appendix D.1 describes this approach in more detail.

Demonstration Using these tools a new data set can be created containing Twitter messages interacting with a specific disinformation campaign.

Evaluation Based on the size and precision of this data set, the choice is made to continue or adjust the approach. Existing data sets used in related research range from 12 thousand (W. Y. Wang, 2017) to 60 million (Mitra and Gilbert, 2015) records, the aim is to at least reach the lower bound.

Continuing from RQ1, two parallel approaches are taken to identify communities within the gathered data. The first approach (RQ2) focuses on communities that can be extracted directly from the set. The second approach (RQ3) focuses on communities based on diaspora membership. RQ4 combines the result of both approaches to see whether any relationships can be found.

5.2.3 RQ2 Social media communities

RQ2. Which Twitter user communities exist interacting with the selected disinformation campaign surrounding the Russo-Ukrainian conflict? Using a selection of known disinformation campaigns gathered as part of RQ1, the research aims to adapt the existing approach used by (Helmus, 2018). This approach uses a *mention network* to identify interactions between users and then applies the Clauset-Newman-Moore community detection algorithm (Clauset, Newman, and Moore, 2004) to find any communities of interest. Lexical analysis is then performed on the message content of any communities of interest to determine their stance towards the Russo-Ukrainian conflict. Such methods are further studied in Sections 4.1 and 4.3.

Solution objectives The primary objective is to use the data set gathered as part of RQ1 to find and visualize communities for different stances towards the selected disinformation campaign(s).

Design and development Code has to be written to take the data gathered earlier and apply different community detection algorithms on the relationship between users, this should be relatively straightforward when representing the initial data set as a graph as defined in Appendix D.1. The textual content of messages belonging to these communities can then be plugged into lexical analysis methods.

Demonstration The results of applying these algorithms will be a list of communities identified as well as a visualization showing each community. Furthermore, the result of lexical analysis on the messages posted within each community will show their overall stance towards the Russo-Ukrainian conflict.

Evaluation The successful completion of this stage depends on the types and number of communities detected. In case of a large number of distinct communities, these steps can be repeated using different parameters and or algorithms. It further depends on how well lexical analysis can identify the overall stance of each community.

5.2.4 RQ3 Diaspora membership

RQ3. Which machine learning methods are suitable for predicting diaspora membership Twitter user profiles? Building on top of existing research into social media profiling using machine learning, multiple machine learning methods have to be tested to see how well diaspora membership can be predicted. Section 4.3 explores social media profiling and applicable machine learning methods in more detail.

Solution objectives The primary objective is to be able to correctly identify a diaspora member based on the Twitter message and profile data collected as part of RQ1 using supervised machine learning methods. The goal of this research question is not to find a novel way to perform social media profiling but to apply proven profiling methods to the specific task of predicting diaspora membership.

Design and development Detecting Ukrainian, Russian, or none/other diaspora membership is a *classification* problem. There are a number of machine learning algorithms commonly used for such problems. An important step of performing classification is features engineering, the process of determining which exact features will be used to determine the classification. An example of features based on basic profile information is the usage of demonyms or flag emoji in profile descriptions. However, there is more information available than just basic profile information. Such features and their application within classification problems are further explored in Section 4.3. Having a predefined list of features leads us to specifically focus on supervised Machine Learning algorithms. The large set of available algorithms available (e.g. Logistic Regression, Decision Trees, Random Forest, Gradient Boosted Decision Trees, and Support Vector Machines) makes it unfeasible to cover all. Furthermore, no universally best algorithm exists and each problem might offer its own best solution (Caruana and Niculescu-Mizil, 2006). However, the initial selection falls on Gradient Boosted Decision Trees which have been used for similar classification problems (Pennacchiotti and

Popescu, 2011). For this, we write code to extract the selected features and plug these into the chosen supervised machine learning method. In addition, part of the data has to be tagged by hand for training purposes.

Demonstration The resulting data set uses the data gathered as part of RQ1 enhanced with the diaspora classification of each user.

Evaluation Because the this research specifically focuses on the role of diaspora members, there has to be a very precise indication of diaspora membership. This means that it is necessary to find a set of features and an algorithm that has a high *precision* (the percentage of actual diaspora members among all matches) but does not specifically require a high *recall* (the percentage of matched diaspora members among all actual diaspora members). Or more simply put, it is better to be sure that matched profiles are diaspora members than to match all possible diaspora members. It is also important to note that Russian or Ukrainian diaspora members will probably only represent a small number of overall social media users, so an imbalanced data set is to be expected. Different results can be compared using these metrics or derived metrics like F-scores. To use the same data set for both training and testing a machine learning algorithm, a technique called *cross-validation* can be used, this technique splits the data into blocks, which uses one block at a time to test a model generated from the remaining blocks. These steps can be repeated using different features and or algorithms.

5.2.5 RQ4 Diaspora interaction

RQ4. **How does diaspora membership relate to detected Twitter user communities?** Using the data gathered as part of RQ2 and RQ3 we can study the relationship between the communities found and known diaspora members.

Solution objectives The primary objective is to find a way to identify any existing relationship between the communities found as part of RQ2 and the diaspora membership found as part of RQ3.

Design and development A visualization has to be created that combines the results of RQ2 and RQ3. The actual technical implementation of this research question greatly depends on the data resulting from previous questions.

Demonstration This visualization can be used to gain a high-level overview of possible relationships. Any relationships found can then be further studied. The exact method of identifying and defining any possible relationship is still unclear but will most likely be a statistical relationship between the community a user belongs to based on their stance (RQ2) and the community a user belongs to based on their diaspora membership (RQ3).

Evaluation The result of this research question will provide input for the final conclusion of the research and therefore lacks specific validation criteria. Simply speaking, the result might be that no significant relationship can be identified using the data and means used as part of this research.

5.3 Limitations

5.3.1 Bias in related works

All English-language research found as part of this proposal, while objective, still mostly looks at pro-Russian disinformation campaigns. To the best of our knowledge, no such research exists into pro-Ukrainian disinformation campaigns. While this does not affect the way in which this research itself is conducted, it might result in an implicit bias against the pro-Russian side.

5.3.2 Social media platforms

A number of factors were used to determine which social media platforms would be best suited for this research. Most platforms do not provide in-depth usage statistics, but third-party platforms¹ were used to determine the market share of different platforms. For each platform, the following requirements had to be met:

English-language The platform must have an English-speaking community, this eliminated a number of large Chinese-language social media platforms.

User identity To more easily identify diaspora membership, the platform must provide access to some sort of profile for each user including group memberships or followed accounts.

Text-based content Some platforms are primarily video or image-focused, analyzing disinformation using such formats is still an immature area of research, for this reason only text-based platforms were considered.

Existing research Platforms that have not been widely researched (Chapter 4) are not considered, as this prevents leveraging existing results.

Data availability Platform-provided APIs or third-party tools must exist to easily obtain information shared on that platform. In addition, data has to be available around the timeline selected as part of identifying disinformation campaigns (RQ1). Data availability has become an increasing challenge because multiple social media platforms are restricting direct access (Walker, Mercea, and Bastos, 2019).

Based on these requirements, Twitter² is the best-suited platform. While Facebook³ is currently the largest platform, its data policy severely limits access to required information.

5.3.3 Definition of diaspora

Within the context of this research, diaspora membership is defined based on social media profiles and interactions. This either constitutes explicit mentions (like a profile description) or social-media-based interactions and relations with organizations explicitly promoting the original homeland. This approach eliminates the ambiguity around the term but might lead to false positives

¹<https://gs.statcounter.com/social-media-stats>

²<https://www.twitter.com>

³<https://www.facebook.com>

and false negatives which have to be taken into account. Two expected examples can be found below:

False positive A person with a strong interest in the Ukrainian culture, but not part of the diaspora might follow different organizations promoting Ukraine.

False negative A person who emigrated from Ukraine, who interacts with information surrounding the conflict, but does not follow any organizations and does not explicitly associate with Ukraine in their profile description.

5.4 Machine learning techniques

To answer RQ3, different Machine Learning (ML) techniques will be employed. To gain a better understanding of the ML domain, the primary types of techniques and methods relevant to this research are described below.

5.4.1 Machine learning paradigms

Broadly speaking, three paradigms of Machine Learning techniques exist: supervised learning, unsupervised learning, and reinforcement learning, with a fourth category, semi-supervised learning, being made up of a combination of supervised and unsupervised learning. The most suitable type can be chosen based on the data available and the intended goal (Sarker, 2021). Figure 5.2 shows an overview of these different types and primary use-cases.

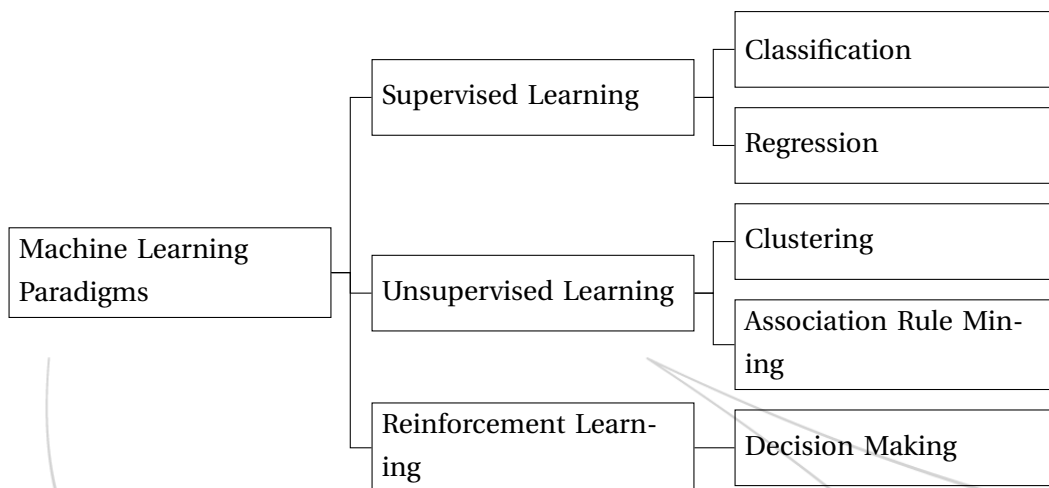


Figure 5.2: Classification of machine learning techniques

5.4.1.1 Supervised learning

Supervised learning relies on annotated (tagged or labeled) data, this means that not only the inputs but also the outputs have to be known for a subset of the data. This is referred to as the training set. Since having this data is a requirement to use Supervised Learning methods, it is often gathered from historical data (e.g. patient records where the diagnosis is already available) or created

by manually annotating a data set (data tagging).

5.4.1.2 Unsupervised learning

Unsupervised learning does not require any output data to be present ahead of time and can be used to find structure or patterns in data using just the input data. One example use-case is the clustering of data which aims to extract groupings from unstructured data.

5.4.1.3 Reinforcement learning

Reinforcement learning takes another approach, where agents make small adjustments to the environment. These adjustments are then continuously verified against the outcome, rewarding positive results and punishing negative results. This approach, therefore, requires a feedback loop and a definition of a positive or negative result.

5.4.2 Applicability

RQ3 focuses specifically on classifying users as diaspora members. Or more specifically, it focuses on categorizing users as being members of either the Russian or Ukrainian diaspora. Because not all users will fall under this category, a third non-diaspora user category is also required. In this case, the expected output is a discrete set of categories, and it is therefore considered a multi-class classification problem. Such classification problems are typically solved by using Supervised Learning algorithms and therefore the choice for the types of techniques considered for this research. In this approach, it is required that an existing annotated data set already exists or that one has to be created to serve as training data.

A plethora of classification algorithms are available within the supervised learning paradigm, the most commonly used algorithms for multi-class classification problems are listed below (Sarker, 2021; Aly, 2005).

(Multinomial) Naive Bayes (NB) Naive Bayes works by calculating the probability of each of the input values for each category in the training data. Furthermore, a prior probability is calculated based on the training data. This is based on the distribution of the different categories in the training data. To classify data, the input specific probability and the prior probability are used in conjunction to calculate the probability that data falls in each category. These values are then compared to select a category.

Logistic Regression (LR) Logistic Regression works by using a sigmoid function to map input values and predict a binary output value. In the case of multi-class classification, each class is treated as a binary one-versus-rest classification. The results of each class are then compared to make the prediction.

K-Nearest Neighbours (KNN) K-Nearest Neighbour classifies data by looking at its distance (using a distance measure) to existing training data. The closest existing data point is called a

neighbor. The number K identifies the number of neighbors to consider. If K is larger than one, the category that occurs the most among neighbors is chosen.

Support Vector Machines (SVM) Support Vector Machines work by finding a hyperplane that divides data between categories. When no clear hyperplane can be drawn, the data can be lifted to a higher dimension using a transform (kernel function) until a hyperplane can be drawn between values. For example: For single-dimensional data points this means using a function to make them 2-dimensional, therefore allowing a dividing line to be drawn between them.

Decision Tree (DT) Decision Trees classify data by creating a tree structure, where each of the nodes introduces a split based on the training data. The conditions for the split are inferred from the input data (features). The leaf nodes of the tree indicate the final classification of the input data. When non-training data is used, the tree is simply followed until a leaf node is reached. Decision trees algorithms have multiple parameters that affect the tree's properties, like the maximum depth or number of leaf nodes.

Random Forest (RF) Random Forests are an ensemble method. Ensemble methods are a category of machine learning techniques that combine multiple other techniques to create a model. Random Forests are made up of many Decision Trees based on random samples of the training data. Data is classified by checking it against each generated tree, the category with the most matches is then selected. This process is referred to as *bagging*.

Gradient Boosted Decision Trees (GBDT) Like Random Forests, Gradient Boosted Decision Trees are also an ensemble method. However, instead of creating multiple decision trees side-by-side, gradient boosting uses multiple trees sequentially, each being adjusted according to the classification errors in the previous tree, this process is referred to as *boosting*.

Further analysis and selection of these methods using tagged training data is described in Section 8.4 (RQ3).

Chapter 6

Identifying suitable disinformation campaigns

Based on the overall research method described in Chapter 5, this chapter will address the identification of disinformation campaigns and the raw data collection process (RQ1). This process is outlined by first defining an overall scope for the collected data. It then defines how disinformation campaigns are identified and how data was collected. Concluding the chapter we will present two data sets consisting of twitter data collected using the presented methodology.

6.1 Scope and context

6.1.1 Time frame

The overall time frame of the Russo-Ukrainian conflict stretches from early February 2014 up to the current-day frozen conflict. However, to study targeted campaigns initial phase of the conflict was picked as a number of key events happened during this period. Starting with the annexation of the Crimean peninsula (February 2014) and followed by the war in Donbas up until the end of 2015 (Chapter 2).

6.1.2 Data sources

Based on the different social media platforms analyzed (Section 5.3), Twitter presents itself as the most useful direct source to study the spread and interaction of disinformation campaigns on social media. Twitter, however, does not provide insights into the disinformation campaigns themselves. This means that further resources are required to find suitable campaigns. As part of this process, multiple data sets tied to existing research into online disinformation were checked but did not provide any data applicable to the Russo-Ukrainian conflict (Section 4.2). Because of this, the choice was made to split the data collection into multiple steps. First of it was necessary to find and identify suitable disinformation campaigns, subsequently, a way had to be found to identify interactions with those campaigns.

6.1.3 State-controlled media

Russia and Ukraine both rank on the lower end of the World Press Freedom Index, receiving *difficult situation*¹ and *problematic situation*² scores respectively. In both countries, the private media outlets are controlled by a small group of oligarchs. An important component of the definition of disinformation campaigns as presented earlier (Section 1.3) is the role of state-actors of states directly involved in the Russo-Ukrainian conflict. In this case, we specifically focus on state-controlled media, since they both provide a way to spread information and are directly being controlled by their respective states. Table 6.1 lists all media outlets regarded as state-controlled within the scope of this research.

¹<https://rsf.org/en/ukraine>

²<https://rsf.org/en/russia>

Country	Organisation	Type
Ukraine	Ukrinform	State enterprise ³
Ukraine	Suspilne (UA:Pershyi)	State enterprise ⁴
Russia	VGTRK (Russia 1, Vesti)	State enterprise ⁵
Russia	TASS	State enterprise ⁶
Russia	Rossiya Segodnya (RIA Novosti, Sputnik)	State enterprise ⁷
Russia	Channel One Russia	State majority shareholder ⁸
Russia	RT (Russia Today)	State funded non-profit ⁹

Table 6.1: State-controlled media

6.1.4 Fact-checkers

While there are multiple fact-checking websites focusing on US politics¹⁰¹¹, there are not many specifically focusing on the Russo-Ukrainian conflict. One fact-checker that does extensively focus on the Russo-Ukrainian conflict is StopFake¹². Because of its extensive coverage, StopFake was picked as an initial source for possible disinformation campaigns. It has to be noted that the StopFake project was started by a group of Ukrainian professors¹³, and has been part of recent controversies surrounding its ties to far-right organizations¹⁴. In the context of this research, StopFake is only used to provide an initial reference to the source articles of widely corroborated disinformation campaigns.

6.2 Data collection process

Using the scope defined above, it becomes possible to define a clear process for identifying and selecting suitable disinformation campaigns and in turn use those campaigns to collect data from Twitter.

6.2.1 Gathering disinformation sources

StopFake was used to collect fact-checked news stories within the given time frame. Only stories containing links to the original source of the story were considered. Furthermore, these links must

³<https://zakon.rada.gov.ua/laws/show/749-97-n#Text>

⁴<https://zakon.rada.gov.ua/laws/show/749-97-%D0%BF#Text>

⁵<https://vgtrk.com/>

⁶<https://tass.com/today>

⁷<https://rossiyasegodnya.com/mediagroup/>

⁸<https://www.interfax.ru/russia/742516>

⁹<https://www.interfax.ru/russia/678102>

¹⁰<https://www.politifact.com/>

¹¹<https://www.factcheck.org/>

¹²<https://www.stopfake.org/>

¹³<https://www.politico.eu/article/on-the-fake-news-frontline/>

¹⁴<https://medium.com/@zaborona.media/neo-nazi-links-of-a-facebook-fact-checker-exposed-d1215dae6c66>

refer to the websites of any of the state media outlets listed in Table 6.1. To process used to achieve this is described in Figure 6.1.

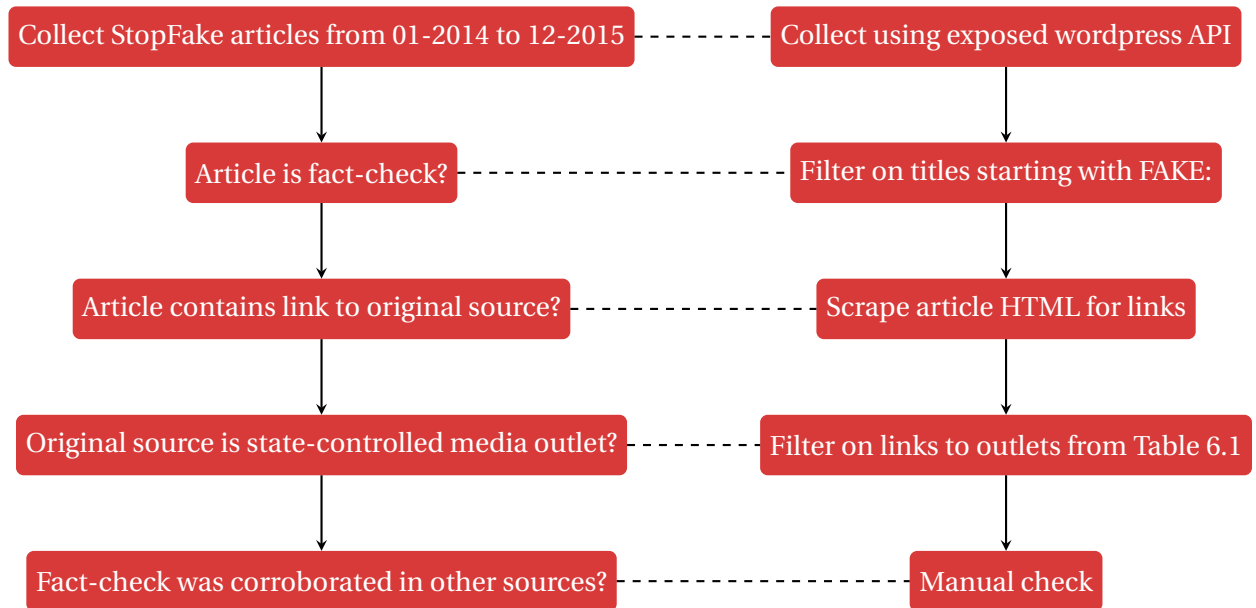


Figure 6.1: Fact-check selection process

In some cases, the articles contained a link to a YouTube copy of an original TV report. In this case, only the articles for which a link to the original news story could be found were included. This process resulted in a number of fact-checked stories originating from state media. These stories are listed in Appendix A.

6.2.2 Twitter interactions

Twitter supports a number of ways of creating and interacting with messages and other users. Each of them uses a slightly different mechanic:

Tweet A tweet is a basic message posted to the Twitter platform, all tweets used as part of this research are publicly available.

Retweet A retweet is a re-posting of a tweet. Users can re-post tweets by others as well as by themselves

Quoted tweet When retweeting a user re-posting the message can optionally add their own comment. In this case, the retweet is referred to as a quoted tweet.

Reply A reply is a tweet responding to another tweet. Replies to replies are grouped together in conversations.

Like A like applies to a single message and does not result in any new content being created.

Mention A mention is a direct reference to another user by including an @ followed by their user-name.

To gather data on interactions with disinformation campaigns it is essential to define what it means for a Twitter user to interact with them. Using the existing interaction mechanics defined above, we can define interaction as follows: an interaction is any tweet containing a direct link to disinformation in the form of a news article. Any retweets, replies, or likes of such tweets are seen as further interactions. This chain of interactions is then followed for each retweet or reply. For each tweet encountered additional metadata is stored. This includes, but is not limited to, geolocation, the language of the tweet, the number of retweets, and the number of likes. Furthermore, the user's profile information is also stored. In this case this concerns at least the user's full name, the user's profile description, the user's location, the number of followers, and the number of friends.

6.2.3 Source URLs and variants

The initial entry point for gathering interactions will therefore be a Twitter search for the URLs to the news stories mentioned in Appendix A. It is possible that multiple URLs point to the same article. To increase the number of possible matches, a number of steps are executed. This process is shown below using the following example URL:

```
https://www.1tv.ru/news/2014-07-12/37175-bezhenka_iz_slavyanska_vspominaet_k_
- ak_pri_ney_kaznili_malenkogo_syna_i_zhenu_opolchentsa?utm_source=share2
```

Since we are aiming to search Twitter for tweets mentioning the URLs, it is important to reduce it to its most generic variants. We achieve this by stripping the URLs of any non-essential parameters. These are any parts of the URL not required to display the right article, like additional tracking information. After stripping these parameters, our initial example will look as follows:

```
https://www.1tv.ru/news/2014-07-12/37175-bezhenka_iz_slavyanska_vspominaet_k_
- ak_pri_ney_kaznili_malenkogo_syna_i_zhenu_opolchentsa
```

6.2.4 URL-shorteners

The limited maximum length of Twitter messages¹⁵ stimulates the usage of URL shorteners. URL-shorteners are services that generate a new, shorter, URL that will redirect to the original URL. At the same time, this also results in masking the original URL. Many services and websites provide their own URL-shortening service, especially when using Twitter share functionalities.

¹⁵<https://developer.twitter.com/en/docs/counting-characters>

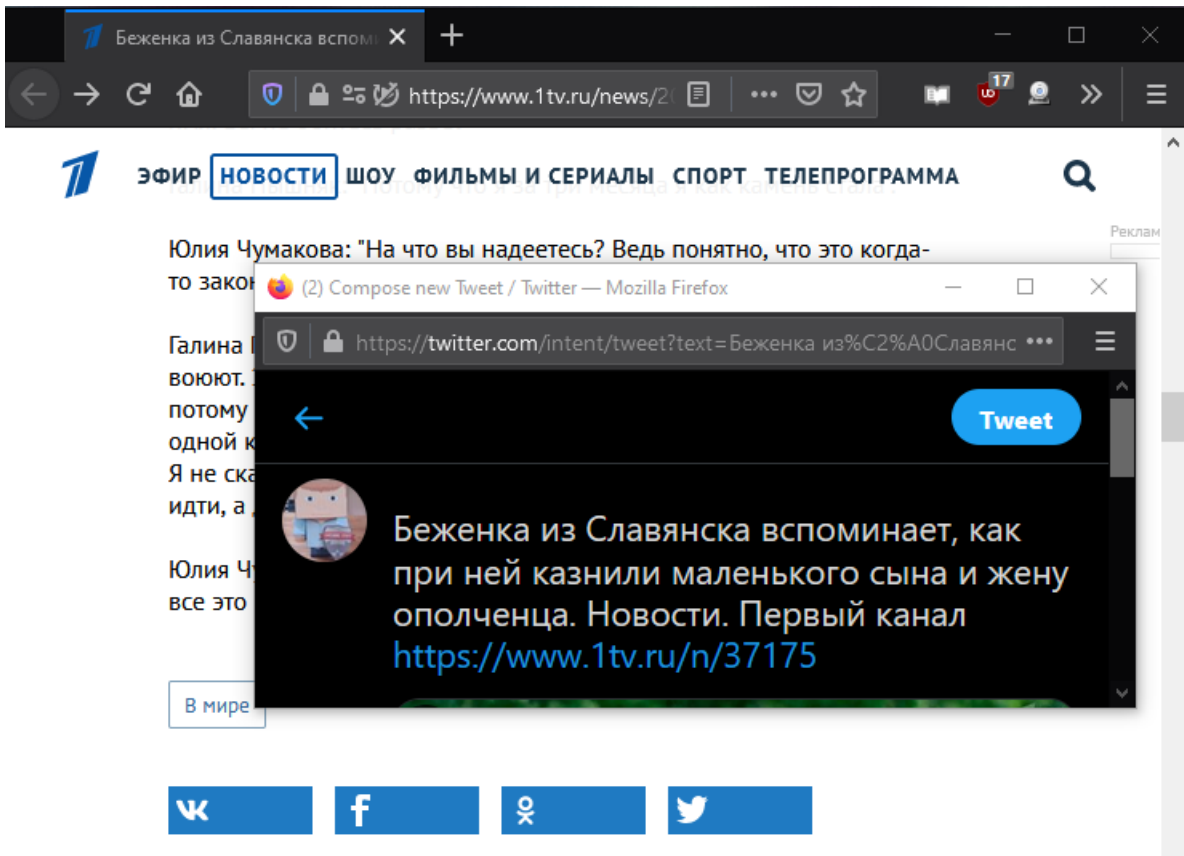


Figure 6.2: Twitter share

Figure 6.2 displays such a sharing functionality for our example URL, clearly resulting in an alternative URL to the same article:

`https://www.1tv.ru/n/37175`

In addition to the URL-shortening provided by news outlets or external services, Twitter also applies its own URL shortening service to any URL shared on the platform, further masking the original search URL¹⁶. Twitter only offers searching for fully unwound URLs (i.e. following all intermediary shortening services and redirects until a final URL location is found) as an enterprise feature¹⁷.

