UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1786

Techniques for analyzing digital environments from a security perspective

AMENDRA SHRESTHA

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2019

ISSN 1651-6214
ISBN 978-91-513-0605-6
urn:nbn:se:uu:diva-379605

Dissertation presented at Uppsala University to be publicly examined in Room 2446, ITC, Lägerhyddsvägen 2, Uppsala, Friday, 17 May 2019 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Adam Wierzbicki (Polish-Japanese Institute of Information Technology).

Abstract

Shrestha, A. 2019. Techniques for analyzing digital environments from a security perspective. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1786. 64 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-0605-6.

The development of the Internet and social media has exploded in the last couple of years. Digital environments such as social media and discussion forums provide an effective method of communication and are used by various groups in our societies. For example, violent extremist groups use social media platforms for recruiting, training, and communicating with their followers, supporters, and donors. Analyzing social media is an important task for law enforcement agencies in order to detect activity and individuals that might pose a threat towards the security of the society.

In this thesis, a set of different technologies that can be used to analyze digital environments from a security perspective are presented. Due to the nature of the problems that are studied, the research is interdisciplinary, and knowledge from terrorism research, psychology, and computer science are required. The research is divided into three different themes. Each theme summarizes the research that has been done in a specific area.

The first theme focuses on analyzing digital environments and phenomena. The theme consists of three different studies. The first study is about the possibilities to detect propaganda from the Islamic State on Twitter. The second study focuses on identifying references to a narrative containing xenophobic and conspiratorial stereotypes in alternative immigration critic media. In the third study, we have defined a set of linguistic features that we view as markers of a radicalization.

A group consists of a set of individuals, and in some cases, individuals might be a threat towards the security of the society. The second theme focuses on the risk assessment of individuals based on their written communication. We use different technologies including machine learning to experiment the possibilities to detect potential lone offenders. Our risk assessment approach is implemented in the tool PRAT (Profile Risk Assessment Tool).

Internet users have the ability to use different aliases when they communicate since it offers a degree of anonymity. In the third theme, we present a set of techniques that can be used to identify users with multiple aliases. Our research focuses on solving two different problems: author identification and alias matching. The technologies that we use are based on the idea that each author has a fairly unique writing style and that we can construct a *writeprint* that represents the author. In a similar manner, we also use information about when a user communicates to create a *timeprint*. By combining the writeprint and the timeprint, we can obtain a set of powerful features that can be used to identify users with multiple aliases.

To ensure that the technologies can be used in real scenarios, we have implemented and tested the techniques on data from social media. Several of the results are promising, but more studies are needed to determine how well they work in reality.

Keywords: digital communities, machine learning, text analysis, linguistic features, linguistic analysis, warning behaviors, Internet, social media, extremism, terrorism, psychological state, author identification, alias matching

Amendra Shrestha, Department of Information Technology, Computer Systems, Box 337, Uppsala University, SE-75105 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-513-0605-6
urn:nbn:se:uu:diva-379605 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-379605)

Dedicated to my family, friends, and my supervisor

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I L. Kaati, E. Omer, N. Prucha, and A. Shrestha. Detecting multipliers of jihadism on Twitter. In *IEEE 15th International Conference on Data Mining Workshops (ICDMW)*, 2016 *

II K. Cohen, L. Kaati, S. Lindquist, and A. Shrestha. Automatic detection of xenophobic narratives: A case study on Swedish alternative media. In *IEEE Conference on Intelligence and Security Informatics (ISI)*, 2016 *

III K. Cohen, T. Isbister, L. Kaati, and A. Shrestha. Linguistic markers of a radicalized mind-set among extreme adopters. In *1st International Workshop on Cyber Deviance Detection (CyberDD)*, 2017 *

IV A. Shrestha, L. Kaati, and K. Cohen. A machine learning approach towards detecting extreme adopters in digital communities. In *1st International Workshop on Advanced ICT Technologies for Secure Societies (AICTSS)*, 2017

V K. Cohen, L. Kaati, and A. Shrestha. Linguistic analysis of lone offender manifestos. In *IEEE International Conference on Cybercrime and Computer Forensic (ICCCF)*, 2016 *

VI L. Kaati, T. Sardella, and A. Shrestha. Identifying warning behaviors of violent lone offenders in written communication. In *IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2016 *

VII N. Akrami, A. Shrestha, M. Berggren, L. Kaati, M. Obaidi, and K. Cohen. *Assessment of risk in written communication: Introducing the Profile Risk Assessment Tool (PRAT)*. The European Commission, Europol, 2018

VIII F. Johansson, L. Kaati, and A. Shrestha. Detecting multiple aliases in social media. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2013 *

IX F. Johansson, L. Kaati, and A. Shrestha. Timeprints for identifying social media users with multiple aliases. *Security Informatics*, 4(1):7, 2015 *

X M. Ashcroft, F. Johansson, L. Kaati, and A. Shrestha. Multi-domain alias matching using machine learning. In *Third European Network Intelligence Conference (ENIC)*, 2016 *

Reprints were made with permission from the publishers.

I initiated the projects, collected data, designed, and conducted the experiments. I am the sole implementer. The ideas were developed in discussion with other authors. All authors contributed to the writing. I implemented the tool described in the report VII.

*Authors names are written in alphabetical order

Other publications

The following papers were not included in this thesis.

XI M. F. Atig, S. Cassel, L. Kaati, and A. Shrestha. Activity profiles in online social media. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2014

XII F. Johansson, L. Kaati, and A. Shrestha. Time profiles for identifying users in online environments. In *IEEE Joint Intelligence and Security Informatics Conference (JISIC)*, 2014 **(Best Paper Award)**

XIII L. Kaati, E. Lundeqvist, A. Shrestha, and M. Svensson. Author profiling in the wild. In *2017 European Intelligence and Security Informatics Conference (EISIC)*, 2017

Authors names are written in alphabetical order

Acknowledgements

First and foremost, I would like to thank my supervisor Lisa Kaati. Thank you for your excellent supervision, guidance, and support during the whole Ph.D. period. Her competence in the field of Security Informatics and her attitude of excellence towards handling the problems have influenced me a lot. The scientific discussions, tips, and advice that she has provided have helped me to solve the research problems. My Ph.D. research could not have been smooth without an advisor like her. I am thankful for her continuous support over the years.

Similar, profound gratitude goes to my co-supervisors, Bengt Jonsson, Mohamed Faouzi Atig, and Michael Ashcroft for their helpful comments and feedback during my doctoral studies and while writing the thesis. I would like to thank Michael Ashcroft for organizing meetings and explaining machine learning algorithms in a clear and thorough manner. Thank you for your brilliant ideas and valuable feedback throughout my Ph.D. period. My special gratitude goes to Parosh Abdulla, without whom I would not have started my Ph.D. at Uppsala University. I would also like to show my appreciation towards members of the Algorithmic Program Verification group: Carl, Diep, Frederic, Jari, Othmane, Phong, Quy, and Yunyun. I appreciate the level of support and friendliness shown by you all. My gratitude also goes to the administrative staff of the Department of Information Technology, Ulrika Andersson, Anna-Lena Forsberg, Elisabeth Lindqvist, and Anne-Marie Jalstrand for helping me on my administrative requests.