6.2.5 Secondary sources

Besides directly referring to the original news article, it is also possible for other news outlets or Twitter users to repeat the same events in their own words and therefore resulting in different URLs, or omitting these URLs altogether. To cover these Tweets and find any secondary source

¹⁶<https://help.twitter.com/en/using-twitter/url-shortener>

¹⁷<https://developer.twitter.com/en/docs/twitter-api/enterprise/enrichments/overview/expanded-and-enhanced-urls>

URLs, a number of manual keyword searches are performed between the publication date of the original article and one month after. These keywords are based on the title of the original news article.

As an example, the following (date restricted) keyword search can be used to collect additional Tweets repeating the narrative of the news article used above:

```
boy crucified ukraine
```

Any URLs resulting from this keyword search are marked as secondary sources and are used for additional searches. Table B.1 shows a number of URLs collected as secondary sources for the example URL used above.

6.2.6 Searching Twitter

The standard Twitter API does not offer the ability to search historical data¹⁸ without special access. While such access is provided for academic purposes in some cases, it was not available while conducting this research. Because of this, the initial data was collected by scraping the search functionality on the Twitter website. This search functionality offers both historical data and filtering by date. The scraping methodology used to collect this data is described in Appendix D.1. Any further interactions can be fetched through the Twitter API. The overall data collection process is described in Figure 6.3.

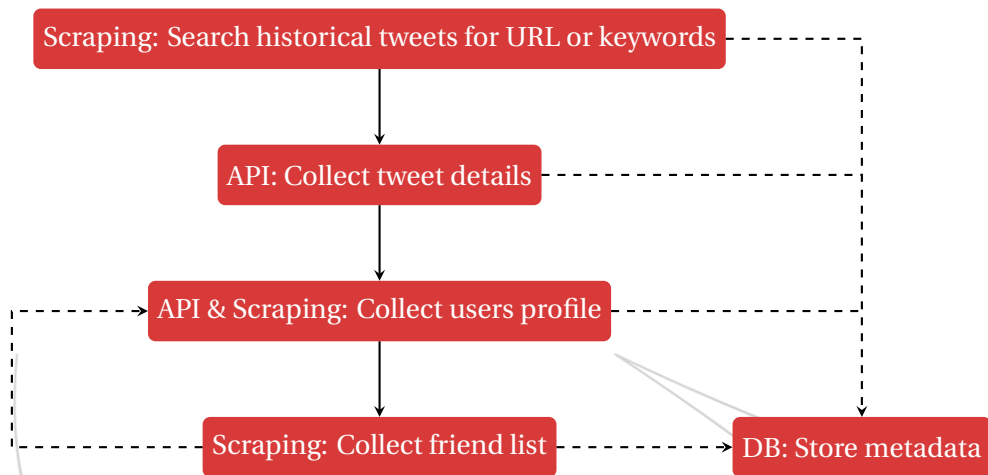


Figure 6.3: Twitter collection process

¹⁸<https://developer.twitter.com/en/docs/twitter-api/search-overview>

6.3 Results

Using the process to collect primary sources only a limited number of Tweets could be found for most news stories. However, a more extensive data set could still be collected from secondary sources using targeted keyword searches (Section 6.2.5). From this result set one specific disinformation campaign is picked, surrounding the downing of passenger flight MH17 over eastern Ukraine. More specifically the theory that the flight was shot down by a Ukrainian air force aircraft, which has been proven false by multiple sources (Higgins, 2015; Ministerie van Justitie en Veiligheid, 2016).

To provide a baseline, a larger data set is collected containing Tweets mentioning the Donbas region. This allows for a better analysis of Twitter users involved in the general discourse surrounding the Russo-Ukrainian conflict, while not only targeting known disinformation. This data set is collected using the same data collection process (Figure 6.3), but only relies on keyword searches.

Each of these data sets consists of the following information:

- The search queries used
- Tweets directly resulting from the search queries
- User profile information for all users involved
- User mentions in the collected tweets
- Hashtags used in the collected tweets
- URLs used the collected tweets

6.3.1 Donbas data set (1)

This data set contains English language tweets mentioning the Donbas region. This was achieved by collecting any tweets containing the keywords *donbass* or *donbas* posted between 02-01-2014 and 31-12-2015 (see Section 6.1). The resulting set contains **155444** tweets, **22068** users, **83142** user mentions, **9993** hashtags and **66565** linked URLs.

6.3.2 MH17 data set (2)

Using the described methods a search is performed that results in English language tweets mentioning the theory that flight MH17 was downed by a Ukrainian fighter jet. This was achieved using the process described above with additional keyword searches for the keywords *MH17 fighter jet*, limited to Tweets posted between 17-07-2014 and 31-09-2014, the day of the flight was shot down and the month after. The resulting set contains **3961** tweets, **3074** users, **1229** user mentions, **369** hashtags and **1701** linked URLs.

6.4 Conclusion

Using the method described in Section 6.2 a data set was formed containing URLs to original disinformation campaign source articles. Directly collecting Twitter data by just using these URLs resulted in a limited number of results. For that reason, alternative links to these articles (shortened URLs) were also collected. Almost all of the articles gathered were published in the native language of the source media outlet, the data set reflected this by not containing any English language tweets referring to the original source articles. To increase the number of results, two steps were taken: Firstly, targeted keyword searches were performed to collect English language tweets up to one month after the publication of the original article. This in turn exposed a number of secondary, English language sources, often consisting of blog articles reciting the same story but not linking to the original source news article. From these results a specific disinformation campaign surrounding the downing of flight MH17 was selected (Data set 2).

To allow for better analysis of RQ2 and RQ3, a much larger additional data set was also created, containing all English language tweets discussing the Donbas region, posted within the selected time frame specified in Section 6.1 (Data set 1).

While Twitter provides an open platform compared to other social media (Section 5.3), some data still could not be retrieved using the publicly available APIs. Because direct search access to the Twitter API was essential, a scraping tool was developed that allows for scraping data directly from the Twitter website (Appendix D.1). Both data set are based on a combination of API responses and scraping.

Chapter 7

Applying mention networks

This chapter further explores the data set collected using the process described in Chapter 6. More specifically it aims to gain a better understanding of relationships between users and topics discussed by these users (RQ2). Firstly this is achieved by looking for structured communities using network analysis. The results of this analysis are then used to perform text analysis for different communities of interest.

7.1 Mention network

Using the data set collected as part of Chapter 6 further analysis can be performed. To gain a better understanding of relationships between users we can study their interactions. In the case of Twitter messages, such interactions are limited to the list provided in Section 6.2.2. By default, Twitter search results do not contain retweets. When only considering tweets themselves, this narrows the interactions down to a single type, user mentions. This approach is based on earlier research by Helmus (2018).

A tweet can contain zero or more mentions to other users. Since a tweet is also created by a user, this allows for creating a directed graph of user-to-user mentions, where each user is presented by a vertex and each mention is presented by a directed edge. By assigning a weight of 1 to each edge, multiple mentions from one user to another user can be represented by taking the sum of these weights. This process results in a graph similar to the one shown in Figure 7.1, the vertices u_1, \dots, u_n indicating unique users, the directed edges indicating the mentions-relationship and the weight assigned to each edge indicating the number of mentions between users.

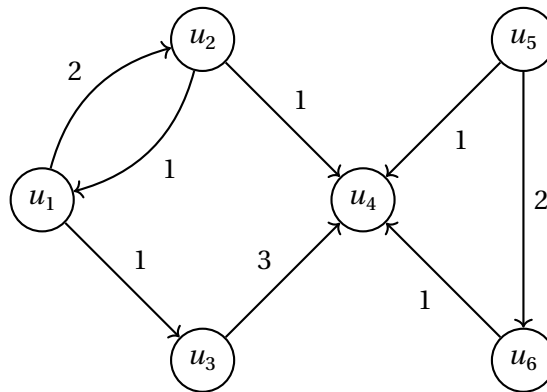


Figure 7.1: Simplified mention network

7.2 Visualization

The visualization shown in Figure 7.1 works well for small graphs, but less so for the size of the collected data sets. For this reason, the Gephi¹ graph visualization tool was used to create more complex visualizations.

Before analyzing the relationships in the mention network, a number of adjustments were made to the data and visual representation. Edge weights are represented by giving each edge a width relative to its weight. Furthermore, the overall activity of a user within the mention network is expressed by the out-degree of each vertex, represented by giving the vertex a size relative to its out-degree. Out-degree was explicitly chosen over the overall degree or in-degree because it better represents initiative. One instance of this difference can be seen with public figures or large organizations (e.g. @BarackObama, @CNN) who are often on the receiving end of mentions but rarely mention individual users themselves, in which case it is more valuable to see who initiates the mention than who receives it.

The mention network is visualized using the force-directed layout algorithm developed by Jacomy et al. (2014). This type of algorithm tries to minimize the number of overlapping edges while retaining an equal edge length for each edge, with each edge behaving like a spring.

The overall result and visualisation for Data set 1 is shown in Figure 7.2 and consists of a total of **18093** vertices and **49454** edges. Based on the visualization it can be concluded that there are a number of mentions that reference users that are not present in any other mention. These are represented by a single vertex fanning out into a set of vertices with a degree of 1. It also shows that there is a large set of vertices that is well connected, which due to the force-directed layout algorithm, shows up as the compact center of the visualization.

¹<https://gephi.org/>

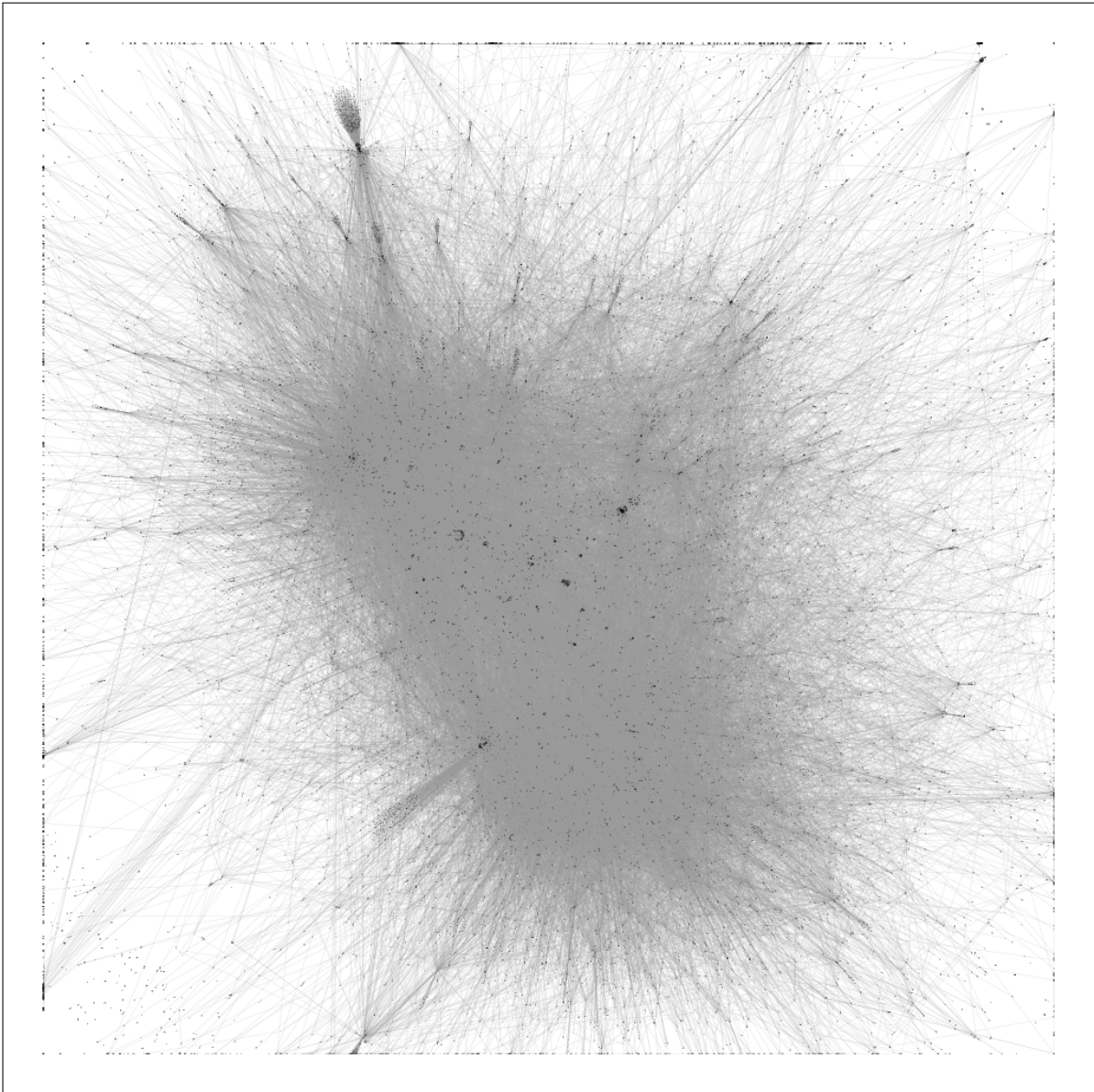


Figure 7.2: Full mention network for Data set 1

Due to the smaller size of Data set 2 the resulting mention network is also a lot smaller mention network as shown in Figure 7.3. This network consists of **1236** vertices and **1095** edges. Overall this mention network is much less connected, as implied by the lower vertex to edge ratio. Nonetheless, a number of clusters stand out.

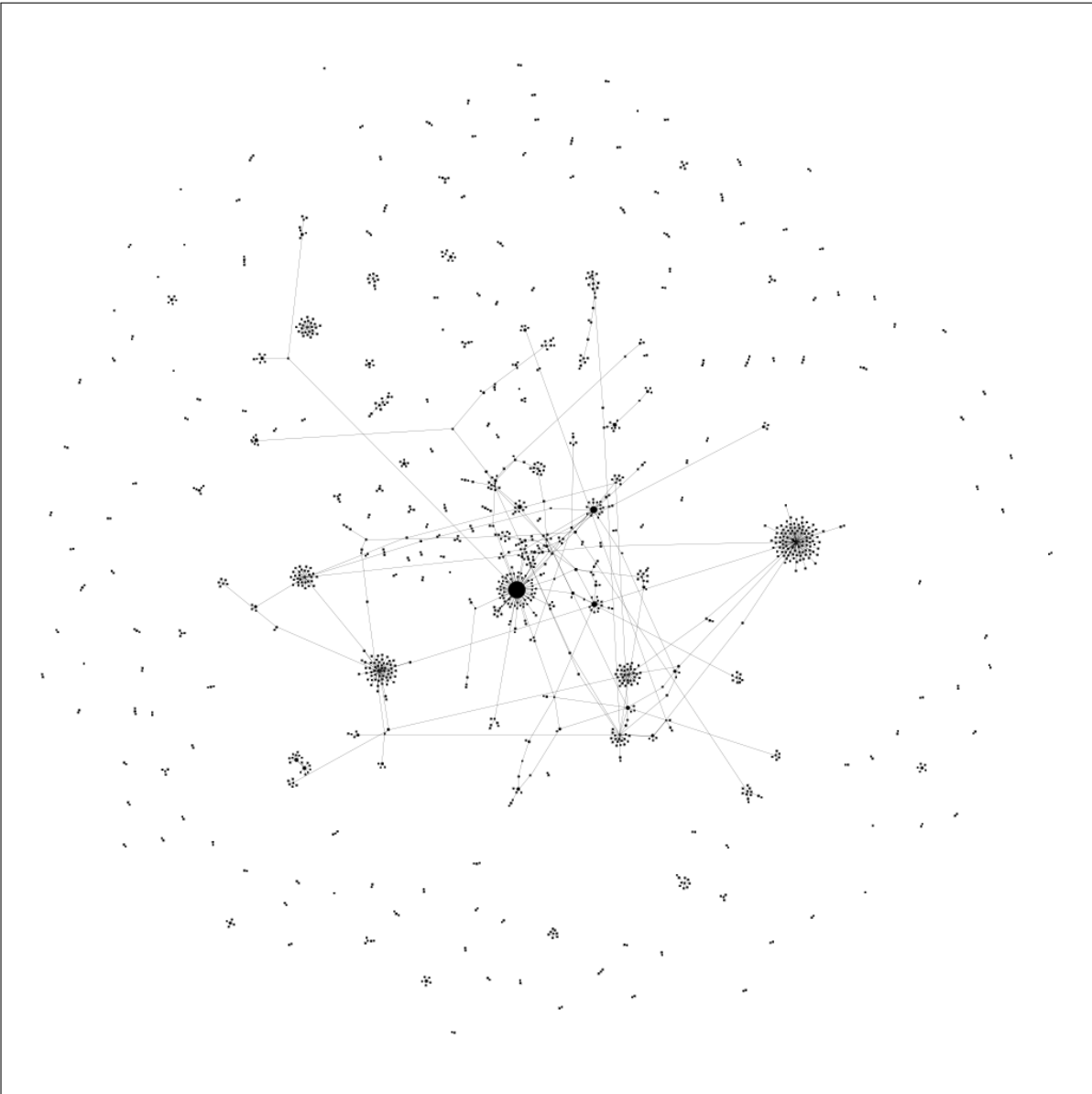


Figure 7.3: Full mention network for Data set 2



7.3 Noise reduction

A large amount of vertices makes it hard to further study and identify relationships between users. Therefore the choice was made to apply noise reduction.

For Data set 1, to reduce the number of overall relationships, all vertices with an in- or out-degree less than 5 or more than 1020 are removed. This eliminates all users that have not initiated any mention from the graph as well as a single outlier that was manually identified as a spam account. Because the graph consists of multiple disconnected components, only the primary (weakly connected) component is considered. This noise reduction method reduces the overall graph to **1727** vertices and **12195** edges as shown in Figure 7.4. The resulting visualization gives a better understanding of the well connected vertices in the center of Figure 7.2, it also gives an initial indication of some form of communities being present by looking at where vertices are clustered together by the force-directed layout algorithm.

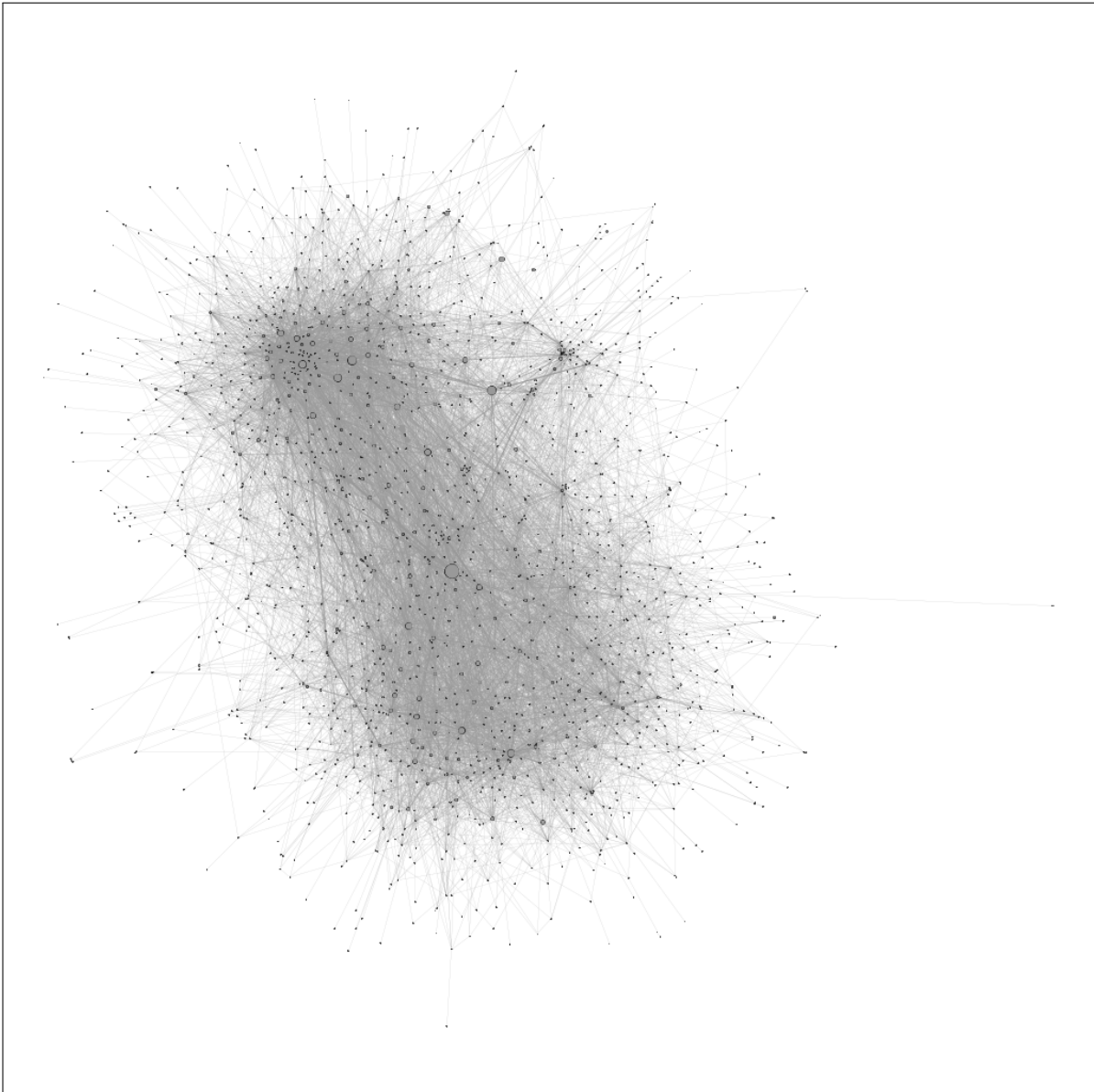


Figure 7.4: Reduced mention network for Data set 1

Due to the smaller size of Data set 2 the graph is only reduced in size by selecting the primary (weakly connected) component. The resulting graph consists of **690** vertices and **733** edges. The resulting graph is shown in Figure 7.5.

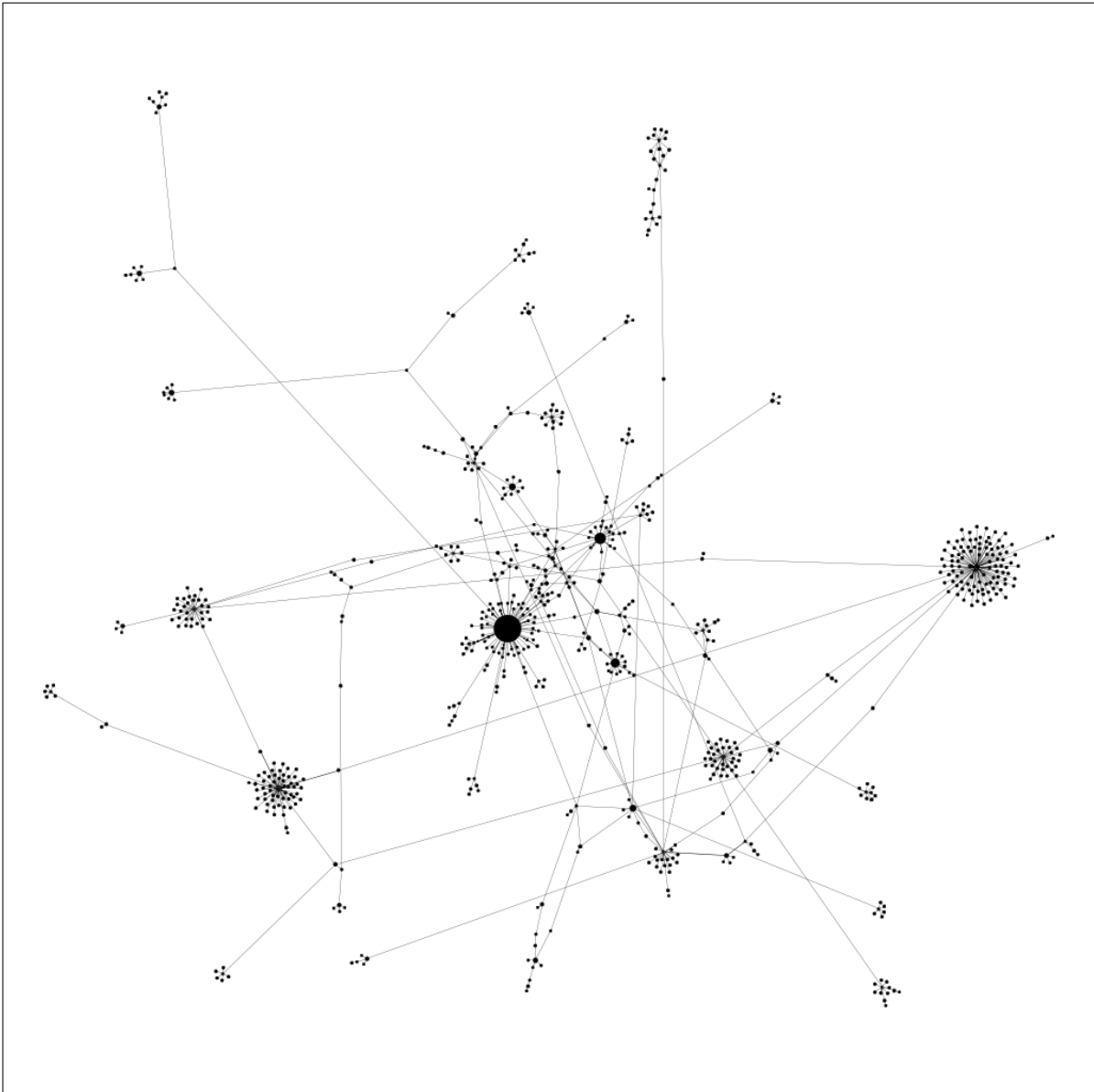


Figure 7.5: Reduced mention network for Data set 2

After performing noise reduction the resulting graphs only retain the strongest relationships, and allow for studying these relationships more closely. This is achieved by using a community detection algorithm to identify the primary communities in the graph.

7.4 Community detection

Community detection algorithms are designed to find groups of vertices within a graph that are more connected internally and less connected to other vertices in the graph. The approach used in this research is adapted from Helmus (2018), however, instead of relying on the Clauset-Newman-Moore community detection algorithm (Clauset, Newman, and Moore, 2004), a different algorithm is used. In this case, the Leiden community detection algorithm developed by Traag, Waltman, and Eck (2019) is employed. This is an extension to the Louvain algorithm developed by Blondel et al. (2008). The Louvain algorithm greatly outperforms Clauset-Newman-Moore in terms of computation time (Blondel et al., 2008), the Leiden algorithm improves this even further (Traag, Waltman, and Eck, 2019) In addition, the Leiden algorithm offers support for directed graphs and its original author provides an implementation in python².

7.4.1 Results

The Leiden algorithm allows for a number of parameters. Firstly the algorithm was instructed to take edge weights into account. In addition, a seed value of 15 was used to ensure consistent results across multiple runs.

Out of the 16 communities found, the six largest are selected for additional analysis because they contained more than 100 users. The selected top six communities and their properties are listed in Table 7.1.

Community	Color	Number of users	%
0	■ (Magenta)	317	18.36
1	■ (Light green)	259	15.00
2	■ (Orange)	241	13.95
3	■ (Light blue)	188	10.88
4	■ (Purple)	152	8.80
5	■ (Mustard)	142	8.22

Table 7.1: Top 6 communities found using Leiden community detection

After applying the community detection algorithm, the Gephi visualization can be improved further by displaying each community in a randomly assigned distinct color, resulting in Figure 7.6 for Data set 1.

²<https://leidenalg.readthedocs.io>

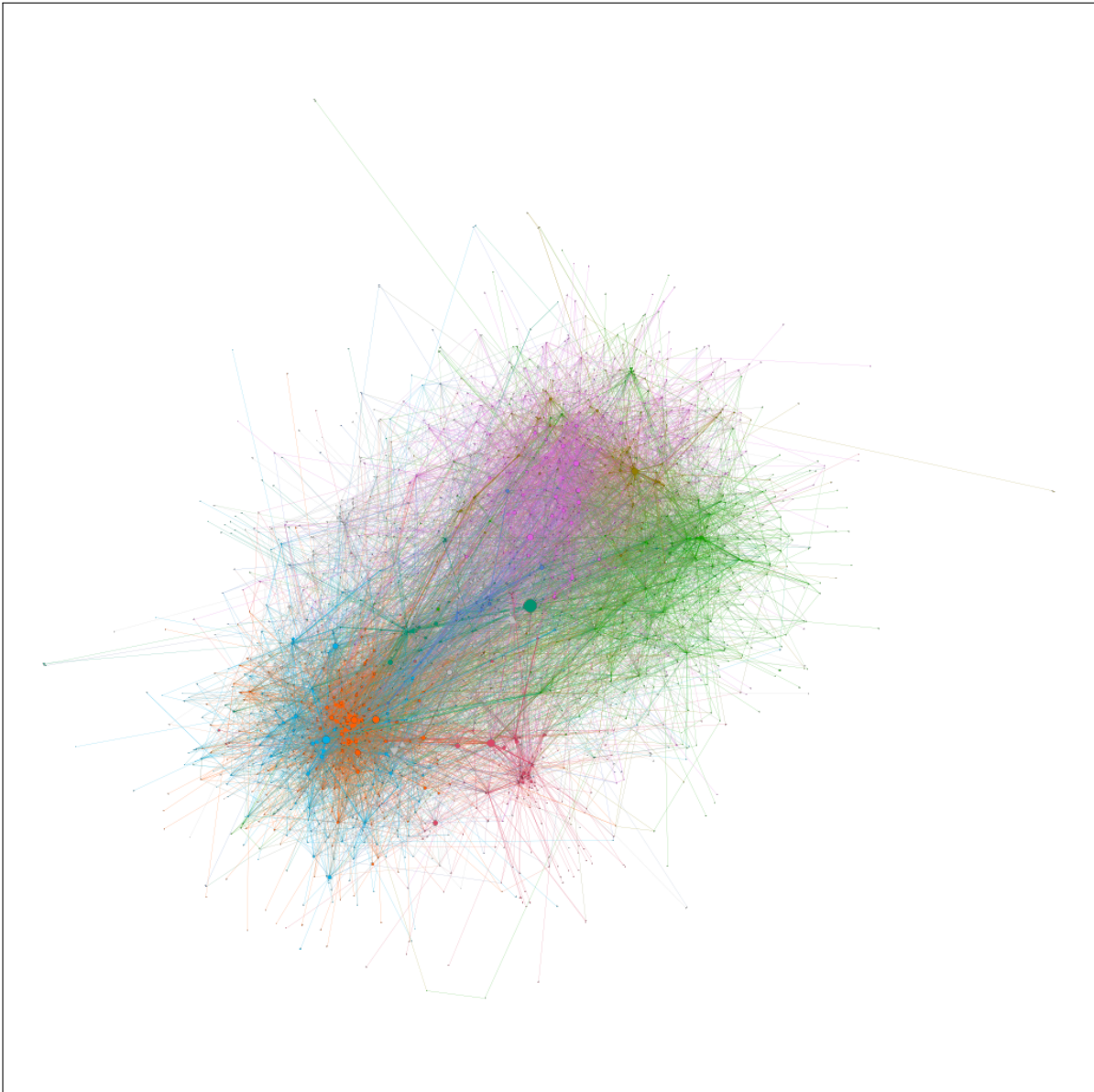


Figure 7.6: Communities found using Leiden community detection

To better understand the sentiment and topics discussed in each community, the contents of the collected messages can be analyzed. For Data set 1 this analysis is performed for each of the top six communities. The limited size of Data set 2 meant that detected communities were very small and consisted of single Tweets and their retweets, not adding any additional insights. For this reason per community analysis is not required for this data set, and therefore performed on the full data set. In the next sections the contents of the collected messages are further studied by looking at their textual content as well as any URLs shared in them.

7.5 Text analysis

To get a better understanding of the overall conversation topic and keywords within such communities, text analysis is applied. Using the communities previously extracted from Data set 1, text analysis is performed on the contents of each message. This analysis was not performed for Data set 2, as it was based on targeted keyword searches and therefore had a predictable textual content with searched keywords being the most commonly occurring terms. The following analysis therefore only focuses on Data set 1.

For each user that is part of one of the top six communities all tweets were collected. These tweets were then further analyzed to extract the top 10 word pairs (bigrams) for each community. This was achieved by using the natural language toolkit (NLTK) library³.

To improve the overall results, a number of preprocessing steps need to be performed: firstly the initial tweet text was stripped of any URLs since they provide no semantic value. In addition, any present HTML encoded entities were decoded (e.g. & becomes &). As a final step, all non-alphanumeric characters were removed with the exception of @ and #, which are used to identify user handles and hashtags in the following steps.

The resulting tweet text was then split using a process known as tokenization as described below. NLTK offers a special Twitter-aware tokenizer⁴ specifically created to correctly treat common Twitter vocabularies such as hashtags and emoticons.

During this process, tokens were converted to lowercase, user mentions (e.g. @CNN) were removed, and repeating characters were reduced to a maximum number of three (e.g. both *helloooo* and *helloooooo* become *hellooo*). To further reduce the number of unique tokens, the # character was removed (e.g. treating both *#ukraine* and *ukraine* as the same token) and tokens were lemmatized. Lemmatization is the process used to reduce a word to a common base form by removing any inflectional endings (e.g. *treaties* becomes *treaty*), which allows for treating different inflections of the same word as one word. Finally, common English stopwords and a list of explicitly excluded terms (Table 7.2) were removed.

³<https://www.nltk.org/>

⁴https://www.nltk.org/_modules/nltk/tokenize/casual.html#TweetTokenizer

Excluded term	Description
via	Often used to mention a user or share a URL
rt	Often used to prefix retweets
donbas	Used as a keyword while collecting Data set 1
donbass	Used as a keyword while collecting Data set 1
ukraine	Very high frequency across all posts
russia	Very high frequency across all posts
luhansk	Very high frequency across all posts
lugansk	Very high frequency across all posts
donetsk	Very high frequency across all posts

Table 7.2: Explicitly excluded tokens

An example of the process described is shown using the fictional tweet below:

```
RT @SomeUser So @CNN, why are you silent on the WAR CRIMES committed in
- donbas, #ukraine?!?! via https://t.co/XXXXXX
```

Which will result in the following tokens:

```
[silent, war, crime, committed]
```

The resulting tweet data was then further analyzed for term frequency. More specifically, for each community, the most common word pairs (bigrams) were identified using using *term frequency-inverse document frequency* (TF-IDF), a method used to create a numeric representation of a set of documents by counting the frequency of specific n-grams relative to their total occurrence across all documents, with bigrams being used as the specific n-gram, and each tweet being considered a document respectively. The advantage of using this method is that it helps reduce the importance of terms that occur very frequently across all documents. The usage of bigrams, as opposed to single words, further reduces the effect of standalone terms and allows for a better understanding of the context in which those words are used.

7.5.1 Results

The results below show the top 20 bigrams for the top 6 communities

Bigrams
russian army
russian terrorist
russian troop
ukrainian army
special status
humanitarian aid
russian force
empr news
russian soldier
ukrainian soldier
russian military
ato ukraine
battalion commander
russian invasion
soldier killed
ukrainian military
russian tank
conflict zone
humanitarian convoy
military equipment

Table 7.3: Cluster 0 top 20 bigrams

Bigrams
дoнбaс дoнбaсс
special status
днр лнр
ukrainian army
russian military
frozen conflict
people republic
nov 2014
russian soldier
look like
humanitarian aid
conflict zone
us uk
russian troop
doesnt want
war crime
russian army
putin want
foxnewsfacts parismarch
kill world

Table 7.4: Cluster 1 top 20 bigrams

Bigrams
msnbc foxnews
ukrainian army
foxnews nyt
kiev junta
novorossiya donbassagainstnazi
nato eu
donbassagainstnazi antifa
kiev regime
humanitarian aid
sputnik international
eng sub
nyt wsj
nyt nato
ethnic cleansing
kiev nazi
save people
right sector
nyt nbc
eu nato
self defense

Table 7.5: Cluster 2 top 20 bigrams

Bigrams
humanitarian aid
novorossiya mariupol
ukrainian army
save_donbass_children ukrarmy
us ows
eng sub
sputnik international
eu uk
fort rus
usaterrorist usafashist
ukrarmy usaterrorist
mariupol kharkiv
special status
kiev novorossiya
humanitarian convoy
novorossiya kiev
dpr lpr
nazi vs
victory people
save people

Table 7.6: Cluster 3 top 20 bigrams

Bigrams
save people
ukrainian army
war crime
kill people
child woman
killed child
child ukrainian
save child
poroshenko fascist
ukrainian troop
fascist poroshenko
people killed
sos people
woman elderly
russian military
poroshenko kill
junta killed
fascist junta
look poroshenko
people look

Table 7.7: Cluster 4 top 20 bigrams

Bigrams
russian troop
special status
russian army
humanitarian aid
russian soldier
ukrainian army
right sector
humanitarian convoy
russian military
look like
artillery strike
conflict zone
osce mission
minsk agreement
ukrainian soldier
people republic
occupied area
putin want
frozen conflict
martial law

Table 7.8: Cluster 5 top 20 bigrams

Clusters 0, 1, and 5 (Tables 7.3, 7.4 and 7.8) use relatively neutral language, considering the fact that the messages mention a military conflict. Furthermore, cluster 0 seems to favor terms commonly used by the Ukrainian government to describe the breakaway republics where the conflict takes place, like *ATO* (Anti-terrorist operation zone).

Clusters 2, 3 and 4 (Tables 7.5 to 7.7) show different patterns. First of all stronger terms with a commonly accepted negative connotation like *nazi*, *junta*, or *fascist* are preferred. In addition, the term *Novorossiia* (New Russia) is used to refer to the breakaway republics. Another interesting pattern that stands out, is a large number of mentions of western powers, geopolitical organizations, or news outlets, like *EU*, *NATO*, *NYT*, or *WSJ*.

7.6 URL analysis

In addition to the text content of the collected tweets, other properties can be extracted from each message, one of such being the URLs shared within these messages. These URLs were excluded as part of the text analysis performed in Section 7.5.

For Data set 1 each community was individually studied for these properties. Since no community detection was performed for Data set 2, the full data set is used.

While Twitter does provide the expanded version of URLs provided by their own URL shortening service, they do not expand any underlying shortened URLs. For this reason, all URLs collected were first normalized using a number of steps.

Initially, each URL is followed until it the request resulted in a non-redirect. Technically speaking this means that HTTP redirects (HTTP status code 301 or 301) are followed until either a correct response (HTTP status code 2XX) or an error (HTTP status code 4XX or 5XX) was presented. When a URL results in an error, it is flagged as such.

The fully expanded URLs can be normalized further by stripping the protocol (i.e. http:// or https://) and only retaining the domain name.

7.6.1 Results

Using these steps the following top 20 URL domains were found for the clusters of Data set 1 identified in Section 7.4.

Domain	Count
www.youtube.com	785
euromaidanpress.com	632
twitter.com	580
j.mp	559
fb.me	480
liveuamap.com	386
www.facebook.com	363
www.unian.info	299
www.interpretermag.com	246
translate.google.com	166
en.censor.net.ua	159
24today.net	152
www.kyivpost.com	133
khpg.org	127
youtu.be	126
maidantranslations.com	88
postmodernnews.com	88
www.rferl.org	75
inforesist.org	70
www.pravda.com.ua	69

Table 7.9: Cluster 0 top 20 URL domains

URL	Count
www.youtube.com	241
twitter.com	181
www.kyivpost.com	118
belsat.eu	70
www.unian.info	67
en.interfax.com.ua	63
www.facebook.com	51
bit.ly	43
www.pravda.com.ua	41
euromaidanpress.com	40
goo.gl	40
www.rferl.org	40
youtu.be	37
millennialmonitor.com	32
www.president.gov.ua	31
zn.ua	25
windowoneurasia2.blogspot.com	24
www.nytimes.com	23
www.theguardian.com	21
uatoday.tv	19

Table 7.10: Cluster 1 top 20 URL domains

Domain	Count
www.youtube.com	1629
tass.com	569
twitter.com	349
www.facebook.com	318
sputniknews.com	301
slavyangrad.org	267
youtu.be	206
rt.com	138
tass.ru	138
translate.yandex.net	134
www.translatetheweb.com	131
www.rt.com	114
news.yahoo.com	103
russia-insider.com	102
en.itar-tass.com	92
shar.es	81
ln.is	80
z5h64q92x9.net	80
ria.ru	74
misc...galleries.shutterstock.com	58

Table 7.11: Cluster 2 top 20 URL domains

URL	Count
www.youtube.com	770
tass.com	357
russia-insider.com	355
sputniknews.com	299
twitter.com	269
ln.is	232
m.vk.com	139
shar.es	112
www.facebook.com	102
www.rt.com	100
youtu.be	84
novorossia.today	77
fb.me	72
russian.rt.com	60
slavyangrad.org	59
fortruss.blogspot.com	55
rt.com	55
j.mp	52
www.kyivpost.com	47
english.pravda.ru	44

Table 7.12: Cluster 3 top 20 URL domains

Domain	Count
www.youtube.com	354
youtu.be	212
twitter.com	146
vk.com	106
www.facebook.com	98
uatoday.tv	87
cs624023.vk.me	59
img-fotki.yandex.ru	59
www.globalresearch.ca	59
wp.me	45
ofa.bo	40
monitor.net.ua	37
goo.gl	33
www.unian.info	33
fb.me	31
racurs.ua	31
news.eizvestia.com	29
www.linkedin.com	28
sputniknews.com	27
pravdoiskatel77.livejournal.com	26

Table 7.13: Cluster 4 top 20 URL domains

URL	Count
liveuamap.com	429
twitter.com	180
www.unian.info	149
www.kyivpost.com	121
www.ukrinform.ua	111
www.youtube.com	103
en.censor.net.ua	85
24today.net	60
euromaidanpress.com	42
youtu.be	33
en.interfax.com.ua	29
joinfo.com	29
translate.google.com.ua	29
inforesist.org	22
www.facebook.com	19
fb.me	18
tass.com	17
pltw.ps	15
www.mfs-theothernews.com	15
zik.ua	15

Table 7.14: Cluster 5 top 20 URL domains

When extending the findings from Section 7.5 to the top URL domains the data shows that links to social media like Twitter, YouTube, or Facebook are very common across all clusters. In these cases, studying the actual information shared was often not possible because the original content had already been removed. Clusters 0, 1 and 5 (Tables 7.9, 7.10 and 7.14) show a strong preference for American and Ukrainian media (and a general pro-Ukrainian tendency), while clusters 2 and 3 (Tables 7.11 and 7.12) show a strong preference for Russian media, most of which were identified as state-controlled in Section 6.1 (Table 6.1) (and therefore have a general pro-Russian tendency). Cluster 4 is an exception as it has a less distinct preference for specific media (Table 7.13), while the text analysis for this cluster (Table 7.7) shows a pro-Russian tendency.

Performing the same analysis for Data set 2 results in the top URL domains listed in Table 7.15.

Again social media and URL-shortening services are very common, leading to similar problems with deleted content. When analyzing the other URL domains, they broadly fit into two categories. Firstly, western news media (i.e. Business insider, The Guardian, CNN, Bloomberg, Telegraph), for each of which the linked articles were still available. These articles contain the initial reporting on the downing of flight MH17 and often refer to the involvement of Ukrainian fighter aircraft as a

URL	Count
www.youtube.com	206
shar.es	176
www.facebook.com	157
www.globalresearch.ca	155
whereisthefuckingplane.com	114
www.businessinsider.com	98
www.theguardian.com	82
edition.cnn.com	75
www.informationng.com	71
www.bloomberg.com	67
www.infowars.com	60
feeds.theguardian.com	55
news.google.com	45
www.rt.com	44
goo.gl	40
watchinga.com	36
linkis.com	31
breitbart.com.feedsportal.com	30
www.presstv.ir	30
telegraph.feedsportal.com	28

Table 7.15: Data set 2 top 20 URL domains

theory proposed by the Russian government. In the second category, alternative news media (i.e. Global Research, RT, Breitbart, PressTV), source articles were often removed. The articles that were still available directly reported on the involvement of Ukrainian fighter aircraft.

7.7 Community sentiment

The results gathered in Section 7.5 and Section 7.6 allow us to broadly classify the sentiment of each community as either pro-Western (including pro-Ukrainian) or explicitly pro-Russian. Table 7.16 lists these sentiments for each community.







Community	Color	Sentiment
0	 (Magenta)	pro-Western / pro-Ukrainian
1	 (Light green)	pro-Western / pro-Ukrainian
2	 (Orange)	pro-Russian
3	 (Light blue)	pro-Russian
4	 (Purple)	pro-Russian
5	 (Mustard)	pro-Western / pro-Ukrainian

Table 7.16: Community sentiment

7.8 Conclusion

Using the data collected as part of Data set 1, we were able to construct and visualize a mention network of the interactions between the most active users in the data set. This was achieved by applying a number of noise reduction strategies. This mention network was then used for further analysis aimed at identifying distinct communities, by applying the Leiden community detection algorithm (Traag, Waltman, and Eck, 2019). Using this algorithm the six largest communities (clusters) were studied for their linguistic properties. Firstly by extracting the most commonly used bigrams as well as extracting the most commonly shared URL domains.

Based on the most commonly used bigrams presented in Section 7.5 it becomes clear that different clusters have a distinctive preference for specific terminology. Based on this terminology the sentiment for each cluster can be identified.

Shared URL domain analysis conducted in Section 7.6 shows that social media URLs are popular across all clusters, but that it is often impossible to analyze the shared materials. When excluding social media URLs, the results do further corroborate the cluster sentiments found in Section 7.5.

The combined results lead to an overall classification of the sentiment in each community extracted from Data set 1, either being a neutral pro-Western (including pro-Ukrainian) or explicitly pro-Russian stance as shown in Section 7.7.

Due to the smaller size of Data set 2 performing community detection was not conducted. Furthermore, due to the targeted keyword search that was used to create the data set, text analysis did not result in any meaningful data. Shared URL domain analysis shows that both Western news articles, as well as alternative news articles, were widely shared after the MH17 crash. One distinct

pattern that can be identified, is that the alternative media had often removed the original article, while the Western news media still offered it. In the cases where the original article could be retrieved, the alternative news media directly supported claims later proven to be false.

Chapter 8

Profiling diaspora membership

This chapter focuses on how to classify Twitter users as Russian diaspora members, Ukrainian diaspora members, or a third group containing users not belonging to either of the previous groups using machine learning methods (RQ3). To do so, firstly part of the data is tagged by hand to be able to train and use supervised machine learning methods. The tagged data is then analyzed to identify and extract features. Three machine learning methods are selected, tuned, and compared with the intent of selecting the best-performing method. The best-performing method is then used to predict diaspora membership for the full data set.

8.1 Data labeling

Supervised machine learning methods require training data to be present, this means that a subset of the analyzed data should be labeled. The labeled data is acquired by taking a random subset of the profiles in Data set 1 assigned to one of the communities found in Section 7.4 and manually determining whether the user is part of the Ukrainian, Russian, or neither diaspora. The latter group was also used for inconclusive results. The tagging process involved a manual analysis of the user's tweets, shared media, given location, followed accounts and, full name to test against the definition of a diaspora member given in Section 5.3.3. In the most clear-cut cases, a user self-identifies in one of their messages or in their profile. Oftentimes users follow accounts from news media in their country of origin or special interest groups for the diaspora. Where needed, a wider search across other social media was conducted. One such search was conducted on LinkedIn, which allows for checking a person's education history and location against the location of their current employment. For example, a user with a name matching Ukrainian naming conventions and a user-specified location in Canada can be tied to a LinkedIn profile with the same name and picture, listed as employed in the same location. This LinkedIn profile is then used to check the user's education history, which lists a high school and university in Lviv, Ukraine.

8.1.1 Dataset balance

Both Russia and Ukraine have a large diaspora community, their total combined number surpassing 15 million people (Chindea, 2008a; Chindea, 2008a; Division, 2020; Ministry of Foreign Affairs of Ukraine, 2019). This is significantly less than the 100+ million monthly active users on Twitter during 2014¹. In addition, not every diaspora member will be active on Twitter. Even when considering that these groups might be represented more in data sets of messages connected to their countries of origin it can still be assumed that the number of users for each of the three categories will be unbalanced.

¹https://s22.q4cdn.com/826641620/files/doc_financials/2014/q4/Q414_Selected_Company_Metrics_and_Financials.pdf

8.1.2 Results

In total **390** profiles out of Data set 1 were tagged by hand. Figure 8.1 shows the proportions of the tagged data.