During my Ph.D. period, I had an opportunity to meet so many outstanding people other than from the department. I would like to thank Fredrik Johansson, Katie Cohen, Tim Isbister, Johan Fernquist, and Björn Pelzer from Swedish Defence Research Agency (FOI) and Nazar Akrami, Mathias Berggren, and Milan Obaidi from Department of Psychology, Uppsala University. Thank you all for the collaboration and exciting discussion. It has been great working with you all.

Last but not least, I would like to express my deepest gratitude to my family and friends for their unconditional love and encouragement. Thank you, my wife, Monika. I am deeply overwhelmed by the support and endless tolerance shown by her during my Ph.D.

Sammanfattning på svenska

Utvecklingen av Internet och sociala medier har fullkomligt exploderat de senaste åren. Olika digitala miljöer och sociala medier gör det möjligt för användare att kommunicera med likasinnade oavsett vart i världen man befinner sig. De digitala miljöerna har många viktiga funktioner i de flesta människors liv, men det finns också ett flertal exempel på hur digitala miljöer används för att sprida allt från hat och hot till våldsbejakande propaganda genom text, bilder, symboler, musik och film. I den här avhandlingen beskrivs ett antal tekniker som kan användas för att analysera digitala miljöer med fokus på hot, hat och våldsbejakande extremism. På grund av den stora mängden data som finns på internet är det i princip omöjligt att genomföra manuella analyser. Med hjälp av datorstöd ökar våra möjligheter att tolka och förstå innehållet i data vilket i sin tur leder till ökade möjligheter till relevanta analyser av digitala miljöer. Datorstöd innefattar en mängd olika tekniker, allt ifrån enklare former av ordräkningsbaserade textanalyser till mer avancerade analyser baserade på maskininlärning.

I ett flertal av våra studier är vi intresserade av psykologiska indikatorer som tankar, känslor och motivationsfaktorer som finns bakom en text. För att studera psykologin bakom orden har vi använt oss av ett textanalysverktyg som heter Linguistic Inquiry and Word Count (LIWC). LIWC är utvecklat av psykologer med syfte att automatiskt kunna studera psykologiska fenomen i en text. LIWC har använts i en mängd olika studier och visat sig vara ett tillförlitligt redskap för att undersöka språkets psykologiska dimensioner.

I avhandlingen beskrivs tekniker som kan användas för att studera digital kommunikation från en grupp men också enskilda individer, till exempel för att göra en riskbedömning av en individ baserat på en text. Vi presenterar också teknik som möjliggör identifiering av individer som använder sig av olika alias med hjälp av deras digitala kommunikation.

Digital kommunikation kan analyseras på många olika sätt. I denna avhandling presenteras tre studier som innehåller analyser av olika fenomen. En av studierna beskriver hur man kan identifiera förekomsten av ett högerextremt narrativ i invandringskritiska alternativmedier med hjälp av ordlisterbaserade tekniker och stöd av experter. En annan studie visar hur man automatiskt kan identifiera propaganda från terrorgruppen Islamiska Staten med hjälp av maskininlärning. I ytterligare en studie använder vi oss av lingvistiska särdrag i en text för att skapa en bild av författarens uppfattningar, relationer och känslomässiga tillstånd. Genom att förena forskning om radikaliseringens psykologi med forskning kring sambandet mellan psykologi och språk-

bruk beskriver vi hur en uppsättning lingvistiska markörer kan användas för att studera radikalisering i digitala miljöer.

I några digitala miljöer återfinns individer som begår våldsdåd, exempelvis ensamagerande våldsverkare. Ensamagerande våldsverkare är individer som på egen hand utövar eller förbereder terror, massmord, seriemord eller allvarlig skadegörelse. Ensamagerande är ett reellt hot mot samhället säkerhet och eftersom de kan förbereda sina dåd i tysthet är de svårare att upptäcka och stoppa i tid. Ytterligare en försvårande omständighet när det gäller att förebygga attacker är det inte går att göra en profil över en typisk ensamagerande. Tidigare erfarenheter av ensamagerande har nämligen visat att det finns en stor variation i motiv, bakgrund och ideologiska övertygelser. I tidigare studier har det visat sig att ett flertal ensamagerande offentliggjort sina åsikter skriftligt, via brev, online-dagböcker, foruminlägg, manifest och liknande innan de begått sina dåd. I de fall dessa texter går att finna i öppna källor har vi tillämpat ett antal metoder för datoriserad textanalys i syfte att lära oss mer om huruvida det finns gemensamma lingvistiska särdrag som speglar deras tillstånd, personlighet och självbild. Målet är att nå en ökad förståelse för de processer som leder ensamma individer till att begå våldshandlingar. Ett annat mål är att undersöka möjligheten att använda automatiserade tekniker som hjälpmedel i våldsriskbedömningar. De tekniker som vi använt oss av för att riskbedöma texter är baserade på dels ordlistor och dels på maskininlärning. Resultatet är implementerat i prototypen till verktyget Profile Risk Assessment Tool (PRAT) som används för att riskbedöma texter.

Ett av problemen med kommunikation på internet är användningen av olika användarkonton som inte explicit talar om vilken person som ligger bakom kontot, utan erbjuder någon form av anonymitet. För att kunna avgöra om en författare använder sig av flera alias för att kommunicera eller för att känna igen om två texter är skrivna av samma författare kan man använda sig av ett flertal olika tekniker. För att kunna jämföra texter kan man använda sig av något som kallas för *skrivavtryck*. Tanken med skrivavtryck är att alla människor har ett mer eller mindre specifikt sätt att skriva på och ett skrivavtryck kan användas för att identifiera individer på samma sätt som man använder sig av fingeravtryck för identifiering. Skrivavtrycket baseras på stylometriska särdrag i texter exempelvis subtila skillnader av användandet av funktionsord och kommatering. På samma sätt kan man tänka sig att de tidpunkter när individer väljer att kommunicera är något som kan användas för identifiering. Genom att använda sig av de tidpunkter som en individ kommunicerat på (givet att det finns tillräckligt mycket data) kan man skapa ett tidsavtryck som precis som skrivavtrycket kan användas för att identifiera en användare. I den här avhandlingen presenteras en uppsättning tekniker som kan användas för att skapa och jämföra både skrivavtryck och tidsavtryck.

De tekniker som presenteras i den här avhandlingen är framför allt anpassade för att analysera textbaserad kommunikation. För att säkerställa att teknikerna kan användas i riktiga scenarios har vi implementerat och testat

teknikerna på data från sociala medier. Flera av resultaten är lovande men det behövs fler studier för kunna avgöra hur väl de fungerar i verkligheten.