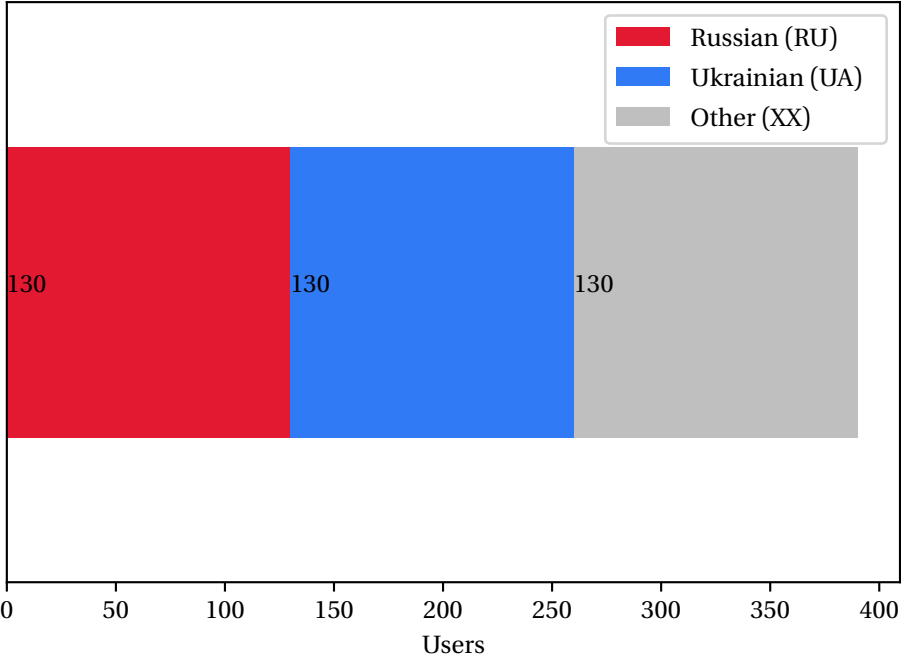


Figure 8.1: Tagged data distribution

8.2 Feature overview

Based on the works presented in Section 4.3, it is determined that using *profile features* and *tweeting behavior* are the most suitable to profile diaspora membership, while *linguistic content* is used for determining community sentiment (Section 7.5). Another reason for omitting the use of *linguistic content* is the requirement to have an overall representation of messages or behavior for a specific user over time. Both Data sets 1 and 2 were collected using targeted keyword searches, and therefore are not necessarily representative for all messages placed by a particular user.

A Twitter profile offers the following information:

- A username
- A full name
- A description
- A profile image
- A location

Pennacchiotti and Popescu (2011) exclude the profile picture as it has often shown to be inconsistent with the rest of the profile (e.g. a picture of a celebrity or model). Based on this information the features presented below are defined based on the **username**, **full name**, **description** and **location** fields (and common patterns seen in those fields) during the data tagging process described in Section 8.1.

The following sections detail each of these features and their collection method.

8.2.1 Direct country mentions

User profile descriptions often contain some sort of indication about the user's profession (e.g. Journalist covering Ukraine for CNN) or location (e.g. Russian living in London) therefore a direct mention to a country name as a noun or an adjective form can provide an indication of a relationship to that country.

Feature	Regular Expression	Description
has_ua_country_mention	ukrain(e ian) (case insensitive)	Username, full name or description contain either <i>ukraine</i> or <i>ukrainian</i>
has_ru_country_mention	russian? (case insensitive)	Username, full name or description contain either <i>rus-sia</i> or <i>russian</i>

Table 8.1: Country mention features

8.2.2 Flag Emoji

With the widespread support of unicode emoji, a common way to express one's affiliation with a country is by using the flag emoji² of that country.





Feature	Regular Expression	Description
has_ua_country_flag		Full name or description contain the Ukraine flag emoji ()
has_ru_country_flag		Full name or description contain the Russia flag emoji ()

Table 8.2: Country flag emoji features

8.2.3 Country code

Since Twitter usernames can only contain alphanumeric roman characters and underscores³, two-letter ISO-3166⁴ country codes are often used in place of country flag emoji. To distinguish the country code from the rest of the name, it is common to use uppercase letters (e.g. SomeUserNL) or underscores (e.g. someuser_nl).

Feature	Regular Expression	Description
has_ua_country_code	<code>[\s _]+(ua UA) </code> <code>(ua UA) [\s _]+ </code> <code>[a-z]+(UA) </code> <code>(UA) [a-z]+</code>	Username, full name or description contain the country code of Ukraine (UA): preceded or succeeded by white space or underscores, or the uppercase code preceded or succeeded by a lowercase string
has_ru_country_code	<code>[\s _]+(ru RU) </code> <code>(ru RU) [\s _]+ </code> <code>[a-z]+(RU) </code> <code>(RU) [a-z]+</code>	Username, full name or description contain the country code of Russia (RU): preceded or succeeded by white space or underscores, or the uppercase code preceded or succeeded by a lowercase string

Table 8.3: Country code features

²<https://unicode.org/emoji/charts/emoji-list.html#country-flag>

³<https://help.twitter.com/en/managing-your-account/twitter-username-rules>

⁴<https://www.iso.org/obp/ui/#search>

8.2.4 Cyrillic Characters

While this research focuses on English language messages on Twitter, the usage of either the Russian or Ukrainian language in the user profile indicates a connection to either country. Both languages use variants of the Cyrillic script which makes it possible distinguish them from other European languages that use variants of the Roman or Greek scripts respectively. However, the general usage of Cyrillic script does not directly indicate either language, since there are many other languages that use variants of it. There are, however, a number of characters that are unique to both Ukrainian and Russian⁵. The whole set of Cyrillic characters can be identified by their Unicode block⁶.

Feature	Regular Expression	Description
has_cyrillic_characters	<code>[\u0400-\u04FF]</code>	Full name or description contain any Cyrillic character based on their unicode range
has_ua_characters	<code>[Ґ ґ І і Ї ї Є є]</code>	Full name or description contain the any of the Cyrillic characters unique to the Ukrainian alphabet: Ґ ґ, І і, Ї ї, or Є є
has_ru_characters	<code>[Ы ы Ё ё Ъ ъ]</code>	Full name or description contain any Cyrillic characters unique to the Russian alphabet: Ы ы, Ё ё, or Ъ ъ

Table 8.4: Cyrillic character features

8.2.5 Naming conventions

A number of naming conventions exist that uniquely identify names of Ukrainian and names of Russian origin. Due to the large internal movement of people within the Soviet Union, naming conventions by themselves can not be considered a very accurate way of identifying diaspora membership. It can, however, add more certainty to the results based on other features. Instead of relying on our own method to detect ethnicity based on naming conventions, we use NamSor⁷ (Carsenat, 2013). NamSor is a data mining tool that predicts the cultural origin of a name based on alphabet and naming conventions. The NamSor API was used to analyze the full name as provided in the users' profile.

⁵<https://web.library.yale.edu/cataloging/music/cyrillic>

⁶<https://unicode.org/charts/PDF/U0400.pdf>

⁷<https://namsor.com/>

Feature	Description
has_ua_name_namsor	Full name is predicted to most likely be a Ukrainian name by NamSor
has_ru_name_namsor	Full name is predicted to most likely be a Russian name by NamSor

Table 8.5: Naming conventions

8.2.6 User-specified location

Twitter allows users to specify their location using a free text field. This location can be a good indicator of whether someone is a diaspora member, as it might show them as living outside of their country of origin. However, since it is a user-provided text field, the location is often left blank or contains a non-existing place. The Google Maps Geocoding API⁸ was used to normalize the location value by converting user-provided locations into a country code.

Feature	Description
google_maps_country_code	The first country code returned by the Google Maps Geocoding API

Table 8.6: User specified location

8.2.7 Government and state-affiliated accounts

8.2.7.1 Twitter-annotated accounts

Twitter annotates government and other state-affiliated with a special label⁹. While this information is not available through the Twitter API, it is gathered using scraping as part of the data collection process described in Section 6.2.6 and the data collection method described in Appendix D.1. Using this annotation, the affiliated country can easily be extracted from the Twitter-provided label (e.g. *Russia government organization* as shown on the right hand side of Figure 8.2).

8.2.7.2 Non-annotated accounts

Manual analysis of the retrieved annotations shows that they are not consistently applied. Figure 8.2 shows a side-by-side view of two Russian embassy accounts, of which only one is annotated as a government organization.

⁸<https://developers.google.com/maps/documentation/geocoding/overview>

⁹<https://help.twitter.com/en/rules-and-policies/state-affiliated>



Figure 8.2: Non-annotated and annotated accounts

Further analysis of these accounts did show other useful patterns. Embassy accounts are of particular interest here, because these account names often follow a preset format. Table 8.7 shows a number of Embassy account names and their profile descriptions.

Username	Full name	Description
UKRinMKD	UKR Embassy in MKD	Ukrainian Embassy in the Republic of North Macedonia/Посольство України в Республіці Північна Македонія/Амбасада на Украина во Република Северна Македонија
UKRinMNE	UKR Embassy in MNE	Ukrainian Embassy in Montenegro / Посольство України в Чорногорії /Ambasada Ukraїne u Crnoj Gori
UKRinNLD	UKR Embassy in NLD	Embassy of Ukraine in the Kingdom of the Netherlands/Посольство України в Королівстві Нідерланди
RusEmbEst	Russia in Estonia	льство России в Эстонии / Russian Embassy in Estonia / Venemaa Suursaatkond Eestis.
RusEmbEthiopia	Russia in Ethiopia	Официальный аккаунт Посольства России в Эфиопии и при Африканском союзе. Official account of the Russian Embassy in Ethiopia and to the African Union.
RusEmbIndia	Russia in India	The official account of the Russian Embassy in India

Table 8.7: Official accounts

Based on this analysis, a search was performed and accounts were manually tagged with the representative country and host country, for example, *RusEmbEst* was tagged as representing Russia in Estonia. The full list of tagged accounts can be found in Appendix C. This way an extensive list of annotated government or state-affiliated accounts was created.

Feature	Description
is_ua_state_affiliated	The profile is annotated as a government or state-affiliated account representing Ukraine
is_ru_state_affiliated	The profile is annotated as a government or state-affiliated account representing Russia

Table 8.8: State affiliation labels

8.3 Feature selection

Before applying a machine learning method, it is important to select the right features. Each of the features presented in Section 8.2 is manually analyzed. Based on this analysis we decided to exclude the *naming conventions* and *location* features from the final model, as both rely on an external service which in turn relies on user-provided data of varying quality. However, these features are used as a guideline during the manual data labeling process. For example, users having names matching Slavic naming conventions were manually checked. Both to quickly find users that could be classified in one of the diaspora groups as well as providing counterexamples for people being part of other Slavic diaspora groups. The user-specified location is used to flag users that are currently located in Russia or Ukraine for manual inspection, since those users should not be considered diaspora members when living in their country of origin.

8.3.1 Profile features

The selection of features directly inferred from a user profile therefore consists of the following features, as defined in Section 8.2:

- has_ua_country_mention
- has_ua_country_flag
- has_ua_country_code
- has_ua_characters
- is_ua_state_affiliated
- has_ru_country_mention
- has_ru_country_flag
- has_ru_country_code
- has_ru_characters
- is_ru_state_affiliated
- has_cyrillic_characters

8.3.2 Followed account features

While a user is in control of the information on their own profile, this is not the case with the profiles of the people they follow. This means that these followed profiles can provide additional insights about a user. Depending on the amount of information such accounts offer, it can provide valuable insights about the origin and/or location of a user and can be used to boost the overall precision of the user classification (Zamal, W. Liu, and Ruths, 2012).

For example: to keep in touch with information related to their country of origin, diaspora members can follow accounts related to their home country. Such accounts can include diaspora organizations and embassies in their current country of residence.

To take this information into account, the feature set mentioned above also applies to each account followed by a particular user. The feature set defined above is therefore extended with the value for each feature across all followed accounts expressed as a percentage of all followed accounts:

- `has_ua_country_mention_friends`
- `has_ua_country_flag_friends`
- `has_ua_country_code_friends`
- `has_ua_characters_friends`
- `is_ua_state_affiliated_friends`
- `has_ru_country_mention_friends`
- `has_ru_country_flag_friends`
- `has_ru_country_code_friends`
- `has_ru_characters_friends`
- `is_ru_state_affiliated_friends`
- `has_cyrillic_characters_friends`

For example: If a user follows 10 accounts, of which 5 have *has_ua_country_mention* marked as 1 (true), this results in a *has_ua_country_mention_friends* value of 0.5.

8.3.3 Feature correlation

To determine the final set of features to be used, any correlation between them correlation is studied using the Pearson correlation coefficient¹⁰. The resulting correlation matrix is shown in Figure 8.3.

¹⁰<https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html>

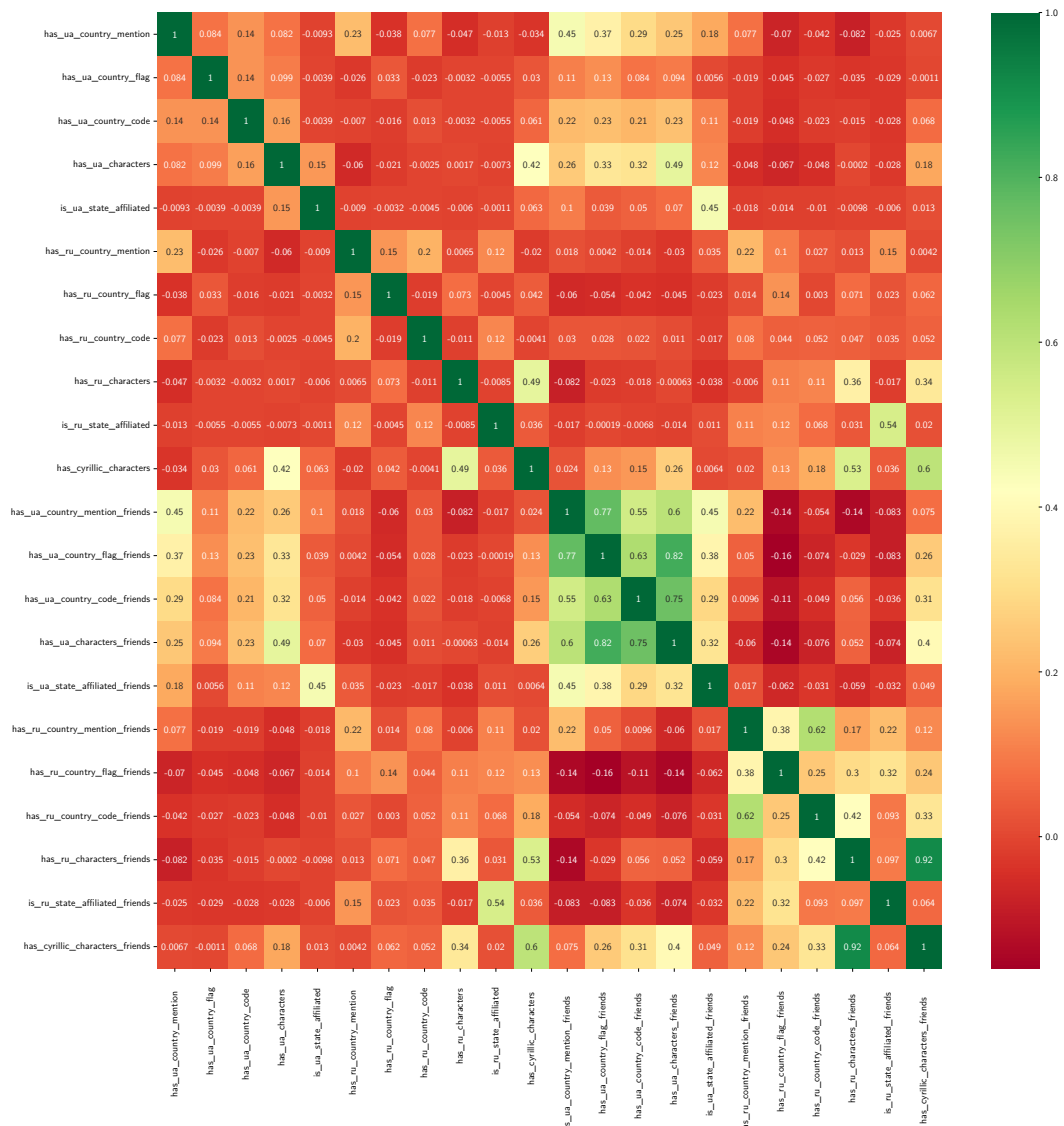


Figure 8.3: Pearson coefficient correlation matrix for all features

When analysing the matrix, it becomes clear that there are features showing some degree of correlation, ranging from moderate (i.e. a correlation coefficient between 0.5 and 0.7) to strong (i.e. a correlation coefficient larger than 0.7). These features are shown in Table 8.9.

Feature A	Feature B	Correlation coefficient
has_cyrillic_characters_friends	has_ru_characters_friends	0.92
has_ua_characters_friends	has_ua_country_flag_friends	0.82
has_ua_country_flag_friends	has_ua_country_mention_friends	0.77
has_ua_characters_friends	has_ua_country_code_friends	0.75
has_ua_country_code_friends	has_ua_country_flag_friends	0.63
has_ru_country_mention_friends	has_ru_country_code_friends	0.62
has_ua_characters_friends	has_ua_country_mention_friends	0.60
has_cyrillic_characters_friends	has_cyrillic_characters	0.60
has_ua_country_code_friends	has_ua_country_mention_friends	0.55
is_ru_state_affiliated	is_ru_state_affiliated_friends	0.54
has_ru_characters_friends	has_cyrillic_characters	0.53

Table 8.9: Strongly correlating features

These results show that strong correlation can be detected between different country indicators, this is especially visible for any features related to friends, as these are represented as continuous variables (a percentage across all friends) as opposed to the user specific discrete Boolean values. This does explain why the different indicators referencing the country of origin show a strong correlation across friends. Furthermore the features identifying the Russian and Ukrainian script are subsets of the feature identifying the use of any Cyrillic characters. However, because all features are eventually based on discrete Boolean values, the presence of one indicator does not automatically imply the presence of another indicator for a single user. In this case having multiple indicators is intended to result in a stronger bias towards a specific classification.

8.3.4 Results

The overall selection results in the exclusion of two features from the originally proposed feature set, resulting in the 22 features listed in Section 8.3.1 and Section 8.3.2. These features are therefore used as part of the final model. Using a correlation matrix it is determined that correlation indeed exists but that the discrete original of these features does not allow for ruling them out of the final model.

8.4 Model selection

Using tagged data defined in Section 8.1 and the features defined in Section 8.3, a selection of supervised learning algorithms (Section 5.4) can be made.

8.4.1 Performance metrics

Due to the unbalanced nature of the number of diaspora members amongst all Twitter users, F_1 -score is used as the primary performance metric. The F_1 -score is defined by the harmonic mean of the *precision* (the percentage of actual diaspora members among all matches) and *recall* (the percentage of matched diaspora members among all actual diaspora members) metric. In the case of multi-class classification tasks, these metrics are calculated for each class. For these metrics, a split of **70%** training data and **30%** testing data is made. In addition, there are a number of aggregate scores that can be used to measure the performance of the model as a whole. In the scope of this research the aggregated micro F_1 -score (μF_1) is used as this measure works well with imbalanced data sets (Sokolova and Lapalme, 2009), which is expected due to the large total number of Twitter users versus the size of the diaspora of each country (Section 8.1). When all classes are considered, the micro F_1 -score is equal to the accuracy of the model.

8.4.2 Performance results

Each of the algorithms described in Section 5.4 is trained and tested to get an indication of their performance for the given classification task. Table 8.10 shows the micro F_1 -score for each of the different algorithms using their default parameters as implemented in SciKit Learn¹¹.

Algorithm	Accuracy / μF_1
Random Forest (RF) ¹²	0.85
Gradient Boosted Decision Trees (GBDT) ¹³	0.81
Support Vector Machines (SVM) ¹⁴	0.81
Logistic Regression (LR) ¹⁵	0.80
Naive Bayes (NB) ¹⁶	0.79
K-Nearest Neighbours (KNN) ¹⁷	0.78
Decision Tree (DT) ¹⁸	0.78

Table 8.10: Supervised learning algorithm performance

¹¹<https://scikit-learn.org>

In addition to the results found, Gradient Boosted Decision Trees have been used successfully for a similar classification tasks performed by Pennacchiotti and Popescu (2011). Furthermore, Caruana and Niculescu-Mizil (2006) show that Random Forests, Gradient Boosted Decision Trees, and Support Vector Machines were the best performing algorithms across 11 different classification problems. Based on findings the top three methods are selected for further analysis: Random Forest (RF), Gradient Boosted Decision Trees (GBDT), and Support Vector Machines (SVM). The following sections study the detailed performance results of each of these three algorithms.

8.4.2.1 Random Forest

Table 8.11 shows the performance metrics of the Random Forest classifier. Overall the model performs well with a high accuracy of **0.85**. Not all categories perform equally, with the non-diaspora category (XX) performing the worst. This can be corroborated by studying the confusion matrix, shown in Figure 8.4. The confusion matrix indicates that the classifier never incorrectly classified a Russian diaspora user as Ukrainian and vice versa. However, there were **5** cases in which a non-diaspora user was classified as Russian, **4** cases in which a non-diaspora user was classified as Ukrainian, **4** cases in which a Russian diaspora user was classified and being a non-diaspora user, and **5** cases in which a Ukrainian diaspora user was classified as being a non-diaspora user. Section 8.4.3 studies these classification errors in more detail.

	Precision	Recall	F1-score	Support
RU	0.87	0.89	0.88	38
UA	0.89	0.87	0.88	38
XX	0.78	0.78	0.78	41
Macro avg	0.85	0.85	0.85	117
Weighted avg	0.85	0.85	0.85	117
Accuracy / μF_1	0.85			

Table 8.11: Classification report for RF

¹²<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

¹³https://xgboost.readthedocs.io/en/latest/python/python_api.html?highlight=scikit#xgboost.XGBClassifier

¹⁴<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>

¹⁵https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

¹⁶https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

¹⁷<https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>

¹⁸<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>

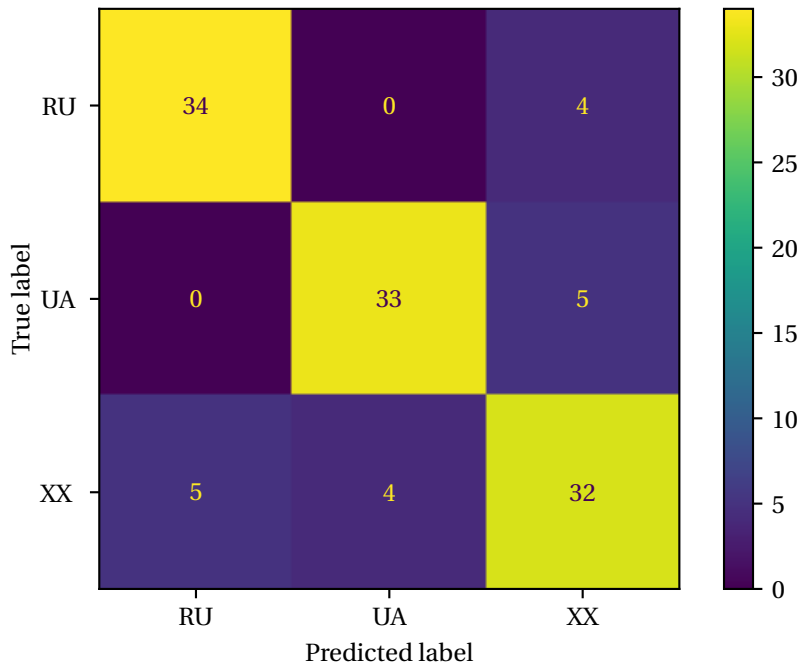


Figure 8.4: Confusion matrix for RF

8.4.2.2 Gradient Boosted Decision Trees

Table 8.12 shows the performance metrics of the Gradient Boosted Decision Tree classifier. Like the Random Forest classifier, the model performs well with a high accuracy of **0.81**. Not all categories perform equally, with the non-diaspora category (XX) performing the worst. This can be corroborated by studying the confusion matrix, shown in Figure 8.5. The confusion matrix indicates that the classifier never incorrectly classified a Russian diaspora user as Ukrainian and vice versa. However, there were **6** cases in which a non-diaspora user was classified as Russian, **6** cases in which a non-diaspora user was classified as Ukrainian, **3** cases in which a Russian diaspora user was classified as being a non-diaspora user, and **7** cases in which a Ukrainian diaspora user was classified as being a non-diaspora user. Section 8.4.3 studies these classification errors in more detail.