Contents

1. Introduction

Internet and social media are integrated into the everyday lives of people and used by both individuals and groups for many different purposes. For example, for sharing and gathering information, expressing and exchanging views, and for education and e-commerce. The advancement in technologies makes it possible for a large number of individuals from all over the world to communicate. Social media users are no longer just passive consumers; they also participate in a digital community by producing media content. Users can share their personal stories, life experiences, and opinions making the shared content more personalized and subjective.

Most of the user-generated content is produced with the intention to communicate with other people. However, Internet and social media are not only used for harmless communication they also provide individuals and groups with the possibility to engage in criminal activities [49] [71]. The Internet has, for example, played a role in both radicalization processes and when it comes to finding information on how to conduct violent acts and terrorist attacks. A study by Gill et al. [32] showed that in a group of 223 convicted United Kingdom-based terrorists there was evidence of online activity related to their radicalization and/or attack planning in 61% of the cases. Online activities may precede future threats toward the society [15] [23] [55] and therefore monitoring the Internet and social media is an important task for intelligence analysts and law enforcement agencies. By analyzing online environments and individuals that might pose a threat towards the security of society, there is a possibility to prevent future attacks before they take place.

This thesis focus on three aspects of analyzing data from the Internet and social media. Firstly, we show how different technologies can be used to analyze different aspects of digital environments. We study the possibilities to identify propaganda from the so-called Islamic State on Twitter. More precisely, the use of an extreme right narrative in alternative media, and the presence of linguistic markers of radicalization among writers in a discussion forum. Secondly, we address the problem of risk assessment of written communication. We show how different technologies can be used to assess the risk that an individual will commit targeted violence. Finally, we present a set of techniques that can be used to identify users with multiple aliases. These kinds of technologies are important when studying online communication since the same individuals have the possibility to use different aliases when they communicate.

1.1 Data sets and Challenges

The research field of the thesis belongs to *Intelligence and Security Informatics* (ISI) [19]. ISI is an interdisciplinary research field that involves academic researchers from different disciplines as well as practitioners from law enforcement and intelligence agencies. The aim is to use information technology to address security-related problems related to counter-terrorism, homeland security missions, and responses to terrorist acts. The problems that we have addressed in our research are generally difficult and do not have a straight forward solution. Working on solutions for the problems requires an interdisciplinary approach that combines knowledge from different field of research like terrorism research, psychology, and computer science.

Our research is based on data from various forms of social media such as Twitter, forums, blogs, and news. This is common for a lot of ISI research. One of the major problem when working with security-related problems and Internet data is the lack of suitable data sets that can be used to develop and evaluate algorithms. For the problems that we consider, it is a great challenge to find appropriate data. In cases where relevant data sets exist the amount of data is usually very small, which is a problem when developing and evaluating algorithms. One of the major challenges with ISI research is to find relevant data sets and develop algorithms despite lack of data. We have used different approaches to find relevant data. Most of the data is collected by ourselves, but in some cases, we have used data sets that were made public by scientific conferences.

1.2 Ethics and Social Media Analysis

The use of social media data in research poses essential ethical concerns. One concern with social media data is if the data should be considered public or private data. Whether online communication should be regarded as public or private is determined by the online community and the social media user's expectation of privacy. A discussion forum where you have to be a member and log in with a password can, for example, be considered private while an open discussion on Twitter in which people use hashtags to associate messages to subjects can be regarded as public.

A component in all types of research is to get informed consent by participants. Informed consent is a problem in social media research since (in many cases) the data is collected and analyzed without having the participants informed consent. The participants (social media users) are not aware of their participation, and it is not possible to get consent. Instead in this research, we have minimized the intrusion of people's privacy by anonymizing their real identity. Anonymity is something else that needs to be considered when analyzing social media. The identity of the participants becomes important when

the data that is analyzed could contain sensitive information that is exposed in a context that was not intended.

To ensure privacy and comply with GDPR[1], we have taken measures to protect the integrity and privacy for individuals. Our analyses are limited to open source data, meaning that the data we have used in our studies are collected from sources that are accessible to everyone. No data has been gathered from password-protected sites, closed Facebook pages, or other types of websites or social media where posted material is kept accessible only to a closed audience. Some of the data sets that we have used are released by scientific conferences and competitions while some of the data are downloaded from sources such as news, blogs, Twitter, and discussion forums.

When using data from social media, it may be the case that people reveal their real identity. The identity of research subjects is anonymized, and the personal information or information that could be used to identify a user is not published or misused. When possible, the analysis is done on an aggregated level. All data are kept confidential. Most of the research described in this thesis includes creating numerical feature vectors of data. The vectors are used to train machine learning models or to measure the use of different words. The data is stored in a numerical format as feature vectors, and the data in its original format was deleted as the research experiments were completed. The data sets that we have used are cited or acknowledged in the publications.

Automatically monitoring the Internet for security reasons raises many concerns. The research presented in this thesis is intended to be used in an intelligence and security perspective, but the technologies described may pose a threat to privacy and online anonymity. If the suggested techniques can be used to reveal the true identity of a potential lone offender, there is a risk that the same techniques can also be used for other purposes. There is always a possibility that the technologies are used in an undesirable manner by actors with unfriendly intentions.

1.3 Thesis Overview

The research presented in this thesis is divided into three different themes. Each theme summarizes research that has been done in a specific area. In Chapter 2, we provide definitions of terms that are used in the thesis. The chapter describes what kind of features we use to build our classification models and how the classification models are evaluated.

In Chapter 3, we present three different studies on digital communication. The first study focuses on detecting propaganda from IS on Twitter. The second study focuses on identifying references to a narrative containing xenophobic and conspiratorial stereotypes in alternative immigration critic media. In

[1]General Data Protection Regulation (EU) 2016/679 is a regulation in EU law on data protection and privacy for all individuals within the European union

the third study, we define a set of linguistic features that we view as markers of a *radicalized mind-set*. A radicalized mind-set is a certain style of understanding and relating to the world that has often been observed among violent extremists.

In Chapter 4, we focus on risk assessment of written communication. We have studied texts written by violent lone offenders before they attack with the aim to learn more about them: about their personality, their emotional state and how they see themselves. We use different technologies including machine learning to experiment on the possibilities to detect potential lone offenders before they attack. Our risk assessment approach is implemented in the tool PRAT (Profile Risk Assessment Tool).

In Chapter 5, we present research done with the aim to identify users with multiple aliases on social media. Our research focuses on solving two different problems: author identification and alias matching (or similarity detection). We present different techniques that can be used to address these two problems and the different sets of features that we have used.

Finally, in Chapter 6, we summarize the contributions of our research and provide some concluding remarks.

2. Preliminaries

In this chapter, we provide some definitions that are used throughout the thesis. In particular, we define terms like features, feature vectors, and feature transformation. We explain different sets of features that we use, namely *stylometric feature*, *time-based feature* and *emotion-based feature*. We also describe the text analysis tool *Linguistic Inquiry and Word Count (LIWC)* that we have used in our research. Finally, we describe the terminology that we use to evaluate classification models.

2.1 Features

When analyzing text data, it is common to convert the text into a set of features. The set of features are attributes that constitute a characteristic property or set of properties that are unique, measurable, and differentiable. Ideally, a set of features should be informative, discriminating, and independent. For example, when detecting spam, features may include the sender's email address, the presence or absence of certain email headers, the email language and structure, the subject of an email, the frequency of specific terms, and the grammatical correctness of the text. In our case, we create an n-dimensional vector of numerical features that represents an object (in our case a user/alias). The vector is called a *feature vector*.

When creating a numerical feature vector for a user, we take the individual posts made by the user. Next, we aggregate all posts into one text. The feature vector created from the text represents the user. Each position in the feature vector commonly corresponds to the number of occurrences of each feature expressed with its relative frequency. It is also possible to use a function of such features as a feature. The process of replacing original feature counts with functions of features is called *feature transformation*. The new features might not have the same interpretation as the original features, but they might have more discriminatory power in a different space than the original space. Some of the feature scaling transformation techniques used in this work are:

I. **Scaling and centering:** Scaling is the process of dividing all values of a feature by its sample standard deviation. Centering of a feature is done by subtracting its sample mean from all values. The process of scaling and centering brings all variables on equal standing, i.e., all the feature values are treated similarly.

II. **Normalization:** Normalization is done to prevent features with large numerical values from dominating in distance-based objective functions. In our case, normalization is done by dividing the actual count of a feature by the total count of words in the text.

2.2 Types of Features

Our research focuses on content-based analysis along with metadata related to the content, and we have used several types of features. The features we consider are stylometric, time-based, and emotion-based.

2.2.1 Stylometric features

Stylometry refers to the statistical analysis of writing style [82]. Stylometric features measure computable and countable language features like word length, phrase length, sentence length, vocabulary frequency, distribution of words of different lengths. The stylometric features are a combination of lexical features, structural features, syntactic features, and content specific features.

I. **Lexical features** measure the vocabulary richness of an individual's writing by counting words and characters in their text. Some of the lexical features could be the total number of words or characters per sentence or frequency of use of certain letters or words. An individual might use some words more frequently than others which helps to reveal a unique idiosyncrasy of an individual. The advantage of these features is that they can be applied to any corpus in any language and with no additional requirements.

II. **Structural features** are used to analyze the organization and layout of the text. These types of features are important when analyzing short text like emails and online messages [30]. An individual might have a habit of constructing their paragraphs, such as always starting a sentence with lowercase letters or usual way of using greeting and farewell words in their emails.

III. **Syntactic features** focus on identifying patterns used to construct the sentences. To distinguish the individual, it analyzes the usage of all-purpose words like function words and punctuations. The use of punctuations like the comma (",") varies between authors since it is not restricted to be used in specific places. Research [17] has shown that punctuations are useful in discriminating authors.

IV. **Content specific features** refer to words that are relevant to a particular topic domain, discussion forums or a group of individuals. These types of features are used minimally since it is difficult to generalize in cross-topic settings.

A list of the sets of features we have used in our work is shown in Table 2.1. The count column in the table refers to the number of dimensions a specific set of features holds.

Table 2.1. *Examples of features for stylometric matching.*

Category	Description	Count
Text length	Frequency of number of characters in text	1
Word length	Relative frequency of words with 1-20 characters	20
Letters	Relative frequency of *a* to *z* (ignoring case)	26
Digits	Relative frequency of 0 to 9	10
Punctuation	Relative frequency of characters . ? ! , ; : ( ) " - ´	11
Function words	Relative frequency of various function words	488

2.2.2 Time-based features

Apart from examining the text, meta-data associated with the text can also be used for authorship analysis. In the case of social media posts, the meta-data could be the username of an author, time-stamp of a post, a location from where the post is created or a profile picture of an author. In this thesis, we use the time-stamp of a post as a feature. We have used the following sets of time-based features:

- **Hour Of Day:** each hour of the day,
- **Period Of Day:** four-hour intervals (early morning, morning, midday, evening, night, midnight)
- **Month:** each month of year
- **Day:** each day of week
- **Type Of Day:** weekdays and weekends

2.2.3 Emotion-based and Twitter specific features

Emotion-based and Twitter specific features are intended to capture various kinds of emotions and Twitter-specific content. The emotion and Twitter specific features that we have used are described in Table 2.2. Example of emotion words that we have used are *sad*, *happy*, *angry*, *mad*, etc.

Table 2.2. *Example of feature set for emotions and tweet-specific content.*

Category	Description	Count
Emotion words	Relative frequency of various sentiments words	108
Smilies	Relative frequency of various smilies :') :-) ;-) :P :D :X <3 :) ;) :@ :* :| :$ %)	14
Hashtag	Relative frequency of hashtags	1
User Mention	Relative frequency of user mentions	1
URLs	Relative frequency of URLs	1

2.3 Linguistic Inquiry and Word Count (LIWC)

Linguistic Inquiry and Word Count (LIWC) [63] is a text analysis tool that generates psychologically meaningful categories from a given text. LIWC was developed by James W. Pennebaker at the University of Texas and has been evaluated and tested in a number of different studies [61] [74]. LIWC has 73 psychological dimensions of the language related to emotions, social and cognitive processes, and attentional processes. LIWC is a dictionary based tool that consists of a number of different dictionaries (also called categories). Some of the categories and example words are illustrated in Figure 2.1. LIWC checks each word contained in a document against an internal dictionary of almost 6400 words (e.g., *with*, *but*, and *they*) and word stems (e.g., *punish*∗, and *revenge*∗). The use of LIWC has been expanded to several contexts of psychology and word use. The dictionaries are currently translated into more than twelve different languages. The research presented in this thesis employs the English version of LIWC 2015 and a Swedish translation of the English dictionaries in LIWC 2007.

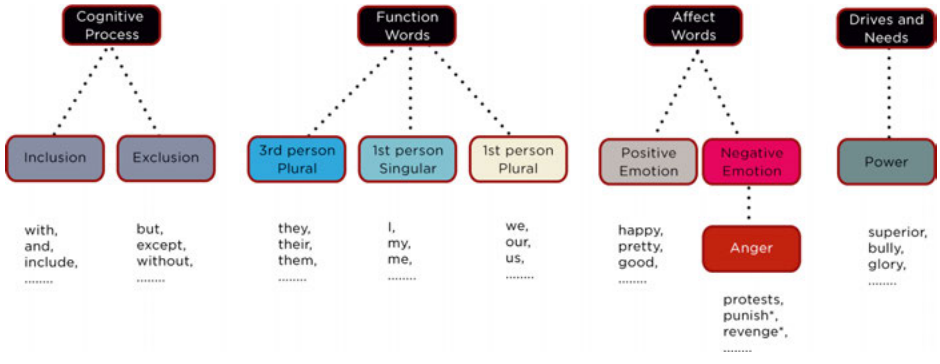

Figure 2.1. A subset of the LIWC 2015 categories and some example words.