	Precision	Recall	F1-score	Support
RU	0.85	0.92	0.89	38
UA	0.84	0.82	0.83	38
XX	0.74	0.71	0.72	41
Macro avg	0.81	0.81	0.81	117
Weighted avg	0.81	0.81	0.81	117
Accuracy / μF_1	0.81			

Table 8.12: Classification report for XGB

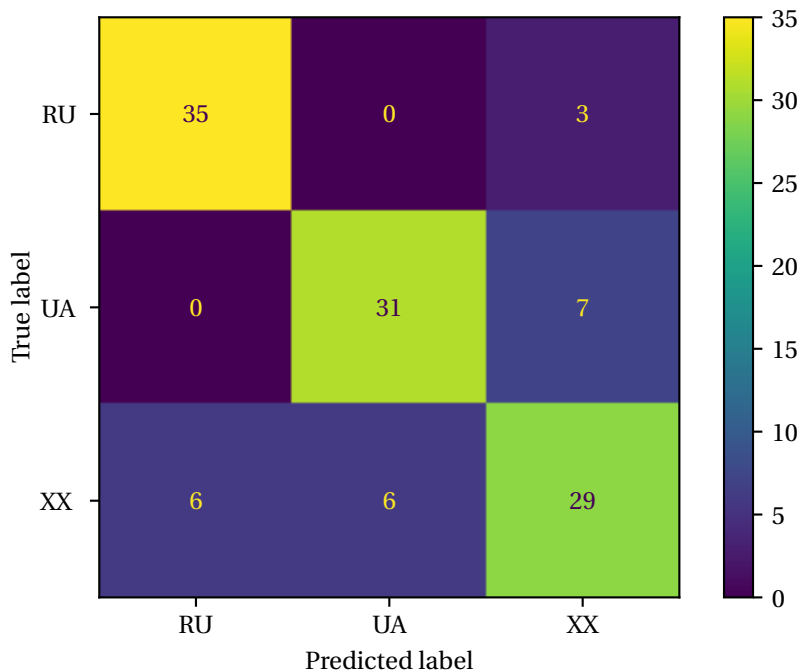


Figure 8.5: Confusion matrix for XGB

8.4.2.3 Support Vector Machines

Table 8.13 shows the performance metrics of the Support Vector Machines classifier. Like the previous two models, the model performs well with a high accuracy of **0.81**. By studying the confusion matrix, shown in Figure 8.6, a number of observations are made. As opposed to the previous two models, the Support Vector Machines classifier did incorrectly predict Ukrainian diaspora users as Russian in 4 cases. No Russian diaspora users were classified as Ukrainian. In addition, there were 3 cases in which a non-diaspora user was classified and Russian, 4 cases in which a non-diaspora user was classified as Ukrainian, 5 cases in which a Russian diaspora user was classified and be-

ing a non-diaspora user, and **6** cases in which a Ukrainian diaspora user was classified as being a non-diaspora user. Section 8.4.3 studies these classification errors in more detail.

	Precision	Recall	F1-score	Support
RU	0.82	0.87	0.85	38
UA	0.88	0.74	0.80	38
XX	0.76	0.83	0.79	41
Macro avg	0.82	0.81	0.81	117
Weighted avg	0.82	0.81	0.81	117
Accuracy / μF_1	0.81			

Table 8.13: Classification report for SVM

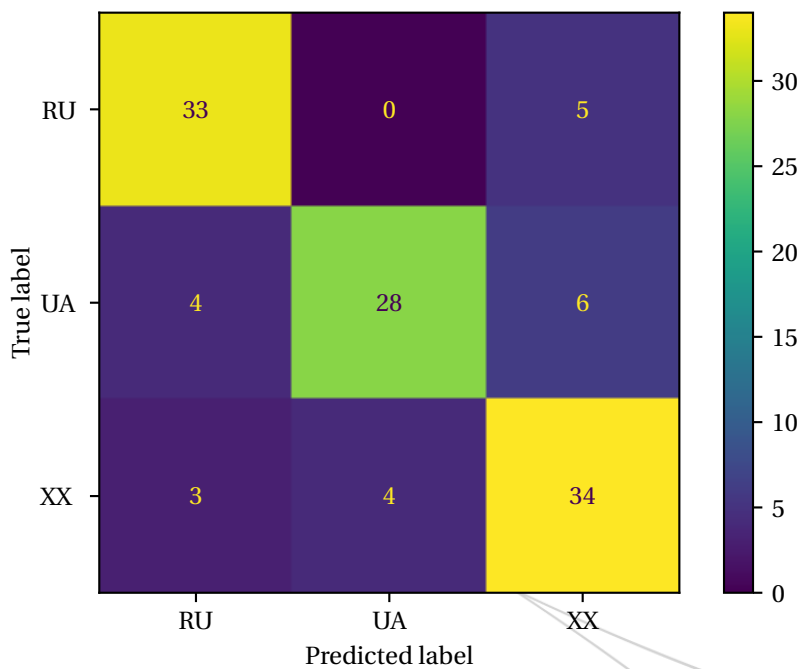


Figure 8.6: Confusion matrix for SVM

8.4.3 Classification errors

Each model shows a similar distribution of incorrect classifications. The incorrect classification for the non-diaspora category (XX) can be explained by the fact that this category encompasses any user not fitting in either of the diaspora groups and might therefore represent a much more diverse range of feature values. The incorrect classification of non-diaspora users as either Russian or Ukrainian can be explained by the set containing Russian or Ukrainian users that live in their

home country. These users might show similar values for some of the selected features as diaspora members while being labeled as non-diaspora.

8.4.4 Cross-validated results

The reported metrics for each of the selected algorithms (using default parameter values) show promising results with an accuracy / μF_1 -score of over **0.80**, with Random Forests performing the best reaching **0.85**.

Cross-validation is used to further compare each of the models. Cross-validation provides two distinct benefits over a regular split of training and test data. First of all, it has the benefit of being able to use the complete set of tagged data. In addition, it reduces the risk of overfitting the model to the training data which can lead to issues generalizing the model to unseen data. KFold achieves this by splitting the training data set into multiple subsets (folds) and using each subset for validating the model trained on the other subsets. In the scope of this research, cross-validation is performed using Stratified KFold. Stratified Kfold uses the same approach as KFold but also ensures that each individual subset contains a number of labels reflecting the training data set. While not strictly necessary here, this approach covers cases where the training data set contains an unbalanced set of labels (Alpaydin, 2010, p. 486). A total of **5** folds is used, as this ensures each fold, consisting of **78** users, with an equal distribution of each label, is large enough relative to the total number of tagged users.

Figure 8.7 display the resulting Accuracy / μF_1 -scores for the cross-validated models. Here it becomes clear that Random Forests show the largest variation in scores of **0.71-0.92**, but the highest median score at **0.83**. Gradient Boosted Decision Trees show a slightly smaller variation of **0.71-0.90**, but has the lowest median score at **0.78**. The results for Support Vector Machines the least variation in scores of **0.75-0.87** combined with a high median score at **0.81**. Both Random Forests and Gradient Boosted Decision Trees show a lower Accuracy / μF_1 -score after cross-validation.

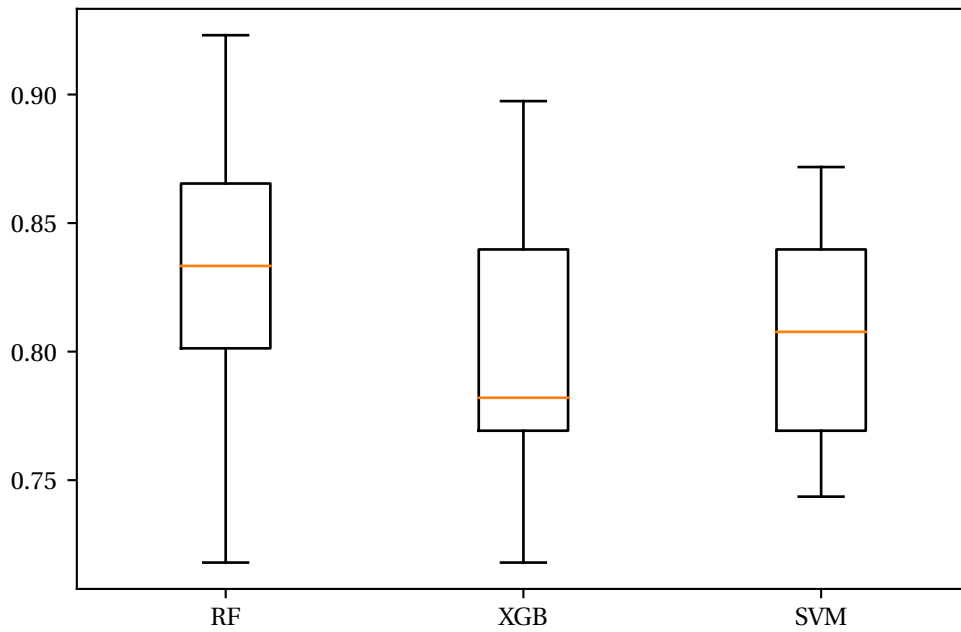


Figure 8.7: Cross-validated model Accuracy / μF_1 -scores

Overall the results look promising with high median values. However, the overall range of Accuracy / μF_1 -scores shows noticeable differences between folds. To improve the selected models, it is important to not only increase the median value but also to decrease the variation in scores between folds.

8.5 Model tuning

While the initial analysis gives an initial indication of which algorithms performed best, each of the algorithms provides a set of so-called hyper-parameters used to tweak the algorithm. Manually finding the ideal set of these parameters for each of the selected algorithms is infeasible, therefore automated techniques for hyper-parameter tuning are used.

Even when automated, the operation of fitting and validating a new model for each possible combination of parameters is computationally expensive, therefore a less demanding tuning technique is used first in the form of a random search¹⁹. Random search takes all possible parameter configurations within the given ranges and randomly selects a predefined number of combinations to train the model with. The results of each of these models are then compared using a scoring function, resulting in the best scoring configuration (within the selected set) being returned. After determining a sensible parameter range for each algorithm using random searches, an exhaustive grid search²⁰ is conducted to find the best performing configuration within those ranges. The exhaustive search works similar to a random search, but instead tests every possible parameter configuration within the given ranges. To validate the results during the parameter model tuning process, each of the searches uses the same cross-validation technique and scoring function as is used for the initial analysis.

8.5.1 Performance metrics

In this section, the results of each of the optimized models are studied in detail and compared to the results of the models before optimization.

8.5.1.1 Random Forest

Table 8.14 shows the performance metrics of the Random Forest classifier after optimization. The overall accuracy increased from **0.85** before hyper-parameter tuning to **0.88**. The confusion matrix, shown in Figure 8.8, indicates that this is caused by a correctly classifying **4** non-diaspora user previously classified as Russian diaspora users. Table 8.15 describes the non-default parameter values found during optimization.

¹⁹https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV

²⁰https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

	Precision	Recall	F1-score	Support
RU	0.97	0.89	0.93	38
UA	0.89	0.87	0.88	38
XX	0.80	0.88	0.84	41
Macro avg	0.89	0.88	0.88	117
Weighted avg	0.89	0.88	0.88	117
Accuracy	0.88			

Table 8.14: Classification report for optimized RF

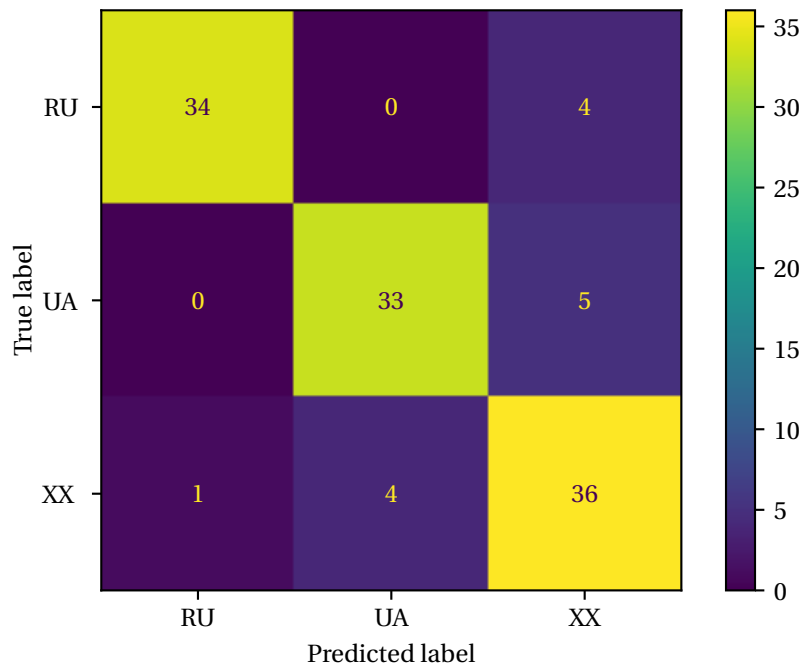


Figure 8.8: Confusion matrix for optimized RF

Parameter	Optimized value	Description
min_samples_split	5	The minimum number of samples required to split an internal node
min_samples_leaf	2	The minimum number of samples required to be at a leaf node
n_estimators	1200	The number of trees in the forest

Table 8.15: Parameters used to optimized RF

8.5.1.2 Gradient Boosted Decision Trees

Table 8.16 shows the performance metrics of the Gradient Boosted Decision Tree classifier after optimization. Like the Random Forest classifier, the model performed better after optimization, increasing the overall accuracy from **0.81** to **0.85**. Again this performance increase primarily comes from a better classification of non-diaspora users, correctly classifying **4** more users as shown in Figure 8.9. Table 8.17 describes the non-default parameter values found during optimization.

	Precision	Recall	F1-score	Support
RU	0.97	0.89	0.93	38
UA	0.82	0.87	0.85	38
XX	0.79	0.80	0.80	41
Macro avg	0.86	0.86	0.86	117
Weighted avg	0.86	0.85	0.86	117
Accuracy	0.85			

Table 8.16: Classification report for optimized XGB

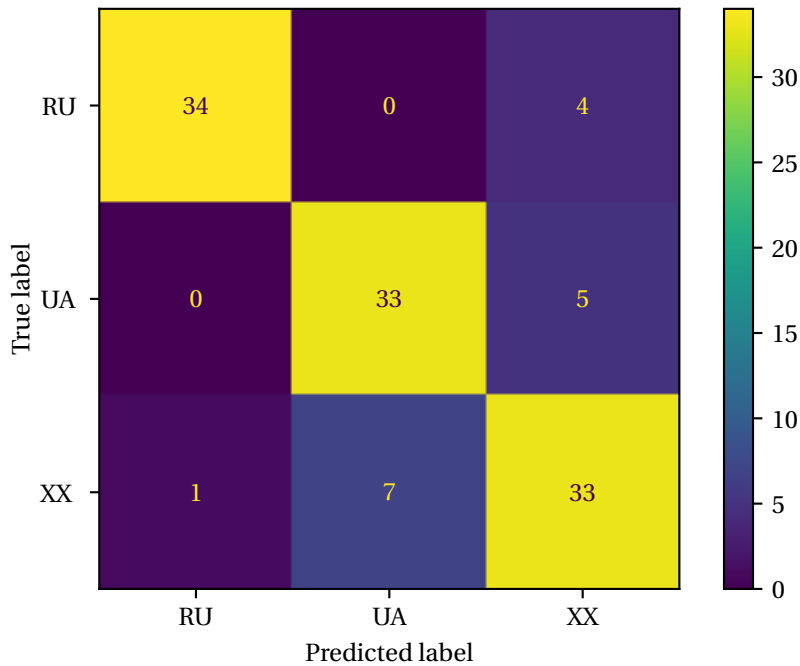


Figure 8.9: Confusion matrix for optimized XGB

Parameter	Optimized value	Description
tree_method	hist	Histogram optimized approximate greedy algorithm
subsample	0.8	The subsample ratio of the training instances
n_estimators	100	Number of gradient boosted trees
min_child_weight	10	Minimum sum of instance weight needed in a child
max_depth	5	Maximum depth of a tree
learning_rate	0.01	Step size shrinkage used in update to prevents overfitting
gamma	1.5	Minimum loss reduction required to make a further partition on a leaf node of the tree
colsample_bytree	0.3	The subsample ratio of columns when constructing each tree

Table 8.17: Parameters used to optimized XGB

8.5.1.3 Support Vector Machines

Table 8.18 shows the performance metrics of the Support Vector Machines classifier after optimization. This classifier shows the largest increase in overall accuracy, increasing from **0.81** to **0.87**. The confusion matrix, shown in Figure 8.10, shows that this classifier performs the best at correctly

classifying non-diaspora users. In three cases this classifier performed worse than the other models, incorrectly classifying Ukrainian diaspora users as Russian diaspora users. Table 8.19 describes the non-default parameter values found during optimization.

	Precision	Recall	F1-score	Support
RU	0.89	0.87	0.88	38
UA	0.94	0.82	0.87	38
XX	0.81	0.93	0.86	41
Macro avg	0.88	0.87	0.87	117
Weighted avg	0.88	0.87	0.87	117
Accuracy	0.87			

Table 8.18: Classification report for optimized SVM

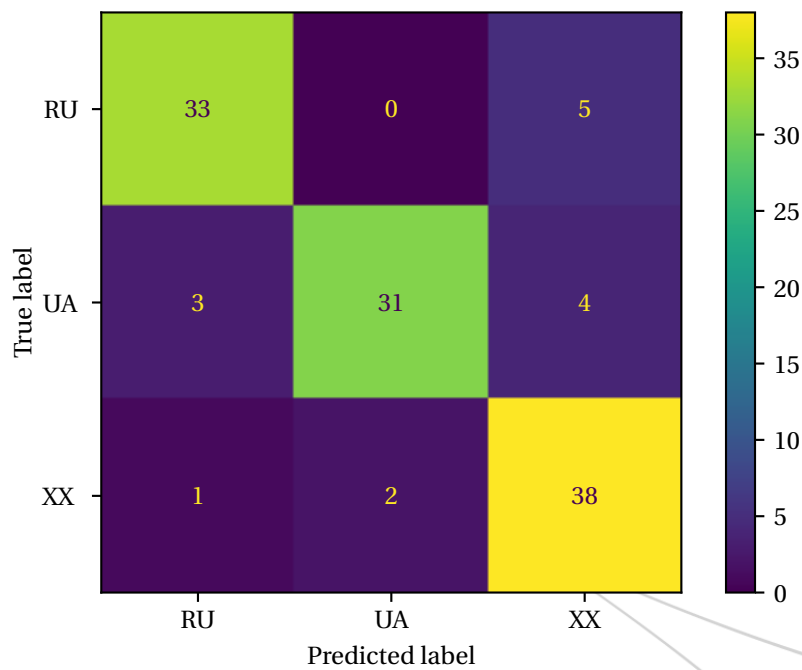


Figure 8.10: Confusion matrix for optimized SVM

Parameter	Optimized value	Description
C	100	Regularization parameter
kernel	linear	The kernel type to be used in the algorithm

Table 8.19: Parameters used to optimized SVM

8.5.2 Cross-validated results

The reported metrics for each of the optimized algorithms show improved results compared to the default parameters, reaching Accuracy / μF_1 -scores of over **0.85**, with Random Forests performing the best reaching **0.88**. After creating the optimized models, cross-validation is used again to get the Accuracy / μF_1 -scores after model tuning as shown in Figure 8.11.

For Random Forests, the variation in scores is reduced from **0.71-0.92** to **0.80-0.92**, while the median score increases from **0.83** to **0.85** and becoming the best performing model. The Gradient Boosted Decision Tree classifier also shows an improvement in the overall variation of the scores, changing from **0.71-0.90** to **0.75-0.90**, but does become the model with the largest variation of all optimized models. However, the median score increases from **0.78** to **0.86**, the highest of all optimized models. The Support Vector Machines scores increase from **0.75-0.87** to **0.77-0.90**, with the median score remaining the same at **0.81**.

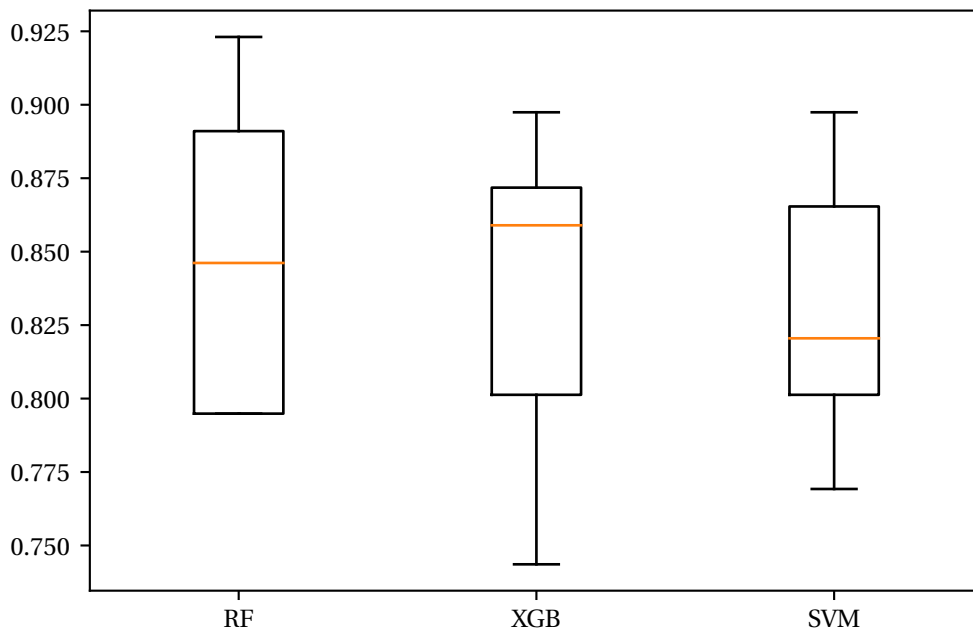


Figure 8.11: Cross-validated optimized model Accuracy / μF_1 -scores

8.6 Conclusion

Based on the user profile information collected as part of Data set 1 we identified a number of features that can indicate a user as being part of the Russian diaspora, the Ukrainian diaspora, or neither. These features consist of both direct profile information as provided by the user, as well as features based on profiles followed by that user. After an evaluation, it was determined that **22** of these features are suitable to train different supervised Machine Learning classifiers. To achieve this, a subset of **384** user profiles was tagged as belonging to either of the three aforementioned categories using a combination of manual profile evaluation as well as targeted searches across other social media. Three classifiers were compared, namely Random Forests, Gradient Boosted Decision Trees, and Support Vector Machines (Section 5.4). After evaluating these models using their default parameters, a combination of random search and exhaustive grid search was used to find an optimized set of hyper-parameters. This resulted in an overall increase in accuracy, with Random Forests being the best performing classifier with a median cross-validated accuracy of **0.85**.

Chapter 9

Diaspora interaction

This chapter combines the communities and their properties collected in Chapter 7 and the diaspora classification defined in Chapter 8, to find whether any meaningful relationship between them exists (RQ4). This is studied for Data set 1, which contains messages mentioning the Donbas region, as well as Data set 2, which contains targeted disinformation surrounding the shoot-down of flight MH17.

9.1 Distribution of predicted diaspora membership

Based on the result found in Chapter 8, the optimized Random Forests model is selected as the preferred method to predict diaspora membership.

9.1.1 Results for data set 1

For the initial prediction, all user profiles collected as part of Data set 1 that belong to one of the top 6 communities found in Section 7.4 are used, resulting in a total of **1299** user profiles. Figure 9.1 shows the overall distribution for each category after performing the prediction using the optimized Random Forests model. Here the assumed data imbalance between diaspora and non-diaspora users becomes visible, as the total predicted set consists of **25%** Russian diaspora users, **23%** Ukrainian diaspora users, and **52%** non-diaspora users.

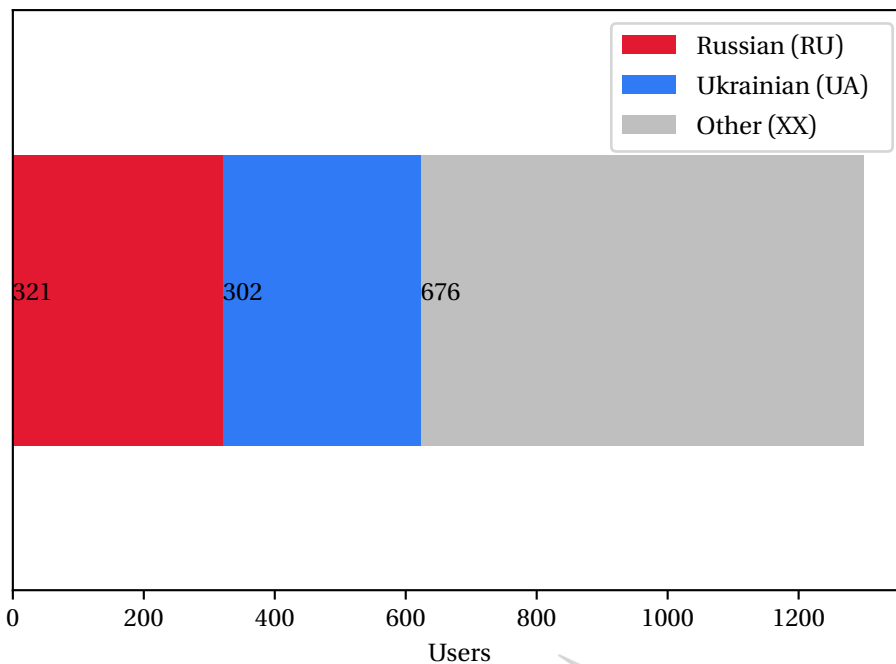


Figure 9.1: Predicted data distribution for Data set 1

9.1.1.1 Results per community

More in-depth insights can be obtained by further splitting the predicted diaspora membership for each community. Figure 9.2 shows the resulting distributions for each community. While each group consists of users from all three categories, the distribution does differ between communities.