2.4 Evaluating Classification Models

Most of the research problem described in the thesis are solved using machine learning classification algorithms. These algorithms build a mathematical model based on some example inputs and use the model to make predictions or decisions. Throughout the thesis, we use the term *model* when we discuss machine learning classification models. A model is a learner which is built using historical data that learn about the distribution of data and facilitates us on making inferences of new unseen objects.

Usually, the results of classification experiments are reported using a confusion matrix, also known as a contingency table or error matrix [21] [38] . A confusion matrix describes the performance of a model in a matrix. Each

row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class (and vice versa). The number of true positives, false negatives, true negatives, and false positives for the binary classification problem is reported as illustrated in Table 2.3. In case of binary classification problem, a new unseen object needs to be assigned to one of the predefined two discrete classes according to its characteristics. Accuracy, specificity, sensitivity, precision, balanced accuracy, etc. can be calculated from the confusion matrix.

Table 2.3. *Confusion matrix for binary classification.*

	Predicted class	
Actual class	**True Negatives (TN)**	**False Positives (FP)**
	False Negatives (FN)	**True Positives (TP)**

Accuracy is the ratio of the number of correct predictions to the total number of input samples. To measure the quality of a predictor, accuracy is a good indicator of how many correct classifications were made. Using the notation of Table 2.3, the overall accuracy is formulated as $\frac{TP+TN}{TP+FP+TN+FN}$. Specificity measures the proportion of actual negatives that are correctly identified while model testing and is defined as $\frac{TN}{TN+FP}$. Sensitivity or recall measures the actual positives that are correctly identified and is defined as $\frac{TP}{TP+FN}$. Precision measures the positive cases that were correctly identified and is identified as $\frac{TP}{TP+FP}$. F–1 score is the weighted average of precision and sensitivity. It takes both false positives and false negatives into account while evaluating models. It finds an optimal blend of precision and sensitivity. Maximizing the F–1 score ensures that we get a reasonably high precision and sensitivity. F–1 score is defined as: $2 * \frac{Precision * Sensitivity}{Precision + Sensitivity}$. Balanced accuracy is used to avoid inflated estimates on imbalanced data sets. The definition is $\frac{0.5 \cdot TP}{TP+FN} + \frac{0.5 \cdot TN}{TN+FP}$

The goal of any model is to learn from examples and generalize some degree of knowledge regarding the task it was trained to perform. Since the performance of a model depends on the problem you are solving and the data you use it is hard to say when a model performs good enough. The performance of different models can only be compared with other models that are trained on the same data.

Even though the accuracy of a model is more than 80%, we cannot assure that the model will work well when it is applied to new data in a realistic setting. In our work, we have used features that are not dependent on a specific data set since we assume that it will increase the performance of the model in the wild. The model needs to be tested and evaluated on realistic data to assure that it works good enough for the application it should be used for.

3. Analyzing Digital Communication

Analyzing digital communication can be done using a number of different approaches. One approach to studying digital communication is to use dictionaries. Dictionaries can be pre-defined (such as in LIWC) or developed specifically for the environment that is studied. By counting the occurrences of the words in the text, it is possible to get an understanding of how much (or how little) is being written about any specific topic. One of the difficulties with the dictionary-based analysis is to come up with all the relevant keywords. If a number of human experts are asked to extract keywords from a text, their choices will probably not agree to any great extent. When analyzing text on the Internet, and especially in social media, this problem becomes severe, since this type of text is usually informal, with much higher variation in vocabulary than what is found in a standardized language such as news text. The problem of vocabulary variation stems not only from the broad diversity of the vocabulary but also from how the choice of words is made. It is clearly domain-specific, which requires in-depth knowledge of the domain. This is the case with different digital groups and communities, which tend to use domain-specific terminology.

Another approach to studying digital communication is to use machine learning technologies. Machine learning can be used to solve many different problems such as sentiment and affect analysis or to determine if a text originates from a specific group or not. One of the challenges when using machine learning in this context is to find relevant training and test data. In this thesis, three different studies of digital communication are included. The first study describes how machine learning can be used to identify Twitter accounts that are distributing propaganda from the so-called Islamic State (IS). The second study aims at identifying references to a narrative containing xenophobic and conspiratorial stereotypes in Swedish immigration critic alternative media sites. The last study focuses on *extreme adopters* of a community-specific jargon on a Swedish immigration critic discussion forum. We are interested in finding out if extreme adopters differ from the rest of the forum users with respect to how they see themselves and if they exhibit certain linguistic features that we view as markers of a *radicalized mind-set* (a certain style of understanding and relating to the world that has often been observed among violent extremists). The rest of this chapter summarizes the work described on Publications I, II, III, and IV (listed in Section **List of papers**).

3.1 Identifying Propaganda from the Islamic State

IS successes in recruiting, as well as inspiring followers can be largely contributed to their skills in producing and disseminating online propaganda. IS has produced a great quantity of propaganda material for different target groups and effectively exploited various digital communication channels for broadcasting their message. IS propaganda can serve as a gateway into a radicalization process, even if propaganda by itself is usually not the only ground for radicalization or recruitment to violent extremist ideologies. One of the most common approaches to stop IS from distributing propaganda is to suspend accounts. In 2016, The Guardian reported that during the previous six months, Twitter had shut down 235,000 IS-friendly accounts that transgressed the company's guidelines concerning the dissemination of terrorism and violent threats [79].

Detecting accounts that distribute IS propaganda requires human analysts to read manually and analyze huge amounts of information on social media. We have studied the ability to automatically detect propaganda from IS on Twitter, i.e., Twitter accounts involved in media mujahideen - the supporters of jihadist groups who disseminate propaganda content online. We use a machine learning approach where we make use of two sets of features: data dependent features and data independent features.

3.1.1 Data dependent features

Data dependent features are constructed from the data and are highly influenced by the specific data set. The subset of tokens or features are chosen based upon their frequency in the data make these feature data dependent features. Some examples of data dependent feature are term-frequency (tf), term frequency-inverse document frequency (tf-idf), the most common hashtags (if the data is from Twitter), most common word bi-grams, most common letter bi-grams and most frequent words in the data set. Using data dependent features could potentially be a problem when running models "in the wild" if the data set that is used for training and testing is not representative of data that exists in the wild. In this work, we are interested in building a model for classifying twitter users that are communicating jihadi content and therefore data dependent features can be valuable. It might be the case that certain hashtags and frequently mentioned words change over time, but many of the data dependent features remain the same and are representative for the group of users that are targeted in this work. The classes of features that we use are common hashtags, common word bigrams, common letter bigrams, and the most frequent words.

3.1.2 Data independent features

Data independent features are the features that are not influenced by the specific data set. Data independent features can be used to estimate how well a model performs on a data set that is different from the training and test data. The data independent features that we have used are stylometric features, a subset of time-based features (e.g., what time or what day a tweet is posted), and emotion-based features. Most of the features are similar for both English and Arabic text.

The reason for using two different sets of features is to investigate if the result depends heavily on features that are specific for a given data set. When using machine learning, there is always the risk that the models that are built are only applicable to the specific data set. By experimenting with both data dependent and data independent features, we hope to get an understanding of the performance of our models in a real scenario.

3.1.3 Related Work

Analyzing terrorist related content on the Internet has been done in several studies. In [1] affect analysis is used to analyze the intensity of emotions in extremist discussion boards. The authors have manually created affect lexicons that are used to measure the usage of violence and hate affects among U.S. and Middle Eastern extremist groups in discussion forums. In [20] the author has analyzed affects in two jihadist discussion boards using machine learning models and a number of data dependent linguistic features like character n-grams, word n-grams, root n-grams, and collocations. Machine learning was used to classify users as potentially supporting or opposing IS in [51]. The authors have collected Arabic tweets referring to IS, and the tweets are then classified into pro-IS and anti-IS. Classification result shows that IS supporters and opposers can be separated with high accuracy using features that are data dependent.

3.1.4 Data set

The data set we have used to train our model is based on a set of 66 users described as "The most important jihadi and support sites for jihad and the mujahideen on Twitter" in a posting on the Shumukh al-Islam forum [31]. We downloaded the latest 3400 tweets from 30 accessible users (the other accounts were suspended by Twitter at the time of downloading). We have also used a set of 45 users that were manually identified to be multipliers of jihadism; all these users are followers of the 66 users and are spreading jihadist propaganda. We also collected a set of tweets containing hashtags that were related to jihadists, in particular IS, and selected users from clusters of

known Jihadist sympathizers [31]. There are 93 English tweeps and 81 Arabic tweeps. A tweep is an individual who has written tweets.

A set of tweeps discussing various topics is also used to train our classification model. To get tweeps discussing various topics we used two approaches. Our study includes both English and Arabic tweets. To collect English tweets, we collected a set of tweets during a certain time period and used some of the tweeps that had written these tweets. For the Arabic tweets, we collected tweeps from a list of Twitter users influential in Arabia and a list of the 100 most influential Arabic female Twitter users. In total, there are 742 English tweeps and 256 Arabic tweeps.

3.1.5 Experiments and results

We have done two sets of experiments where we have used machine learning to identify pro-IS tweets and tweeps on both English and Arabic language data. For each experiment, we use three different sets of features: data dependent features, data independent features, and a mixture of both features. We used stochastic boosting (AdaBoost) with regression trees as base classifiers.

In the first experiment, the goal was to identify English and Arabic Twitter accounts (tweeps) that were pro-IS. While classifying tweeps, the AdaBoost model has been performed with 500 boosting iterations and the rest of the parameters set to default. When classifying English tweeps the result is perfect when data independent features are used and an accuracy of 0.99 using data independent features. The results of the Arabic tweeps are significantly worse. We achieved 92% accuracy, 80% precision, and 94% recall when using data independent features. With data dependent features the accuracy was 98%, the precision 83%, and 100% recall. One of the reasons for getting such a good result could be that the tweeps spreading jihadist propaganda and the random tweeps are totally different from each other in term of expressing views and ideology; a much larger data set would be needed to investigate this further. Another reason for the results could be that the tweets are downloaded during different time periods and there might be a difference in what topics are discussed.

In the second experiment, the goal was to identify English and Arabic tweets expressing support for IS. When classifying individual tweets, the results are not as good as when classifying tweeps (where each tweep had written at least 60 messages). For English tweets, the accuracy was 98% when data independent features were used and 99% when data dependent features were used. Both precision and recall were over 98%. When we tried to classify Arabic tweets, the results dropped. Using data independent features we obtained 82% accuracy and when we used dependent features the accuracy was 85%. The precision decreased significantly to 47% when data independent features were used and 57% for data dependent features.

It is clear that our models work significantly better on classifying English tweeps and English tweets than on Arabic data. The reason could be the complexity of the Arabic language. Arabic is a very specific language, in particular in an orthodox-conservative Sunni Islamic environment, of which groups like IS and al-Qa'ida claim to be part of. Since we have used a small data set in our experiments, it is hard to say anything about how the results would work in a realistic scenario.

3.2 Narratives in Alternative Media

In Sweden, as well as in several other European countries, there has been a recent surge in activity and formation of extreme right movements. These groups show high interactivity on forums and blogs, using the Internet and social media as a means for recruiting and spreading their views. In this study, we have focused on studying a set of alternative media news sites that are critical towards immigration.

Our analysis is focused on studying the occurrence of a narrative by analyzing references to xenophobic and conspiratorial stereotypes. We are also interested in identifying differences in emotional tone and pronoun use in comparison with traditional media. The analysis is done using dictionaries: both dictionaries developed by experts and pre-defined dictionaries from the text analysis tool LIWC (see Section 2.3). The narrative that we study is a common extreme right narrative that is based on a story of conspiracy, including three groups: The Elite, the People, and the Minority. The plot is that in order to gain or preserve power, the Elite uses the Minority against the people. For instance, the elite (liberal politicians and Jewish-owned media) imposes minorities (immigration, multiculturalism) on the people ("real people", Swedish workers, elderly, children), while lying them about the true consequences of immigration (Expo Foundation 2016, personal communication, 3 March 2016). It follows that anyone who accepts this narrative as true is able to construe oneself as a person who has called the bluff and gained insight into the real state of affairs. The English translation of different categories of language variables (LIWC) and narratives that we consider in this analysis are listed in Table 3.1.

To identify the presence of the narrative, we use a dictionary-based approach to find references to the stereotypes that the narrative is based upon. Experienced domain experts from Expo Foundation[1] who have been working in the field and have expertise manually created dictionaries of words used among the extreme right to refer to any of the three groups. The relative frequencies of dictionary words were calculated. We also studied the emotional tone and the use of pronouns using LIWC [62]. The reason for this is that

[1] https://expo.se

Table 3.1. *The different categories we focus on in our analysis and some example words.*

Categories	Example
Language variables	
3rd person plural	they, their, them
Negative Emotions	hate, worthless, enemy, hurt
Narratives	
The Elite	race mixers, anti-swedes
The Minority	luxury immigrants, occupants
The People	nation, people of reality, Swedes

an elevated use of negative emotion-words and third person plural words are features that have been observed in extremist sites [61].