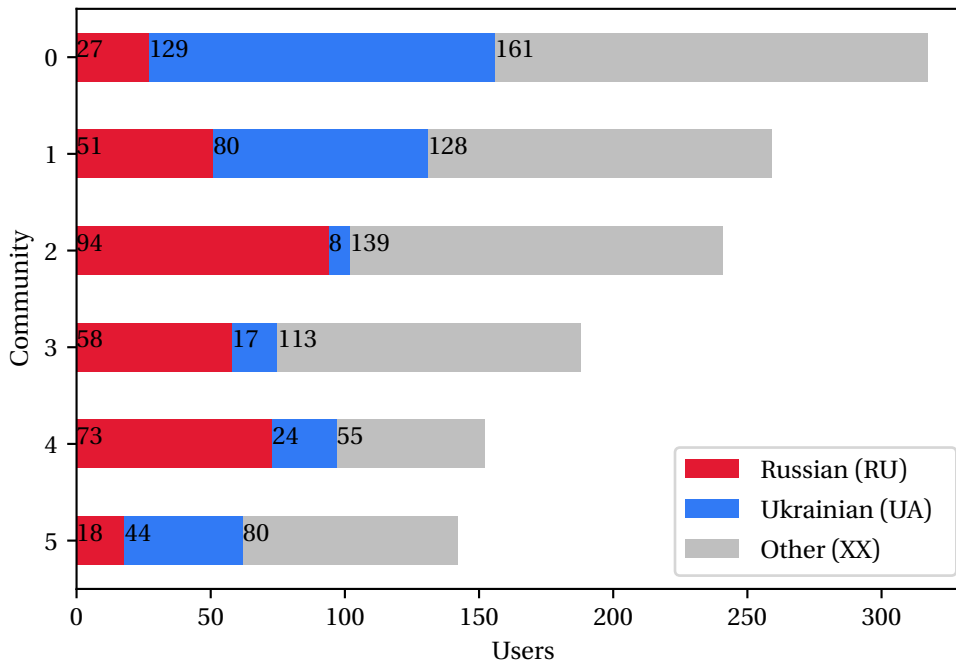


Figure 9.2: Predicted data distribution per community

Community 0, the largest, consists of **9%** Russian diaspora users, **41%** Ukrainian diaspora users and **50%** non-diaspora users. Overall the number of non-diaspora users matches with the overall data set. The Russian and Ukrainian groups show more skewed proportions, with the majority of the diaspora members being Ukrainian. This correlates with the results from Tables 7.3 and 7.9, which show a pro-Ukrainian stance.

Community 1 consists of **19%** Russian diaspora users, **29%** Ukrainian diaspora users and **52%** non-diaspora users. Again the group of non-diaspora users matches with the overall data set. This community once again has a larger group of Ukrainian diaspora members than Russian ones, albeit a weaker preference, matching the findings in Tables 7.4 and 7.10.

Community 2 consists of **39%** Russian diaspora users, **5%** Ukrainian diaspora users and **56%** non-diaspora users. This community has a larger percentage of non-diaspora users than the full data set. Furthermore, the users predicted as being diaspora members strongly skew towards Russian-diaspora users. The heavy skew towards Russian diaspora members matches the findings in Tables 7.5 and 7.11, which has a strong pro-Russian sentiment. Because the sentiment in this community is much stronger it provides an opportunity to perform a qualitative analysis of the non-diaspora members in it (Section 9.2).

Community 3 consists of **30%** Russian diaspora users, **9%** Ukrainian diaspora users and **61%** non-diaspora users. The distribution of this community is very similar to community 2, the same can be said about the findings in Tables 7.6 and 7.12, which again show a strong pro-Russian sentiment.

This group has the largest percentage of non-diaspora users.

Community 4 consists of **47%** Russian diaspora users, **16%** Ukrainian diaspora users and **37%** non-diaspora users. This is the only cluster where the number of non-diaspora users is lower than the users predicted as being diaspora members. It consists of the largest percentage of Russian diaspora members of any cluster. The findings in Tables 7.7 and 7.13 again show a strong pro-Russian sentiment.

Community 5, the smallest, consists of **14%** Russian diaspora users, **30%** Ukrainian diaspora users and **56%** non-diaspora users. This community has a larger percentage of non-diaspora members than the full data set. Of the diaspora members, the Ukrainian diaspora members are the largest group, showing a similar pattern as found in Tables 7.8 and 7.14.

Based on these results, Data set 1 shows a large involvement of diaspora members, with an overall composition of **25%** Russian diaspora users and **23%** Ukrainian diaspora users. Furthermore, the individual communities found as part of Chapter 7 show that the distribution of Ukrainian versus Russian diaspora members correlates with the community sentiments presented in Section 7.7.

9.1.2 Results for data set 2

Using the same trained model, the same prediction is made for Data set 2. In this case, the diaspora membership was predicted for the **1542** user profiles for which all features were available. In this case, the results show a very different distribution, with the number of diaspora members representing a much smaller part of all user profiles as shown in Figure 9.3. In this case, the Russian diaspora users make up **5%** of the total set, while Ukrainian diaspora users make up **2%**. The other **93%** is made up of non-diaspora users.

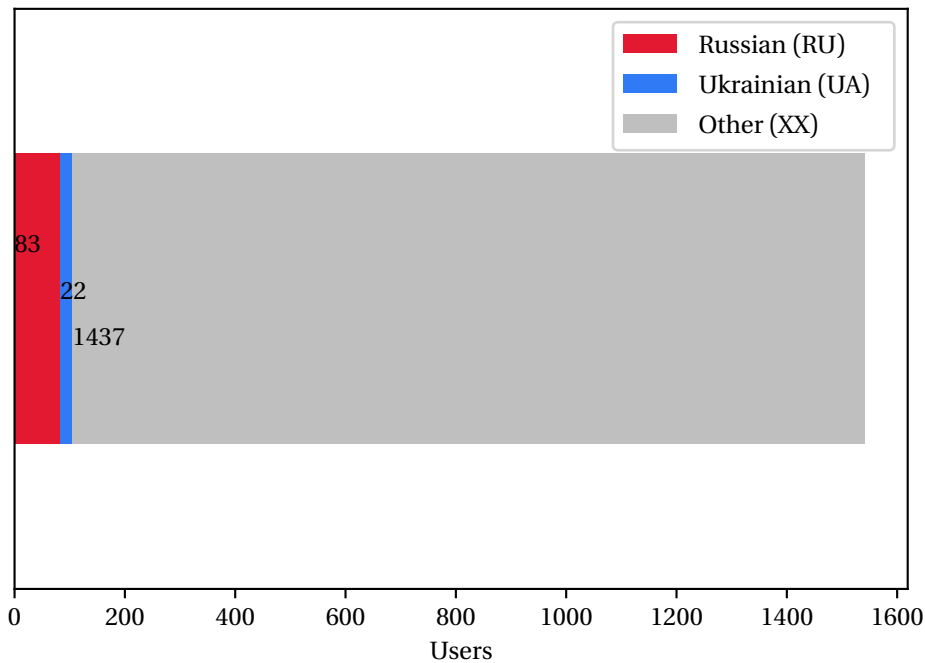


Figure 9.3: Predicted data distribution for Data set 2

9.1.2.1 Data cleaning

Data set 2 shows a very different make-up, with diaspora users only totaling to **7%** of the overall data set. This can partially be explained by the fact that the MH17 incident was widely reported across international media, which lead to a lot of speculation on the cause while the story still developed. Furthermore, no Russian or Ukrainian nationals were on board flight MH17 nor was either country the origin (The Netherlands) or destination (Malaysia) of the flight (Dutch Safety Board, 2015). Because these results disturb further analysis of the interaction with disinformation, additional data cleaning is therefore required filter out any neutral messages and keep only the messages explicitly sharing disinformation. Building on the results from earlier URL domain analysis (Section 7.6), a number of domains of interest are picked, consisting of state media (Table 6.1) and one widely shared alternative media source (*globalresearch.ca*). Oftentimes these messages contain the direct title of the linked article because a standard *Share on Twitter* functionality is used (Figure 6.2), this allows for selecting additional keywords to find messages making the same claim without a direct link to the source article. This leads to a subset of **568** Tweets and **466** users that explicitly spread disinformation.

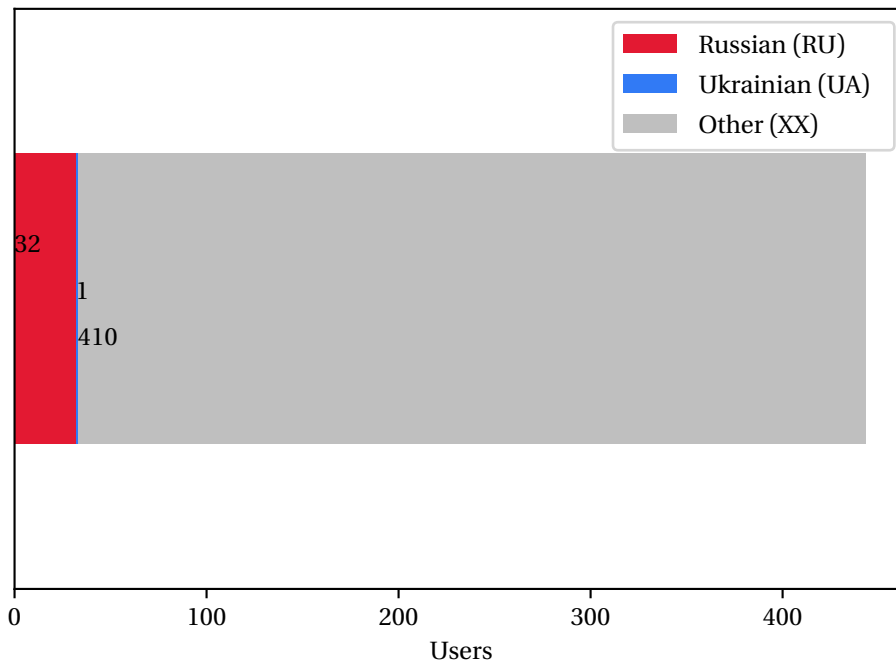


Figure 9.4: Predicted data distribution for Data set 2 after data cleaning

After data cleaning, the overall distribution of diaspora versus non-diaspora only shows a slight decrease in Ukrainian diaspora involvement as shown in Figure 9.4. However, due to the small amount of users considered, it is not possible to conclude whether this has any significant meaning. Based on these results it can be said that the overall interaction of diaspora users with this specific disinformation campaign about flight MH17 is similar to their interaction with the MH17 discourse as a whole, it being very limited compared to non-diaspora users.

9.2 Disinformation interaction

The results found in Section 9.1 show that there is a distinctive difference between the predicted diaspora membership for Data set 1 and Data set 2. Data set 1 which contains Tweets discussing the Russo-Ukrainian conflict in general, shows clear involvement of users predicted as diaspora members. For Data set 2, which contains Tweets referring to a targeted disinformation campaign surrounding the shoot-down of MH17, the set of users predicted as diaspora members is much smaller.

9.2.1 Results for data set 1

While the number of users identified as diaspora users (Ukrainian and Russian) aligns with the overall community sentiment (pro-Ukrainian, pro-Western) (Sections 7.7 and 9.1), not all communities are equal in terms of diaspora versus non-diaspora users. For Data set 1 each community shows a varying number of non-diaspora users, with the overall number ranging from **37%** to **61%** (Figure 9.2). It is expected that a large number of non-diaspora users take part in the discussion surrounding the Russo-Ukrainian conflict, due to the imbalanced nature of these groups on Twitter (Section 8.1). Community 0, 1, 2, and 5 all show values between **50%** and **56%**. Community 3 and 4, however, both show more extreme values with **61%** and **37%** respectively. When studying the messages and message sentiment in detail, no distinct differences can be identified between those communities and communities with a similar sentiment. One possible explanation is homophily among Twitter users (McPherson, Smith-Lovin, and Cook, 2001; Zamal, W. Liu, and Ruths, 2012; Asikainen et al., 2020), where similar people tend to associate. This in turn would cause them to interact more with such users and therefore influence the community detection performed on the mention network (Section 7.4).

9.2.2 Results for data set 2

Based on the predicted diaspora membership and detection methods used in the context of this research, Data set 2 shows very little interaction by users predicted as diaspora members. This would mean that neither Ukrainian nor Russian diaspora members play a significant role in spreading disinformation on English-language Twitter. Similar disinformation has been spread in the native language of the diaspora members (Khaldarova and Pantti, 2016) which raised the question whether such interaction did take place with messages in the users' native languages, which was not considered within the scope of this research.

9.3 Twitter trolls

Section 9.2 shows that the number of users predicted as diaspora members is much lower for the data set containing Tweets explicitly interacting with disinformation. One explanation not yet covered is that the users in this data set might be partially made up of Twitter trolls. These are profiles specifically set up for spreading disinformation in a coordinated way. These profiles can present themselves as concerned citizens (Xia et al., 2019), in this context often without any direct connection to Russia nor Ukraine, as well as media outlets or news aggregators (Doroshenko, 2021).

To gain a better understanding of the involvement of Twitter trolls, an existing data set of known troll accounts used by the *Internet Research Agency* (Section 3.3) published by the *US House of Representatives Intelligence Committee* is used¹. This is achieved by cross-referencing the published user names with the user names of users with that posted Tweets that are part of Data set 1 or Data set 2.

9.3.1 Results for data set 1

The data does not contain any messages by known troll accounts but does contain **20** messages mentioning a total of **6** known troll accounts in a total of **155444** tweets.

9.3.2 Results for data set 2

While Data set 2 contains messages explicitly interacting with disinformation, it does not contain any messages by known troll accounts nor any messages interacting with known troll accounts.

Based on the list of troll accounts uses, the involvement seems to be negligible. It is worth mentioning, however, that the list used was published as part of an investigation in the 2016 US Presidential Election, while this research focuses on the events surrounding the Russo-Ukrainian conflict in 2014 and 2015. A more extensive data set was published by Twitter², but is unusable because user identifiers in this set were anonymized.

¹<https://intelligence.house.gov/social-media-content/>

²<https://transparency.twitter.com/en/reports/information-operations.html>

9.4 Conclusion

Using the Random Forest classifier trained and optimized as part of Chapter 8, the diaspora membership of users in both Data set 1 and Data set 2 was predicted.

Based on these predictions, Data set 1 shows a large involvement of diaspora members, with **25%** Russian diaspora users and **23%** Ukrainian diaspora users. By further studying the individual communities found as part of Chapter 7, it is found that the distribution of Ukrainian versus Russian diaspora members correlates with the community sentiments presented in Section 7.7.

The same prediction shows very different results for Data set 2, with **5%** Russian diaspora users and **2%** Ukrainian diaspora users. To further validate these results, additional data cleaning was performed on the data set to make sure that only explicit disinformation was included (Section 9.1.2.1). However, after performing this data cleaning, the ratio of diaspora versus non-diaspora users remained the same. Further study of the data sets revealed a number of findings and possible explanations. Firstly, while the ratio of diaspora vs non-diaspora users fluctuates per community found in Data set 1, this does not seem to affect the message sentiment. One explanation here is homophily, the tendency of similar people to associate.

The limited involvement of diaspora users in Data set 2 can be explained by the fact that only English-language messages were considered, while disinformation targeted at these groups might be spread in their native language. Furthermore, the imbalance between data sets can be caused by the widespread coverage that the MH17 shoot-down received compared to the Russo-Ukrainian conflict as a whole.

Finally, the possibility of the involvement of known Twitter trolls was checked by cross-referencing the results in Data set 1 and Data set 2 against a list of known Twitter trolls. The results show that only six known troll accounts were found within the data sets (Section 9.3).

Chapter 10

Conclusions

This chapter presents the results of this research, reiterating each research question and detailing the key findings.



10.1 Primary research question

This research explored whether members of diaspora communities play a distinct role in the spread of known disinformation, related to and during conflict, explicitly focusing on the ongoing conflict between Russia and Ukraine. More specifically the period from begin 2014 until the end of 2015 was chosen, as this period saw the most active military conflict. In addition, social media was restricted to English-language Twitter due to the high availability of data compared to other social media platforms (Section 5.3).

The primary research question was therefore formulated as follows:

What is the interaction between users, considered to be diaspora members, and known disinformation campaigns on social media related to the Russo-Ukrainian conflict?

Answering the primary research question was achieved by splitting the problem into four research questions, the conclusions to each of which are listed below.

10.2 Disinformation campaigns

Because the primary premise of this research focuses on the interaction with disinformation campaigns, identifying and selecting a suitable campaign is essential. Within the scope of this research this is achieved by answering the following question:

Which known disinformation campaign surrounding the Russo-Ukrainian conflict is suitable to be analyzed for Diaspora interaction? (RQ1)

A common approach to identifying disinformation campaigns is by relying on fact-checking websites, made up of a panel of experts verifying claims made in news stories (Section 6.1). One fact-checker, *StopFake*, was selected as it specifically focuses on the Russo-Ukrainian conflict. Based on fact-checks extracted from this website within the time frame stated above, a number of source news articles was selected, each corresponding to a news story checked and marked as being false. Based on these stories searches were conducted on Twitter. The primary search was performed using the different variants of the article URL, the secondary search was then performed based on keywords extracted from the article title and Tweets from the first search (Section 6.2). Due to Twitter API limitations, not all data could be collected directly. A scraping tool was developed to collect data directly from the Twitter website (Appendix D.1).

To provide a good baseline on the discussion surrounding the Russo-Ukrainian conflict and aid further research, a primary data set was created containing all messages discussing the *Donbas* region, the region in which the conflict takes place (Data set 1). Based on the available data and existing proof, the claim that *flight MH17 was shot down by a Ukrainian fighter aircraft* was selected as a suitable disinformation campaign. A secondary data set was created containing all messages discussing this claim (Data set 2). The resulting data sets were used as the basis for answering subsequent research questions (Section 6.3).

10.3 Social media communities

Two approaches were taken to identify communities within the gathered data. The first approach focuses on communities that can be extracted directly from the data set based on user behavior, and therefore answering the following question:

Which Twitter user communities exist interacting with the selected disinformation campaign surrounding the Russo-Ukrainian conflict? (RQ2).

Interactions were studied in the form of mention networks, a directed graph that treats each user as a vertex and each time a user mentions another user as an edge between those two users (Section 7.1). On Twitter a user mentions another user by including their username in a message, this is especially useful as it can indicate both the target (message directed at the mentioned user) as well as the source (message retweeted from the mentioned user).

An initial analysis was performed by using the Gephi graph visualization and analysis tool to render the mention network for each of the data sets. The out-degree (number of mentions by a user) was chosen as an indication of activity in this visualization, expressed by the size of the vertices, because it represents initiative from that user (Section 7.2).

Due to the large size of the resulting networks, a number of steps were taken to reduce noise and therefore better study the most active users in the network. For Data set 1, the largest of both, this approach meant removing any users with an in-degree or out-degree less than 5, after which only the primary (weakly-connected) component was retained. For Data set 2 only the second step was required (Section 7.3).

By using a force-directed graph visualization algorithm, clusters of vertices could already be identified. The small size of Data set 2 allowed for manual analysis, showing that vertex clusters mostly indicated single users sending the same message many times (while mentioning different people) or users using a Twitter share functionality to share a social media post or news article that automatically mentions the Twitter account of the source. Data set 1 required additional steps to identify communities, this was achieved by using the Leiden community detection algorithm (Traag, Waltman, and Eck, 2019). Out of the 16 resulting communities, the six largest communities were selected for detailed analysis, because they contained more than 100 users each, a number that significantly dropped for any following communities (Section 7.4).

To better understand the sentiment of each community, two approaches were taken. Firstly the textual content of each message was studied for its textual properties, expressed as the top 20 most common bigrams for each community found (Section 7.5). The resulting data showed clear differentiation between communities in terms of word-choices and sentiment.

The second approach looked at URLs shared. This required additional data collection, as usage of URL-shorteners is standard practice on Twitter, making it harder to study the actual source of a shared URL (Section 7.6). Furthermore, many of the shared URLs referred to deleted social media posts. Still a similar sentiment, although less verifiable, could be seen as was produced by the

text-analysis (Section 7.6).

Based on the sentiment identified using textual analysis (Section 7.5) and similar results found using (Section 7.6) each community was labeled as either pro-Western (including pro-Ukrainian) or pro-Russian (Section 7.7).

10.4 Diaspora membership

The second approach focuses on identifying communities based on diaspora membership using machine learning of user profiling and classification. To do so, the following question was answered:

Which machine learning methods are suitable for predicting diaspora membership from Twitter user profiles? (RQ3)

Because both data sets only contain messages collected using keyword searches, they are not guaranteed to be indicative of a user's overall behavior and message content. It was considered unfeasible to collect all messages from each user in the data set in the selected time frame. Because of this, only static profile information was considered for user profiling. This resulted in a list of features extracted from the username, full name, location, and profile description (Section 8.2).

Further analysis of each feature was performed to ensure data quality, in this process *naming convention* and *location*-based features were excluded from the final model. The resulting features were calculated for each user profile. In addition, because people tend to interact with similarly minded people, the same features were also calculated for each profile followed by the user to uncover any latent profile attributes (Section 8.3).

A tagged data set was created from a subset of Data set 1 was created to aid in training supervised machine learning methods (Section 8.1).

A selection of three supervised machine learning methods was made, namely Random Forest, Gradient Boosted Decision Trees, and Support Vector Machines (Section 5.4). For each method, the performance in predicting diaspora membership was compared. Cross-validation was used to take overfitting into account. Each of the models showed an accuracy of over **0.71** per fold (Section 8.4).

To further optimize each of the models, hyper-parameter tuning was employed with a combination of random and exhaustive grid searches. We presented the improved parameters and their results, which increased the accuracy for each model to over **0.75** per fold. Random Forest was selected as the best performing model with a median accuracy of **0.85** across folds (Section 8.5).

10.5 Diaspora interaction

By combining the result of both approaches, a final question is answered:

How does diaspora membership relate to detected Twitter user communities? (RQ4)

Using the Random Forest classifier trained and optimized as part of, the diaspora membership of users in both Data set 1 and Data set 2 was predicted. For Data set 1 a correlation between the ratio of Ukrainian versus Russian diaspora members and the sentiment of each community was identified. Furthermore, the prediction showed that the involvement of users classified as diaspora members was significantly lower for Data set 2. Additional data cleaning was performed, but lead to similar results (Section 9.1).

These results were then studied in detail to explore possible explanations for the large difference in the involvement of diaspora members between both data sets. Two possible causes were identified. Firstly, Data set 1 is made up of tweets containing the term *Donbas*, the region in which there was an active military conflict during 2014 and 2015 as part of the overall Russo-Ukrainian conflict. This term might be more commonly used by people familiar with the region. Secondly, disinformation like the messages contained in Data set 2 is spread in multiple languages, including the ones in the native language of the diaspora members, while this research only considered English language messages (Section 9.2).

To better understand the users that were involved in spreading the messages in Data set 2, the possibility of the involvement of troll accounts was studied. However, no traces of such known accounts were found in this data set (Section 9.3).

10.6 Conclusion

What is the interaction between users, considered to be diaspora members, and known disinformation campaigns on social media related to the Russo-Ukrainian conflict?

Based on the results discussed above, it can be concluded that both groups of diaspora members have a clear role in the discourse surrounding the Russo-Ukrainian conflict as a whole, but that a distinct role in spreading disinformation could not be identified using the data and methods used in this research.

Chapter 11

Discussion

This chapter presents the key findings and implications of this research. Additionally, it explores limitations and recommendations for future research.



11.1 Key findings

The following key findings can be derived from the results presented in Chapter 10:

- Fact-checkers can be used to find news articles where disinformation was first spread. However, due to the usage of URL shorteners, it is hard to directly gather data from Twitter for any references to these articles. Further keyword searches based on article titles and claims had to be performed.
- Mention networks provide a useful way to identify communities within social media discourse and study the further sentiment in each of these communities. However, this method does not provide very useful insights in smaller data sets like Data set 2.
- Random Forests perform well for classifying diaspora membership based on static user profile data and latent properties based on static user profile data of followed accounts.
- Diaspora members play a clear and distinct role in the general discourse surrounding the Russo-Ukrainian conflict, but do not play an active role in the spread of disinformation on English-language Twitter, given the data sets available for this research.

11.2 Limitations

11.2.1 Data availability

There is a large body of research focusing on disinformation, both in terms of detecting it as well as how it spreads (Section 4.1). Most existing methods of disinformation detection rely on existing manual fact-checkers or tagged data sets for training and verification purposes. The lack of existing data sets containing Twitter data related to the Russo-Ukrainian conflict in the period from the start of 2014 until the end of 2015 meant that it was necessary to first create such a data set (Section 4.2). Over the years multiple social media platforms have limited access to data that was previously publicly available (Walker, Mercea, and Bastos, 2019). In addition, a number of features available through Twitter's enterprise API offering were not available. This particularly posed a problem for URL-shortening as many URL-shortening services used in the data collected have since shut down, making it impossible to find the originally linked URL (Section 6.2). Twitter does store this original URL, but it can not be retrieved using the regular API. A similar limitation can be seen with links to social media posts, which in many cases were already deleted, making it impossible to verify whether it contained disinformation (Section 7.6). Studying more recent events might therefore lead to more precise results.

11.2.2 Data collection

Due to the API limitations mentioned above, alternative data gathering methods (scraping) had to be used, this slowed down the data collection speed and resulted in less reliable data overall as multiple adjustments were needed during the data collection phase reflecting changes made by

Twitter.

Furthermore, the collection method used relied on a combination of fact-checked articles and tweets containing links to such articles. This approach was severely limited by the data availability issues and therefore lead to smaller disinformation data sets than expected in the end resulting in only a single, relatively small, data set containing known disinformation (Data set 2).