3.2.1 Data set

We collected a number of articles published on online alternative media websites during the year 2015 when there was a top surge of the refugee crisis. The set of alternative media sites that we consider are critical towards immigration. The selection of the site was made by domain experts. All articles published in 2015 were collected. For comparison, we have used the websites of Sweden's largest quality press newspaper (DN) and Sweden's largest popular press newspaper (Aftonbladet). For DN and Aftonbladet we have only collected a sample of articles that were published during 2015. The sites and the number of articles are listed in Table 3.2.

Table 3.2. *The different media sites and the number of articles collected.*

Media group	Name of media	No. of articles
Alternative media	Nordfront	619
	Avpixlat	4391
	Exponerat	6239
	Fria tider	4856
	Nyheter idag	1400
	Samtiden	2632
Press newspaper	DN	1747
	Aftonbladet	613

3.2.2 References to the stereotypes

Our analysis shows that references to the stereotypes that the narrative is based upon occur significantly more in the alternative media than in DN and Aftonbladet. Looking at each alternative media site individually, we can observe a variance in the frequency of words referring to the characters in the extremist

right narrative. This indicates that that alternative media should not be classified as one type of media. The most frequent use of references to the narrative can be found in Nordfront - a media source connected to the national socialistic movement the Nordic Resistance movement.

Our analysis also showed that there are significant differences between regular media and immigration critic alternative media. The results from a t-test [52] on the two different groups can be seen in Table 3.3.

Table 3.3. *Significance test for alternative media compared with non-alternative medias.*

Category	p-value	Significant
The Elite	7.809190e-13	Yes
The People	4.388677e-03	Yes
The Minority	8.649609e-26	Yes
Negative emotion	2.683428e-03	Yes
Third person plural	7.961904e-03	Yes

When it comes to the emotional tone and pronoun use these two media groups are significantly different. Overall, immigration critic alternative media are more negative than in regular media. There is a higher use of third person plural on the alternative media sites, which can be a sign of outgroup-focus or even xenophobia. However, it can also be regarded as an artifact of the focus on immigration and immigrants that prevails in all the alternative media considered here.

3.3 Identifying Extreme Adopters in a Discussion Board

The words that we use when communicating in social media can reveal how we relate to ourselves and to others. For instance, within many online communities, the degree of adaptation to a community-specific jargon can serve as a marker of identification with the community. We have singled out a group of so-called *extreme adopters* of community-specific jargon from the whole group of users of a Swedish discussion forum devoted to the topics immigration and integration. We define an extreme adopter of a community-specific jargon as a person who uses substantially more jargon-words or expressions than what is the norm of the community in question. Among extreme adopters, there are both high-status members who are very active in the community and influential in creating the jargon, as well as more peripheral members whose extensive use of jargon reflects their eagerness to fit into the community. What both these groups have in common is a high degree of identification with the community.

3.3.1 Data set

In this study, we wanted to test a number of research hypotheses on a subforum of a Swedish web forum called Flashback: Integration och invandring[2] *(In English: Integration and immigration)*. The forum is characterized by a certain xenophobic jargon. The data that we have collected consists of around half a million posts that were posted between 2007-01-01 and 2015-04-27.

3.3.2 Generating community specific jargons

We have manually created a customized dictionary of community-specific jargon. We choose 14 posts (> 50 words) from 7 different users in different threads from the data set. To make sure that the posts were written by established members (people who could be assumed to be aware of the jargon), only posts written by users who had been registered for more than one year and had written at least 500 posts before were considered. From these posts, all words or strings of words (n-grams) that could be counted as jargon were selected for the new dictionary. The criterion for counting a word or n-gram as jargon is that it is well understood by other community users, i.e., can be used without further explanation, and either (1) not usually found in everyday discourse, or (2) used with a different connotation than in everyday discourse. From manual analyses, we found that these words were derogatory words referring either to ethnic minorities or purportedly immigrant-friendly politicians and media. The words ranged from very crude and insulting ones to words with ironically reversed meanings (e.g., "cultural enrichment"). Some of them were neologisms (e.g. "race-mixers"), others were regular words used metaphorically (e.g. "locusts"). The list of community-specific jargon consists of around 150 words in their base form.

3.3.3 Group identification and a radicalized mind-set

For each user in the web forum, we have summarized the total number of posts written by the user. We are only interested in the forum users that have a high use of the community-specific jargon - the so-called extreme adopters. The set of extreme adopters are selected by considering the term frequency of the community-specific jargon described in Section 3.3.2. This means that for each user, the occurrences of a word from the community-specific jargon were divided with the total number of words used. We selected the set of users that had a term frequency higher than 0.005. The threshold was selected using manual inspection. For future work, we intend to improve this process.

A total of 583 users had a high usage (>0.005) of the forum specific jargon in our data set. We denote theses users as extreme adopters. There is

[2] *https://www.flashback.org/f226*

a noticeable difference in the average use of forum specific jargon between the extreme adopters and the rest of the forum. The group of extreme adopters used an average of 179 forum specific jargon words while the rest of the forum users used an average of around 3 forum specific jargon words.

The group of extreme adopters is compared to a normal group that consists of a total of 49,585 users (the whole sub-forum). We wanted to test two different research hypotheses. The first research hypothesis **(H1)** that we are interested in testing is if the group of extreme adopters differs from the normal group with respect to identity in such a way that they use:

I. less first person singular than the normal group
II. more first person plural than the normal group

How a person perceives oneself can be revealed by the use of first person pronouns. Pronouns are good indicators of social perceptions since they can only be correctly understood when the speaker and listener share a common knowledge of what they refer to. A speaker's use of pronouns conveys some information about the speaker's perception of shared knowledge with the listener. The psychometric properties of pronouns have been validated in several studies where pronoun use has been linked both to different emotional processes and to social ones, such as self and identity, stereotypes and intergroup evaluations [60]. For instance, when people use less first person singular words in the context of a community, it often marks a sense of affiliation with the community [61]. Research on online behavior shows that new members of an online group will use less first-person singular pronouns over time, a change accounted for by increased identification with the group [29]. Correspondingly, an increase in first person plural, *we*, *us* , *our* etc., can often be observed as people spend time together [60]. "We-words" are predictive of group cohesion.

The second research hypotheses **(H2)** that we want to test is if extreme adopters of this jargon also exhibit certain linguistic features that we view as markers of a *radicalized mind-set*. A radicalized mind-set is a certain style of understanding and relating to the world that has often been observed among violent extremists. This cognitive style is marked by conspiracism, rigid binary thinking, a strong differentiation between *us* and *them*, and a sense of grievance and injustice. Having a radicalized mind-set does not necessarily mean one is at risk for joining an extremist group or perpetrating acts of radical violence, but without it is unlikely that a person would engage in radical action. In other words, a radicalized mind-set could be seen as a necessary, but not sufficient precursor of radical violence.

To test our two hypotheses, we use a number of different LIWC categories. For the nine selected LIWC-categories, we compared how much they are used by extreme adopters to how much they are used on the whole forum. Table 3.4 shows the different dictionaries including the LIWC categories that we have used along with some sample words from each category.