Homophily in communities has been proven a powerful tool to discover latent properties when performing user profiling (Zamal, W. Liu, and Ruths, 2012). To leverage these latent properties, half of the features selected for diaspora detection rely on features derived from the friends of a user (i.e. accounts that a user follows). This requirement significantly increased the data collection time as profile data had to be collected for each user followed by a user appearing in the data sets. This, again, limited the possibility for collecting larger data sets within the time frame of this research.

11.2.3 Mention networks

Mention networks have proven to be a valuable tool to detect communities and sentiment in large data sets of social media messages Helmus (2018), and therefore provided valuable insights (Section 7.7) when applied to the data set containing messages discussing the Russo-Ukrainian conflict in general (Data set 1). The data set containing messages spreading known disinformation (Data set 2), however, did not result in any additional insights due to its small size.

11.2.4 Diaspora classification

Because the data was collected using targeted keyword searches (Section 6.2), the number of messages per user was relatively low. This rules out any user profiling methods based on behavioral or linguistic properties, as this would require a larger set of tweets per user. For this reason, the features presented in Section 8.2 all rely on limited static profile information.

An attempt was made to use third-party services to parse the user-provided location field (Section 8.2). However, due to the unstructured and often made-up nature of the data, this led to low-quality results deemed unfit to use for training the machine learning model (Section 8.3). By definition, diaspora members can only be identified as such if they do not live in their country of origin (Section 1.3), excluding the location field might therefore have led to incorrectly classifying people living in their country of origin as diaspora members. While this does not invalidate results presented in Section 9.2, as the number of diaspora members interacting with known disinformation was low in general, it is worth noting.

11.3 Recommendations

Based on the key findings presented in Section 11.1 the following recommendations are suggested:

11.3.1 Limitations to historical data

Most major social media platforms offer some form of data access through APIs. However, the trend is clearly going towards more limits on such data being imposed by those platforms. This can cause a big hurdle for future academic research about the spread of disinformation (Walker, Mercea, and Bastos, 2019). While access to additional data can be requested, it makes the social media platform itself the gatekeeper, which might prevent any research critical of the role these platforms play. Developing alternative means of collecting and archiving such data, like the approach to scraping used in this research or data available through the internet archive¹, can help mitigate this. Such alternative data sets also allow for the retrieval of historical data like removed posts or the unwound versions of shortened URLs.

The end of 2021 and beginning of 2022 have seen renewed tensions surrounding the Russo-Ukrainian conflict, with large scale Russian military exercises taking place near the Ukrainian border, new demands and threats being issues, and high level talks between the United States and the Russian Federation being unfruitful. This situation opens up the possibility conduct a similar research on recent data, which is not affected by many of the limitation stated above.

11.3.2 Mention networks and disinformation

Mention networks have proven to be a useful tool for uncovering relationships between users and communities formed by those users. The data set consisting of messages spreading disinformation was too small to extract meaningful information from the mention network. However, the manual study of the behavior in the network did uncover an interesting pattern that could benefit from further research: in many cases where disinformation was spread, users did not mention people in their own network but often directed the messages at (Western-aligned) news outlets and organizations.

11.3.3 User profiling and classification

In this scope of this research, the country of origin was an important discriminator. However, similar methods can be applied to certain minorities or other groups at risk. While it would have been beneficial for the outcome of this research and provide useful insights, we do not specifically recommend further research into the profiling of social media users without a strong ethical framework.

11.3.4 The role of diasporas during conflict

The limited scope of this research has shown that Russian and Ukrainian diaspora members play no distinct role in the spread of disinformation on English language Twitter. Such a narrow definition naturally excludes other cases. Therefore more research is required about the general role that diaspora members play during a conflict. A research direction similar to the one chosen in this research would be the role diaspora members play in spreading disinformation in their native

¹<https://archive.org/>

tongue. However, other research directions should also be considered: political influence exercised pressure in their country of residence, charity organizations set up by diasporas to support their country of origin, or diaspora members being targeted specifically to alter their voting behavior and sway election results in their country of origin.

Bibliography

- Alpaydin, Ethem (2010). *Introduction to machine learning*. en. 2nd ed. Adaptive computation and machine learning. Cambridge, Mass: MIT Press. ISBN: 978-0-262-01243-0 (cit. on p. 78).
- Aly, Mohamed (Nov. 2005). “Survey on Multiclass Classification Methods”. en. In: *Technical Report, Caltech*, p. 9 (cit. on p. 28).
- Asikainen, Aili et al. (May 2020). “Cumulative effects of triadic closure and homophily in social networks”. en. In: *Science Advances* 6.19, eaax7310. ISSN: 2375-2548. DOI: 10.1126/sciadv.aax7310. URL: <https://www.science.org/doi/10.1126/sciadv.aax7310> (visited on 01/09/2022) (cit. on p. 94).
- Bessudnov, Alexey et al. (Oct. 2021). *Predicting ethnicity with data on personal names in Russia*. en. preprint. SocArXiv. URL: <https://osf.io/wf6p4> (visited on 12/08/2021) (cit. on p. 18).
- Bevendorff, Janek et al. (2021). “Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter, and Style Change Detection”. en. In: p. 12 (cit. on p. 19).
- Blondel, Vincent D. et al. (Oct. 2008). “Fast unfolding of communities in large networks”. en. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/P10008. URL: <http://arxiv.org/abs/0803.0476> (visited on 06/05/2021) (cit. on p. 47).
- Carsenat, Elian (Sept. 2013). “Onomastics for Business: can discrimination help development?” en. In: *ParisTech Review*, p. 7 (cit. on p. 65).
- Caruana, Rich and Alexandru Niculescu-Mizil (2006). “An empirical comparison of supervised learning algorithms”. en. In: *Proceedings of the 23rd international conference on Machine learning - ICML '06*. Pittsburgh, Pennsylvania: ACM Press, pp. 161–168. ISBN: 978-1-59593-383-6. DOI: 10.1145/1143844.1143865. URL: <http://portal.acm.org/citation.cfm?doid=1143844.1143865> (visited on 03/01/2021) (cit. on pp. 24, 74).
- Chindea, Alin (2008a). *Migration in the Russian Federation: a country profile*. en. Geneva: International Organization for Migration. ISBN: 978-92-9068-483-1 (cit. on pp. 2, 61).
- (2008b). *Migration in Ukraine: a country profile*. en. Geneva: International Organization for Migration. ISBN: 978-92-9068-486-2 (cit. on p. 2).
- Clauset, Aaron, M. E. J. Newman, and Cristopher Moore (Dec. 2004). “Finding community structure in very large networks”. en. In: *Physical Review E* 70.6, p. 066111. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.70.066111. URL: <https://link.aps.org/doi/10.1103/PhysRevE.70.066111> (visited on 01/03/2021) (cit. on pp. 15, 23, 47).

- Conover, Michael D. et al. (Oct. 2011). "Predicting the Political Alignment of Twitter Users". en. In: *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*. Boston, MA, USA: IEEE, pp. 192–199. ISBN: 978-1-4577-1931-8. DOI: 10.1109/PASSAT/SocialCom.2011.34. URL: <http://ieeexplore.ieee.org/document/6113114/> (visited on 02/27/2021) (cit. on p. 19).
- Division, United Nations Population (2020). *International Migrant Stock | Population Division*. URL: <https://www.un.org/development/desa/pd/content/international-migrant-stock> (visited on 01/19/2022) (cit. on pp. 2, 61).
- Doroshenko, Larissa (2021). "Trollfare: Russia's Disinformation Campaign During Military Conflict in Ukraine". en. In: *International Journal of Communication* 15, p. 28 (cit. on p. 95).
- Dutch Safety Board (Oct. 2015). *MH17 Passenger information*. en. Tech. rep. (cit. on p. 92).
- Fischer, Sabine (2019). *The Donbas conflict: opposing interests and narratives, difficult peace process*. en. Tech. rep. Publisher: German Institute for International and Security Affairs Version Number: 1. Stiftung Wissenschaft Und Politik. URL: <https://www.swp-berlin.org/10.18449/2019RP05/> (visited on 01/22/2022) (cit. on p. 2).
- Galeotti, Mark (July 2014). *The 'Gerasimov Doctrine' and Russian Non-Linear War*. en. URL: <https://inmoscowsshadows.wordpress.com/2014/07/06/the-gerasimov-doctrine-and-russian-non-linear-war/> (visited on 01/10/2021) (cit. on p. 10).
- (May 2019). "The mythical 'Gerasimov Doctrine' and the language of threat". en. In: *Critical Studies on Security* 7.2, pp. 157–161. ISSN: 2162-4887, 2162-4909. DOI: 10.1080/21624887.2018.1441623. URL: <https://www.tandfonline.com/doi/full/10.1080/21624887.2018.1441623> (visited on 01/10/2021) (cit. on p. 10).
- Giles, Keir (2015). "Russia and Its Neighbours: Old Attitudes, New Capabilities". en. In: *NATO CCD COE Publications*, p. 12 (cit. on pp. 8, 10).
- Grant, Thomas D. (Jan. 2015). "Annexation of Crimea". en. In: *American Journal of International Law* 109.1, pp. 68–95. ISSN: 0002-9300, 2161-7953. DOI: 10.5305/amerjintlaw.109.1.0068. URL: https://www.cambridge.org/core/product/identifier/S0002930000001925/type/journal_article (visited on 12/28/2020) (cit. on p. 8).
- Helmus, Todd C. (2018). *Russian social media influence: understanding Russian propaganda in Eastern Europe*. en. Research report (Rand Corporation) RR-2237-OSD. Santa Monica, Calif: RAND Corporation. ISBN: 978-0-8330-9957-0 (cit. on pp. 15, 23, 40, 47, 104).
- Higgins, Eliot (Jan. 2015). *SU-25, MH17 and the Problems with Keeping a Story Straight*. en-GB. URL: <https://www.bellingcat.com/news/uk-and-europe/2015/01/10/su-25-mh17-and-the-problems-with-keeping-a-story-straight/> (visited on 12/03/2021) (cit. on p. 37).
- Horne, Benjamin D., Jeppe Nørregaard, and Sibel Adalı (July 2019). "Different Spirals of Sameness: A Study of Content Sharing in Mainstream and Alternative Media". en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 13, pp. 257–266. ISSN: 2334-0770. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/3227> (visited on 01/03/2021) (cit. on p. 15).
- Jacomy, Mathieu et al. (June 2014). "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software". en. In: *PLoS ONE* 9.6. Ed. by Mark R.

- Muldoon, e98679. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0098679. URL: <https://dx.plos.org/10.1371/journal.pone.0098679> (visited on 07/04/2021) (cit. on p. 41).
- Jin, Zhiwei et al. (2016). “News Verification by Exploiting Conflicting Social Viewpoints in Microblogs”. en. In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)* 13, p. 7 (cit. on p. 15).
- Kergl, Dennis, Robert Roedler, and Sebastian Seeber (Aug. 2014). “On the endogenesis of Twitter’s Spritzer and Gardenhose sample streams”. en. In: *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. Beijing: IEEE, pp. 357–364. ISBN: 978-1-4799-5877-1. DOI: 10.1109/ASONAM.2014.6921610. URL: <https://ieeexplore.ieee.org/document/6921610/> (visited on 01/10/2021) (cit. on p. 17).
- Khaldarova, Irina and Mervi Pantti (Oct. 2016). “Fake News: The narrative battle over the Ukrainian conflict”. en. In: *Journalism Practice* 10.7, pp. 891–901. ISSN: 1751-2786, 1751-2794. DOI: 10.1080/17512786.2016.1163237. URL: <https://www.tandfonline.com/doi/full/10.1080/17512786.2016.1163237> (visited on 01/31/2021) (cit. on pp. 2, 8, 17, 94).
- Khatua, Aparup, Apalak Khatua, and Erik Cambria (Dec. 2020). “Predicting political sentiments of voters from Twitter in multi-party contexts”. en. In: *Applied Soft Computing* 97, p. 106743. ISSN: 15684946. DOI: 10.1016/j.asoc.2020.106743. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1568494620306815> (visited on 01/22/2022) (cit. on p. 19).
- Leicht, E. A. and M. E. J. Newman (Mar. 2008). “Community structure in directed networks”. en. In: *Physical Review Letters* 100.11, p. 118703. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.100.118703. URL: <http://arxiv.org/abs/0709.4500> (visited on 01/03/2021) (cit. on p. 15).
- McPherson, Miller, Lynn Smith-Lovin, and James M Cook (Aug. 2001). “Birds of a Feather: Homophily in Social Networks”. en. In: *Annual Review of Sociology* 27.1, pp. 415–444. ISSN: 0360-0572, 1545-2115. DOI: 10.1146/annurev.soc.27.1.415. URL: <https://www.annualreviews.org/doi/10.1146/annurev.soc.27.1.415> (visited on 01/09/2022) (cit. on p. 94).
- Mejias, Ulises A and Nikolai E Vokuev (Oct. 2017). “Disinformation and the media: the case of Russia and Ukraine”. en. In: *Media, Culture & Society* 39.7, pp. 1027–1042. ISSN: 0163-4437, 1460-3675. DOI: 10.1177/0163443716686672. URL: <http://journals.sagepub.com/doi/10.1177/0163443716686672> (visited on 12/20/2020) (cit. on p. 10).
- Ministerie van Justitie en Veiligheid (Sept. 2016). *JIT presentation of first results of the MH17 criminal investigation (28-09-2016) - MH17 plane crash - Public Prosecution Service*. en-GB. URL: <https://www.prosecutionservice.nl/topics/mh17-plane-crash/criminal-investigation-jit-mh17/jit-presentation-first-results-mh17-criminal-investigation-28-9-2016> (visited on 12/03/2021) (cit. on p. 37).
- Ministry of Foreign Affairs of Ukraine (Dec. 2019). *Ministry of Foreign Affairs of Ukraine - Ukrainians worldwide*. en. URL: <https://mfa.gov.ua/en/about-ukraine/ukrainians-worldwide> (visited on 01/19/2022) (cit. on pp. 2, 61).
- Mitra, Tanushree and Eric Gilbert (Apr. 2015). “CREDBANK: A Large-Scale Social Media Corpus with Associated Credibility Annotations”. en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 9.1, p. 10 (cit. on pp. 16, 17, 23).

- Peffer, Ken et al. (Dec. 2007). "A Design Science Research Methodology for Information Systems Research". en. In: *Journal of Management Information Systems* 24.3, pp. 45–77. ISSN: 0742-1222, 1557-928X. DOI: 10.2753/MIS0742-1222240302. URL: <https://www.tandfonline.com/doi/full/10.2753/MIS0742-1222240302> (visited on 12/28/2020) (cit. on p. 22).
- Pennacchiotti, Marco and Ana-Maria Popescu (2011). "A Machine Learning Approach to Twitter User Classification". en. In: *Proceedings of the International AAI Conference on Web and Social Media* 5.1, p. 8 (cit. on pp. 18, 19, 24, 63, 74).
- Rangel, Francisco, Anastasia Giachanou, et al. (2020). "Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter". en. In: p. 18 (cit. on p. 19).
- Rangel, Francisco and Paolo Rosso (2019). "Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling in Twitter". In: URL: https://pan.webis.de/downloads/publications/papers/rangel_2019.pdf (visited on 01/23/2022) (cit. on p. 19).
- Richey, Mason (Mar. 2018). "Contemporary Russian revisionism: understanding the Kremlin's hybrid warfare and the strategic and tactical deployment of disinformation". en. In: *Asia Europe Journal* 16.1, pp. 101–113. ISSN: 1612-1031. DOI: 10.1007/s10308-017-0482-5. URL: <https://doi.org/10.1007/s10308-017-0482-5> (visited on 12/05/2020) (cit. on pp. 2, 8, 10).
- Roman, Nataliya, Wayne Wanta, and Iuliia Buniak (June 2017). "Information wars: Eastern Ukraine military conflict coverage in the Russian, Ukrainian and U.S. newscasts". en. In: *International Communication Gazette* 79.4, pp. 357–378. ISSN: 1748-0485, 1748-0493. DOI: 10.1177/1748048516682138. URL: <http://journals.sagepub.com/doi/10.1177/1748048516682138> (visited on 12/27/2020) (cit. on p. 10).
- Santia, Giovanni C and Jake Ryland Williams (2018). "BuzzFace: A News Veracity Dataset with Facebook User Commentary and Egos". en. In: *Proceedings of the Twelfth International AAI Conference on Web and Social Media (ICWSM 2018)*, p. 10 (cit. on p. 16).
- Sarker, Iqbal H. (May 2021). "Machine Learning: Algorithms, Real-World Applications and Research Directions". en. In: *SN Computer Science* 2.3, p. 160. ISSN: 2662-995X, 2661-8907. DOI: 10.1007/s42979-021-00592-x. URL: <https://link.springer.com/10.1007/s42979-021-00592-x> (visited on 12/19/2021) (cit. on pp. 27, 28).
- Shu, Kai, Deepak Mahudeswaran, et al. (2018). "FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media". en. In: *ArXiv abs/1809.01286*, p. 12 (cit. on pp. 3, 16, 17).
- Shu, Kai, Suhang Wang, and Huan Liu (Apr. 2018). "Understanding User Profiles on Social Media for Fake News Detection". en. In: *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. Miami, FL: IEEE, pp. 430–435. ISBN: 978-1-5386-1857-8. DOI: 10.1109/MIPR.2018.00092. URL: <https://ieeexplore.ieee.org/document/8397048/> (visited on 01/22/2022) (cit. on p. 19).
- Sifuentes, Jesse (Nov. 2019). *The Propaganda of Octavian and Mark Antony's Civil War*. en. URL: <https://www.worldhistory.org/article/1474/the-propaganda-of-octavian-and-mark-antonys-civil/> (visited on 06/12/2021) (cit. on p. 9).

- Silverman, Craig et al. (Oct. 2016). *Hyperpartisan Facebook Pages Are Publishing False And Misleading Information At An Alarming Rate*. en. URL: <https://www.buzzfeednews.com/article/craigsilverman/partisan-fb-pages-analysis> (visited on 01/02/2021) (cit. on p. 16).
- Sokolova, Marina and Guy Lapalme (July 2009). “A systematic analysis of performance measures for classification tasks”. en. In: *Information Processing & Management* 45.4, pp. 427–437. ISSN: 03064573. DOI: 10.1016/j.ipm.2009.03.002. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0306457309000259> (visited on 11/06/2021) (cit. on p. 73).
- Song, Xuemeng et al. (Aug. 2015). “Multiple Social Network Learning and Its Application in Volunteerism Tendency Prediction”. en. In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Santiago Chile: ACM, pp. 213–222. ISBN: 978-1-4503-3621-5. DOI: 10.1145/2766462.2767726. URL: <https://dl.acm.org/doi/10.1145/2766462.2767726> (visited on 01/23/2022) (cit. on p. 19).
- Starbird, Kate, Ahmer Arif, and Tom Wilson (Nov. 2019). “Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations”. en. In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW, pp. 1–26. ISSN: 2573-0142, 2573-0142. DOI: 10.1145/3359229. URL: <https://dl.acm.org/doi/10.1145/3359229> (visited on 02/06/2021) (cit. on p. 4).
- Traag, Vincent, Ludo Waltman, and Nees Jan van Eck (Dec. 2019). “From Louvain to Leiden: guaranteeing well-connected communities”. en. In: *Scientific Reports* 9.1, p. 5233. ISSN: 2045-2322. DOI: 10.1038/s41598-019-41695-z. URL: <http://arxiv.org/abs/1810.08473> (visited on 06/05/2021) (cit. on pp. 47, 58, 99).
- United Nations Educational, Scientific and Cultural Organization (2018). *World trends in freedom of expression and media development*. en. United Nations Educational, Scientific and Cultural Organization. ISBN: 978-92-3-100242-7 (cit. on p. 10).
- Walker, Shawn, Dan Mercea, and Marco Bastos (Sept. 2019). “The disinformation landscape and the lockdown of social platforms”. en. In: *Information, Communication & Society* 22.11, pp. 1531–1543. ISSN: 1369-118X, 1468-4462. DOI: 10.1080/1369118X.2019.1648536. URL: <https://www.tandfonline.com/doi/full/10.1080/1369118X.2019.1648536> (visited on 11/28/2021) (cit. on pp. 26, 103, 105, 125).
- Wang, William Yang (June 2017). “”Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection”. en. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* 2, pp. 422–426. DOI: 10.18653/v1/P17-2067. URL: <http://arxiv.org/abs/1705.00648> (visited on 11/28/2020) (cit. on pp. 16, 23).
- Wardle, Claire and Hossein Derakhshan (Oct. 2017). *Information Disorder: Toward an interdisciplinary framework for research and policy making*. Tech. rep. Council of Europe (cit. on pp. 3, 4).
- Wong, Kai On et al. (Nov. 2020). “A machine learning approach to predict ethnicity using personal name and census location in Canada”. en. In: *PLOS ONE* 15.11. Ed. by Sreeram V. Ramagopalan, e0241239. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0241239. URL: <https://dx.plos.org/10.1371/journal.pone.0241239> (visited on 01/17/2021) (cit. on p. 18).

- Xia, Yiping et al. (Sept. 2019). “Disinformation, performed: self-presentation of a Russian IRA account on Twitter”. en. In: *Information, Communication & Society* 22.11, pp. 1646–1664. ISSN: 1369-118X, 1468-4462. DOI: 10.1080/1369118X.2019.1621921. URL: <https://www.tandfonline.com/doi/full/10.1080/1369118X.2019.1621921> (visited on 11/28/2021) (cit. on p. 95).
- Zamal, Faiyaz Al, Wendy Liu, and Derek Ruths (2012). “Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors”. en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 6.1, p. 4 (cit. on pp. 19, 20, 69, 94, 104).
- Zhou, Xinyi and Reza Zafarani (Feb. 2019). “Network-based Fake News Detection: A Pattern-driven Approach”. en. In: *ACM SIGKDD Explorations Newsletter* 21.1, p. 13 (cit. on p. 15).
- (Oct. 2020). “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities”. en. In: *ACM Computing Surveys* 53.5, pp. 1–40. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3395046. URL: <http://arxiv.org/abs/1812.00315> (visited on 11/28/2020) (cit. on pp. 13–15, 18).

Appendices



Appendix A

StopFake articles

Date	Title	State media source
02-03-2014	The mass escape of the military forces on the side of the “Crimean government	ria.ru
03-03-2014	Fighting in Sevastopol	www.vesti.ru
08-03-2014	Ukrainian state agencies were ordered to delete Russian versions of their official websites.	itar-tass.com
10-03-2014	Ukrainian Armed Forces soldiers are not receiving salary for more than a month.	ria.ru
29-03-2014	American mercenaries have appeared in Donetsk	russian.rt.com
01-04-2014	Crimean Tatar autonomy is not a territorial autonomy.	www.vesti.ru
04-04-2014	Only Six Citizens of Crimea Chose to Keep the Ukrainian Citizenship	ria.ru
26-04-2014	Ukrainian fighting vehicle entering Donetsk with a swastika	ria.ru
10-05-2014	Donetsk militants seized two Ukrainian rocket launchers “GRAD”	ria.ru
20-05-2014	Suicide of Odessa Schoolgirl is Connected with Events of May 2	www.vesti.ru
21-05-2014	Executive Editor of The New York Times was let go due to an article on Slovyansk	www.vesti.ru
25-05-2014	System “Election” has been disabled	ria.ru
28-05-2014	Ukrainian General Kneels before the Ex-Ambassador of the US	www.vesti.ru
24-06-2014	Unwilling to Fight Ukrainian Paratroopers Resign from Service	itar-tass.com
27-06-2014	Maidan supporters in Toronto have brought children an expletively-inscribed cake	www.vesti.ru

15-07-2014	Crucifixion in Slovyansk	www.1tv.ru
20-07-2014	Ukrainian army transported SAM “Buk” with one missing rocket on the territory controlled by Ukraine	russian.rt.com
29-07-2014	The Netherlands Accused Ukraine of Lying about Downed Boeing-777	www.vesti.ru
06-08-2014	hundreds of Ukrainian soldiers deserted to Russia and asked for asylum there	www.vesti.ru
28-08-2014	American Farmer Supported Russia’s Response to the EU Sanctions	www.vesti.ru
28-08-2014	The Russian army did not Invade Ukraine	ria.ru
05-09-2014	Children’s Combat Battalion was Formed in Prykarpattia	www.vesti.ru
16-09-2014	German Tanks have Crossed the Ukrainian Border and are on the March to the East	www.vesti.ru
24-10-2014	Fakes in Week News issue with Dmitry Kiselev	russia.tv
01-11-2014	Bodies of 286 Women were Found near Krasnoarmiisk	ria.ru
03-11-2014	Ukrainian Militaries are Promised “a Parcel of Land and Two Slaves”	www.1tv.ru
30-11-2014	Faked Video of Homosexuality Propaganda among Children on TV Channel Russia 1	russia.tv

Table A.1: StopFake articles

Appendix B

Unwound URLs

Twitter URL	Intermediary	Unwound URL
t.co/mFQZTarzq8	1tv.ru/news/world/262978	https://www.1tv.ru/news/2014-07-12/37175-bezhenka_iz_slavyanska_vspominaet_kak_pri_ney_kaznili_malenkogo_syna_i_zhenu_opolchentsa
t.co/YrX3P0dbBa	bit.ly/1jnGB8T	http://themadjewess.com/2014/07/12/slovyansk-senjohnmccain-happy-yet-3-yr-old-boyutm_source=twitterfeed&utm_medium=twitter
t.co/tKsMz4Pg64	wp.me/psNtV-e9C	https://themadjewess.wordpress.com/2014/07/12/slovyansk-senjohnmccain-happy-yet-3-yr-old-boy
t.co/v04cDeioDE		https://www.youtube.com/watch?v=0w0hWtaTseQ
t.co/0zm1iQNRiU	bit.ly/1kRdk1u	http://deadcitizensrightssociety.wordpress.com/2014/07/13/slovyansk-senjohnmccain-happy-yet-3-yr-old-boyutm_medium=twitter&utm_source=twitterfeed
t.co/8I1BozNM00	buff.ly/1kRgJxx	http://themadjewess.com/2014/07/12/slovyansk-senjohnmccain-happy-yet-3-yr-old-boyutm_content=buffer577fa&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

t.co/HKINt3TfDo bit.ly/1kVCQCL

http://translate.google.
co.uk/translate?hl=en&sl=
pl&tl=en&u=http://www.tvn24.
pl/wiadomosci-ze-swiata,2/
trzylatek-ukrzyzowany-w-slowiansku-na-oczach-m
449340.html&sandbox=1
http://themadjewess.
com/2014/07/12/
slovyansk-senjohnmccain-happy-yet-3-yr-old-boy

t.co/XPvTQ660Wg

Table B.1: Source URLs

Appendix C

Official accounts

Username	Fullname	Represents country	Located in country
RusEmbBul	Посольство России в Болгарии	RU	BG
rusemb_pl	Russian Embassy, PL	RU	PL
RussianEmbassy	Russian Embassy, UK	RU	GB
RusEmbDK	Embassy of Russia, DK	RU	DK
EmbRusBotswana	RusEmbassy, Botswana	RU	BW
RusConsulateDC	Russian Consulate	RU	US
RusEmbNigeria	Russia in Nigeria	RU	NG
AmbasadaRusije	Russia in Serbia	RU	RS
RusEmb_Iceland	RusEmbassy Iceland	RU	IS
RussianEmbassyC	Russia in Canada	RU	CA
RusEmbGhanaEng	Russian Embassy in Ghana	RU	GH
russiauz	Russia in Uzbekistan	RU	UZ
RusEmbSG	RusEmbassy Singapore	RU	SG
RusEmbPakistan	RusEmbassy_Pakistan	RU	PK
RusEmbBrunei	RussianEmbassyBrunei	RU	BN
RusEmbGhana	Посольство России в Гане	RU	GH
RussiaJamaica	RusEmb in Jamaica	RU	JM
RusEmbTurkey	RusEmbTurkey	RU	TR
RusEmb_Korea	Russian Embassy	RU	KR
AmbRusFrance	Russie en France	RU	FR
RusEmbassyKabul	Russian Embassy in Kabul	RU	AF
RusEmbEst	Russia in Estonia	RU	EE
RusEmbMNG	Посольство России в Монголии	RU	MN

RusEmbSriLanka	Russian Embassy in Sri Lanka	RU	LK
Rus_Emb_Ireland	Russia in Ireland	RU	IE
LV_RUSEMBAS	Посольство РФ в ЛП	RU	LV
RusEmb_LT	RusEmb_LT	RU	LT
RusEmbSriLankaR	Посольство России	RU	LK
RusEmbassyUAE	Russian Embassy, UAE	RU	EA
AmbRusTun	Russian Embassy, Tunisia	RU	TN
rusembassyqatar	rusembassyqatar	RU	QA
RusEmbLux	Russian Embassy, Lux	RU	LU
EmbassyofRussia	Russia in RSA	RU	ZA
RusEmbHungary	Russian Embassy, HU	RU	HU
RusEmbNo	Embassy of Russia in Norway	RU	NO
RusEmbassyMinsk	Russia in Belarus	RU	BY
RusEmbIndia	Russia in India	RU	IN
RusEmbHungaryR	Посольство в Венгрии	RU	HU
rusemb_tm	Посольство России	RU	TM
RusEmbJakarta	Russian Embassy, IDN	RU	ID
RussianEmbFinla	Russian Embassy in Finland	RU	FI
Rus_Emb_Sudan	Russia in Sudan	RU	SD
RusEmb_Ecuador	Embajada de Rusia en el Ecuador	RU	EC
RusEmbBangkok	Russian Embassy TH	RU	TH
RusEmbPeru	Rusia en Perú	RU	PE
RusEmbGer	Владимир Гринин	RU	DE
Rusembegypt	RussianEmbassy EGYPT	RU	EG
AmbrusSlo	RusEmbSlovenia	RU	SI
RusEmbIran	Russian Embassy, IRI	RU	IR
RusEmbAU	Russia in Australia	RU	AU
RusembUkraine	RusEmbassy Ukraine	RU	UA
rusembassy	RusEmbassy in Guyana	RU	GF
AmbRusME	Russia in Montenegro	RU	ME
RusCons_TX	Russia in Houston	RU	US
RusEmbKuw	Russia in Kuwait	RU	KW
rusembitaly	Russian Embassy in Italy	RU	IT
RusEmbJordan	Russia in Jordan	RU	JO
RusEmbBAH	Russian Embassy in Bahrain	RU	BH
rusembassyARM	Посольство РФ в Армении	RU	AM

RusEmbassyIraq	Russia in Iraq	RU	IQ
rusembleb	Russian Embassy in Lebanon	RU	LB
RusEmbVietnam	RusEmbVietnam	RU	VN
RusEmbSyria	Russian Embassy, Syria	RU	SY
rusemb_dushanbe	Посольство РФ в РТ	RU	TJ
RusEmb_Malaysia	Embassy of Russia in Malaysia, Kuala Lumpur	RU	MY
RusEmbassyJ	ロシア大使館	RU	JP
RusEmbMauritius	Посольство России	RU	MU
RusEmbSK	Igor Bratčikov/Vel'vyslanectvo Ruska na SlovenSKU	RU	SK
Rusembchina	Посольство в Китае / Russian Embassy in China	RU	CN
RusEmbKG	Посольство РФ в КР	RU	KG
RusEmbEthiopia	Russia in Ethiopia	RU	ET
CamboRusEmba	Russian Embassy, KHM	RU	KH
RusCG_MZS	Russia in Mazari-Sharif (IRA)	RU	AF
RusEmbUganda	Russian Embassy in Uganda	RU	UG
RusEmb_KSA	Russian Embassy, Saudi Arabia	RU	SA
russembkenya	Russian Embassy in Kenya/Посольство России в Кении	RU	KE
russiaqatar	Russian Embassy in Qatar	RU	QA
IvanVolodinRUS	Ivan Volodin	RU	GB
RusEmbCyprus	RusEmbCyprus	RU	CY
RusEmbSwiss	Russian Embassy Bern	RU	CH
rusgkkirkenes	Consulate General of Russia in Kirkenes	RU	NO
RusEmbBih	Ambasada Rusije u BH	RU	BA
RusEmbCro	Veleposlanstvo Rusije	RU	HR
RusEmbUSA	Russian Embassy in USA	RU	US
rusembtz	Russia in Tanzania	RU	TZ
Russia_in_BEAC	Russia in BEAC	RU	RU
RusEmbMalta	Russian Embassy in Malta	RU	MT
RusiaColombia	RusEmbColombia	RU	CO
RusEmbCanada	Russian Embassy, CA	RU	CA

RusEmbUSApres rcgnewyork	Petr Svirin Consulate General of Rus- sia in New York	RU RU	US US
RuEmbZimbabwe	Russian Embassy in Zim- babwe	RU	ZW
RusEmbNam	Russian Embassy in Namibia	RU	NA
RusEmbDPRK	RussianEmbDPRK	RU	KP
RusEmbManila	Russian Embassy in the Philippines	RU	PH
RusEmbSwe	Russian Embassy, SWE	RU	SE
RusConsNiigata	Consulate General of Rus- sia in Niigata	RU	JP
rusembnz	Russian Embassy, NZ	RU	NZ
RusembApsny	Посольство России в Абхазии	RU	GE
RusEmb_Rwanda	Russian Embassy in Rwanda	RU	RW
Ru_Cons_Kolkata	The Consulate General of Russia in Kolkata	RU	IN
rusembassynl	Russian Embassy in NL	RU	NL
Russia_Toronto	Consulate General of Rus- sia in Toronto	RU	CA
emb_rus	RusEmbVatican	RU	VA
RusEmbSey	Russian Embassy in Sey- chelles	RU	SC
zambia_in	Russia in Zambia	RU	ZM
RusembZ	Russian Embassy in Zam- bia	RU	ZM
RussianEmbYemen	Russian Embassy in Yemen	RU	YE
rusemberitrea	Russia in Eritrea	RU	ER
UKRinSP	UKR Consulate in Sao Paulo	UA	BR
UKRinUN	UKR Mission to UN	UA	US
InnaYehorova	Inna Yehorova	UA	UA
DGovoroune	Dmytro Govoroune	UA	GB
pavlichenko_vm	V.Pavlichenko	UA	GB
UKRinVNM	UKR Embassy in VNM	UA	VN
UKRinMA	UKRAINEambMAROC	UA	MA
borodenkov	Andrii Borodenkov	UA	SI
shaloput	Olena Shaloput	UA	IE

ROgryzko	Rostyslav Ogryzko	UA	DE
UKRinSWE	UKR Embassy in SWE	UA	SE
UKRinChicago	Consulate in Chicago	UA	US
UKRinPRT	UKR Embassy in PRT	UA	PT
UKRinAUT	UKR Embassy in AUT	UA	AT
UKRinMalaga	UKR Consulate Malaga	UA	ES
UKRinKG	UKREmb in Kyrgyzstan	UA	KG
UKRinAUS	UKR Embassy in Aus	UA	AU
UKRinAZE	UKR Embassy in AZE	UA	AZ
RfUkr	UKR Embassy in RF	UA	RU
UKRinCoE	Ukraine in CoE	UA	BE
UKRinEdinburgh	UA Consulate in EDI	UA	GB
UKRinDEU	UKR Embassy in GER	UA	DE
UKRinTR	Ukraine in Turkey	UA	TR
UKRinUNESCO	UKR Mission to UNESCO	UA	FR
UkrInstitutet	Ukrainska Institutet	UA	SE
UKRinSEN	UKR Embassy in SEN	UA	SN
UKRinHRV	UKR Embassy in HRV	UA	HR
UKRinCUB	UKR Embassy in Cuba	UA	CU
UKRinLatvia	UKR Embassy in LVA	UA	LV
emb_sy	UKR Embassy in SAR	UA	SY
UKRinEST	UKR Embassy in EST	UA	EE
EMBUkraine	UKR Embassy in UZ	UA	UZ
UKRinKEN	Ukraine in Kenya	UA	KE
UKRinLBN	UKR_Emb in Lebanon	UA	LB
UKRinNGA	UKR Emb in Nigeria	UA	NG
UKRinIran	UKR Embassy in Iran	UA	IR
UKRinOSCE	Ukrainian Mission to OSCE & UN in Vienna	UA	AT
UKRinSRB	UKR Embassy in SRB	UA	RS
UKRinNOR	UKR Embassy in NOR	UA	NO
UKRinETH	UKR Embassy Ethiopia	UA	ET
UKRinPL	Ukraine in Poland	UA	PL
UKRinBGR	UKR Emb in Bulgaria	UA	BG
UKRinCZE	UKR Embassy in CZE	UA	CZ
UKRinIraq	UKR Embassy in Iraq	UA	IQ
UKRinEGY	Embassy of Ukraine in Egypt	UA	EG
UKRinAGO	UKR Embassy in AGO	UA	AO
UA_Emb_FI	UKR Embassy in FIN	UA	FI
UKRinKWT	UKR Embassy in KWT	UA	KW

UKRinThessalon	UKRConsulate inThess	UA	GR
UKRinSAU	UKR Embassy in SAU	UA	SA
UKRinKAZ	UKR Embassy in KAZ	UA	KZ
UKRinIsrael	Ukr Emb in Israel	UA	IL
UKRinCHE	UKR Embassy in CHE	UA	CH
UKRinPERU	UKR Embassy in Peru	UA	PE
UKRinLTU	UKR Embassy in LTU	UA	LT
UKRinMAS	UKR Embassy in MAS	UA	MY
UKRinJOR	UKR Embassy in JOR	UA	JO
UKRinMilan	UKR Cons in Milan	UA	IT
UKRinESP	UKR Embassy in Spain	UA	ES
UKRinFRA	UA Embassy in France	UA	FR
UKRinMunich	UKR Cons in Munich	UA	DE
UKRinIT	Ukr Embassy to Italy	UA	IT
UKRinARM	UKR Embassy in ARM	UA	AM
UKRinDZA	UKR Embassy in DZA	UA	DZ
UKRinBEL	UKRinBEL	UA	BE
UKRinSGP	UKR Embassy in SGP	UA	SG
UKRinCAN	UKR Embassy in CAN	UA	CA
UKRinRSA	UKR Embassy in RSA	UA	ZA
UKRinUNOG	Ukraine's Mission to UNOG	UA	CH
UKRinGEO	UKR Embassy in GEO	UA	GE
UkrEmbCy	UKR Embassy in CYP	UA	CY
UKRinNLD	UKR Embassy in NLD	UA	NL
UkrEmbLondon	Ukraine's Emb. to UK	UA	GB
UKRinLublin	UKR Consulate Lublin	UA	PL
UKRinMNE	UKR Embassy in MNE	UA	ME
UKRinMDA	UKR Embassy in MDA	UA	MD
UKRinARG	UKR Embassy in ARG	UA	AR
UKRinHUN	UKR Embassy in HUN	UA	HU
UKRinNATO	UKR Mission to NATO	UA	BE
UKRinVAT	UKR Emb to HOLY SEE	UA	VA
UKRinKorea	UKR Embassy in Korea	UA	KR
UkrinToronto	UKRinToronto	UA	CA
UKRinMKD	UKR Embassy in MKD	UA	MK
UKRinSF	UKR in SanFrancisco	UA	US
UkrainavRB	UKR Embassy in BLR	UA	BY
UKRinJPN	UKR Embassy in Japan	UA	JP
UKRinNewYork	UKR Consulate NY	UA	US
UKRinMEX	UKR Embassy in MEX	UA	MX

UKRinSLO	Ukraine in Slovenia	UA	SI
shalkivski	Volodymyr Shalkivski	UA	US
UKRinBRA	UKR Embassy in BRA	UA	BR
UKRinGRC	UKR Embassy in GRC	UA	GR
UKRinLBY	UKR Embassy in Libya	UA	LY
UKRinSR	UKR Embassy in Slovakia	UA	SK
ukr_embassy	UKR Embassy in INA	UA	ID
UKRinPAK	UKR Embassy in PAK	UA	PK
UKRinIRL	UKR Embassy in Ireland	UA	IE
UKRintheUSA	UKR Embassy in USA	UA	US
UKRinPorto	UKR ConsulateinPorto	UA	PT
UKRinBrest	КУ в Бресті	UA	BY
UKRinCIS	Ukr Mission to CIS	UA	RU
UKRinUAE	UKR Embassy in the UAE	UA	AE
UkrInstitute	Ukrainian Institute of America	UA	US
UKRinROU	UKR Embassy in ROU	UA	RO

Table C.1: Official accounts

Appendix D

Technical artifacts and tools

This section describes a number of technical artifacts that have been created as part of the research as well as existing tools used in more detail.

D.1 Data collection

To collect data from social media, a tool was developed that can collect and parse information. This collection is performed using a combination of existing Application Programming Interfaces (APIs) and scraping. Generally speaking, using APIs is the preferred way to gather structured data, but most platforms enforce strict limits or specific permissions to access certain types of data (Walker, Mercea, and Bastos, 2019). In such cases, a scraping tool has to be used. Preliminary research has shown that both Facebook and Twitter have recently made significant changes to their platform, breaking many existing scraping tools.

D.1.1 Scraping process

Traditional scraping tools rely on collecting data from HTTP requests to regular web pages presented in the HyperText Markup Language (HTML), which often contains a mix of actual content as well as elements that control how this content is presented. This can be seen as an alternative to APIs, which provide only the content in a machine-readable format (usually XML or JSON). This introduces a number of challenges with traditional scraping tools, including but not limited to:

Noise The HTML elements used for the presentation of the content cause noise and often make it hard to find specific content within a page without relying on long element selector queries like those written using the XML Path Language (XPath).

Breaking changes With platforms offering an API as the official route to request information in an automated way, the web pages themselves are not considered to be something external parties rely on. For this reason breaking changes can be made without any announcement. A minor change in the layout of a web page can therefore cause a scraping tool to stop working.

JavaScript Modern web applications use many interactive components that load data asynchronously using AJAX. With the rise of JavaScript-based single page applications, this has become the standard for many web sites. This means that scraping tools have to take interactive component and asynchronous requests into account.

Anti-scraping measures Many large websites employ a form of anti-scraping measures. One of such measures uses browser capabilities and other fingerprinting techniques to distinguish programmatic HTTP clients from genuine web browsers.

Each of the challenges above was encountered during preliminary research, and it was thus deemed necessary to develop an alternative approach to scraping. One of the challenges, the prevalence of JavaScript and the usage of asynchronous requests, actually provides a great opportunity for achieving this as such requests usually rely on a form of content-only machine-readable data. Figure D.1 details the method used.

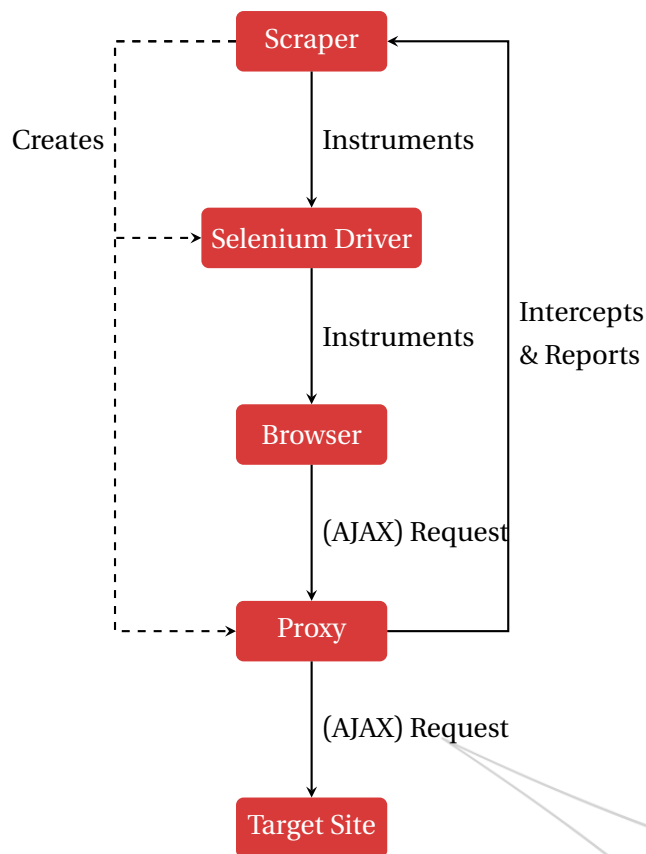


Figure D.1: Scraping with asynchronous request (AJAX) support

This method has a number of advantages over scraping using programmatic HTTP clients:

Real browser A testing toolkit is used to instrument a real browser, circumnavigating possible fingerprinting techniques identifying scraping tools.

Session storage The browser can be instructed to retain login cookies, which prevents multiple logins across sessions.

JavaScript support Support for JavaScript provided by the web browser allows for asynchronous requests.

Proxy support The browser can be instructed to use a proxy to intercept and report on any request made by the web page, including asynchronous requests.

D.1.2 Implementation

The process described in Figure D.1 was implemented by a tool built in the C# programming language using the .NET framework, due to the familiarity of the author with this technology. The tool uses the Selenium web browser automation tool¹ to launch the Chrome² web browser, log in to a genuine account, and browse to the targeted page. When the web browser is launched, it is instructed to use a proxy server, directly hosted from the same C# application using the Titanium Web Proxy³ library. This allows for intercepting any asynchronous requests made by the target page. Many targeted web pages (e.g. Twitter search) use infinite scrolling to keep loading more content, in this case Selenium is used to make the web browser scroll down until a new request for search results is intercepted.

D.2 Data storage

The collected data was stored in a Microsoft SQL database, due to the ease of manual querying and analysis and the author's familiarity with the technology. This also allowed for further processing using other technologies. Figure D.2 shows the entity relationship diagram for the database used to store the collected data.

¹<https://www.selenium.dev/>

²<https://www.google.com/chrome/>

³<https://github.com/justcoding121/titanium-web-proxy>

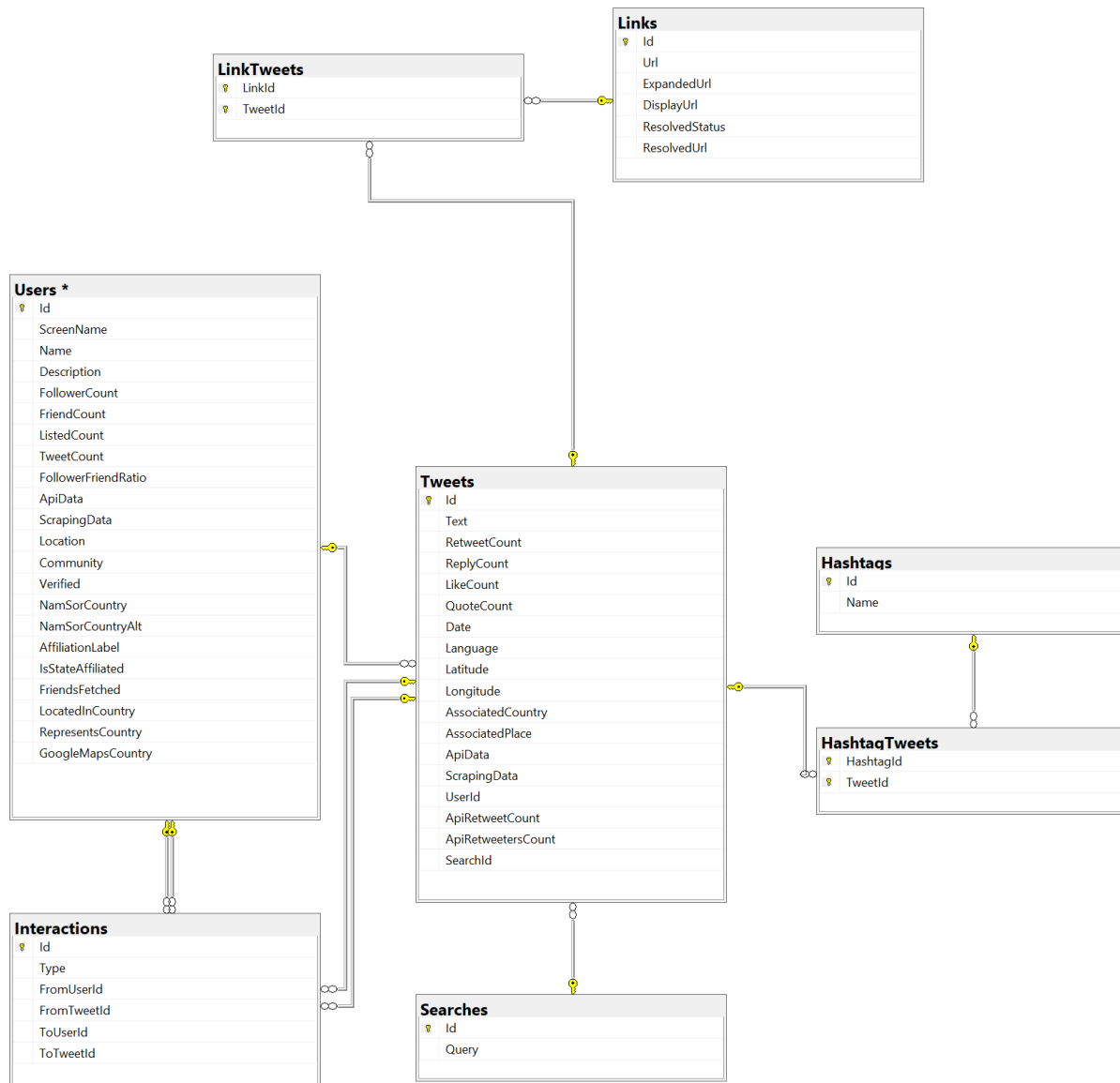


Figure D.2: Entity relationship diagram

D.3 Data analysis and Machine Learning

Data analysis and the application of machine learning was done within a Jupyter Notebook⁴ environment using the python programming language. The Jupyter notebook environment provides a way to quickly analyze data and leverage popular Python libraries like Pandas⁵ and Scikit Learn⁶. The mention network was visualized using Gephi⁷.

⁴<https://jupyter.org/>

⁵<https://pandas.pydata.org/>

⁶<https://scikit-learn.org/>

⁷<https://gephi.org/>

D.4 Other tools

All tools used for the data collection, storage and analysis described above were run in a docker⁸ environment, to keep them self-contained and portable. This document was written in \LaTeX using the Overleaf⁹ text editor and Zotero¹⁰ bibliography manager. All artifacts created as part of this research were stored using the git version control system¹¹.

⁸<https://www.docker.com/>

⁹<https://www.overleaf.com/>

¹⁰<https://www.zotero.org/>

¹¹<https://git-scm.com/>