Table 3.4. *The LIWC categories used in our analysis along with some sample words.*

1st person singular	I, my, me
1st person plural	we, our, us
3rd person plural	they, their, them
Positive Emotions	happy, pretty, good
Negative Emotions	hate, worthless, enemy
Anger	hate, kill, pissed
Inclusive	with, and, include
Exclusive	but, except, without
Power	superior, bully
Forum specific jargon	race-mixers, enrichment

3.3.4 Summary of the result

First, we tested whether there was a difference regarding how they used first person pronouns, and it proved that extreme adopters used less I-words and more we-words than forum users in general. This supports our assumption **(H1)** that these users actually do tend to identify more with the group.

Second, we tested whether the extreme adopters differed from the whole forum regarding those LIWC-categories that we had pre-defined as markers of radicalization; third person plural, emotionality, anger, inclusive words, exclusive words, and power words. Our hypotheses **(H2)** were confirmed insofar as extreme adopters also used more third person plural, anger, and power words [37] [60]. They used less exclusion words and more inclusion words, indicating a cognitive style characterized by a low level of reasoning or analytic thinking [28]. In one aspect, emotionality, our hypotheses were not entirely confirmed. The group of extreme adopters was less emotional overall than forum users in general. There was no difference regarding negative emotion words; however positive emotion words were less frequent among the extreme adopters than among forum users in general. The relative lack of emotionality (the difference in anger words was significant but small) suggests that these users are not more emotionally engaged than others in the issues being discussed. This finding is in line with [53] which proposition that the mechanisms of group conflict and radicalization are not perpetuated by emotion, but rather by perceptions of identity relative to different groups.

3.3.5 Separating extreme adopters using machine learning

Since our two research hypotheses **(H1)** and **(H2)** were proved, we decided to use another approach to investigate different aspects of extreme adopters. We used machine learning to investigate the possibilities to separate extreme adopters from the rest of the discussants, and which features play an important role in the classification. In the experiments, we have used two different clas-

sifiers: support vector machine (SVM) and random forest (RF) with 10-fold cross-validation. There were 583 extreme adopters and 700 forum users. In the classification, the set of extreme adopters are treated as the negative class, and the rest of the forum users are treated as the positive class. The data set used for experiments is scaled and centered to zero mean and unit variance. For SVM, the radial kernel is used while training and predicting. The cost of constraints violation is kept to 10 while the rest of the hyper-parameters are used as default. SVM models were created using e1071 R package and RF models were created using randomForest R package[3].

In the classification, we used both data dependent features such as most frequent words, most frequent bigram words and most frequent bigram letters and data independent features. The reason for using data independent feature set is since we want to investigate if there are some features that can be used to identify extreme adopters regardless of what digital community they belong to and what topics are discussed. The results show that using a Random Forest model it is possible to separate extreme adopters from the rest of the discussants with more than 80% accuracy using data independent features and over 86% using data dependent features. This is an interesting result since it indicates that extreme adopters have a different language compared to other discussants not only because they use more community-specific jargon.

When looking at the features that played an important role in the differentiation between the two groups, we can see that in the case where we used data dependent features racist slurs and the word *white* played an important role in the classification. This is not surprising since the extreme adopters were separated from the rest of the discussants by their high use of a forum specific jargon that included many racist slurs. A more interesting result is the features that played an important role in the classification when we only considered data independent features. In this case, the use of pronouns played an important role. In particular the use of first person singular (e.g *I*, *me*, and *mine*) and the use of third person plural (e.g. *they*, *their*). Pronouns play an important role in our language, and they can be seen as indicators of social perception since they can only be correctly understood when the speaker and listener share a common knowledge of what they refer to.

[3]The packages are freely available from http://CRAN.R-project.org/

4. Risk Assessment of Written Communication

Violent lone offenders pose a threat to modern society and are a serious problem for law enforcement and security services around the world. One of the challenges with lone offenders, compared to terrorist groups, is that they do not need to communicate with others. This makes it much harder for intelligence services and authorities to intercept possible communication, something that could help to identify and capture potential offenders. Another challenge with identifying potential violent lone offenders is their diverse background. For example, they could represent any kind of ethnicity, ideology or subculture [47]. The lack of a common profile among violent lone offenders significantly decreases the possibility of identifying them [9].

In the present work, a violent lone offender is defined as an individual that acts alone (or with a partner) and is not formally connected with an organization. This means that the individual may have been inspired or encouraged by an organization to act without receiving a direct order from a representative for an organization. The process of the attack should include planning, and the motive should not be self-interest or material gain. This definition of a violent lone offender includes school shooters, mass murderers, and ideologically motivated individuals.

Naturally, identifying lone offenders before an attack is fundamental for security agencies and law enforcement officials. A common approach to identify high-risk individuals is to use a risk assessment protocol or a risk assessment instrument. There are a variety of risk assessment protocols available, some of these assess the risk of violent behavior among individuals who already committed violent acts while others assess the risk of violent behavior among first-time or potential offenders. Risk assessment of individuals is commonly done in prisons, hospitals, or when in contacts with social welfare consultants. In the present research, we introduce a method of risk assessment based on text analyses. This method focuses on textual communication by risk individuals. Thus, we use their own words to conduct a risk assessment. Previous research shows that lone offenders signal their upcoming attack by writing about it in, for example, online diaries, manifestos, and discussion forums. In a study by Gill, Horgan, and Deckert [33] it was found that 60% of lone offenders in the study expressed their views and ideologies in written messages. By studying writings from identified lone offenders, we can learn about, for example, their personality, their emotional state, and self-image. The results from such studies can be used to develop new risk assessment tools.

The research conducted within the framework of this thesis focuses on how new technologies based on written communication can assess the risk of an individual that plans to commit targeted violence. In addition, a tool called Profile Risk Assessment Tool (PRAT) was developed to assist analysts, researchers and law enforcement professionals. The tool assesses a text based on a number of different variables. The choice of variables is based on previous research and theory regarding risk behavior but also core personality characteristics that constitute the individual's psychological fingerprint. The extracted variables are aimed to provide a profile that can be compared to, for example, known cases of lone offenders or a theoretical risk profile extracted from a wide range of targets in a variety of populations.

The remaining part of this chapter introduces some of the technologies that can be used in risk assessment based on written communication by presenting two studies that we have conducted. The first study focuses on identifying linguistic markers in texts written by lone offenders and the second study focuses on the possibility to detect (or correctly identify) texts written by known/previously identified lone offenders. The studies also address the problem of imbalanced data sets and the small sample size problem in machine learning. The studies are described in more detail on Publications V, VI, and VII (listed in Section **List of papers**).

4.1 Warning Behaviours for Risk Assessment

One of the difficulties of working with security-related problems is the lack of relevant data sets. Even if there is a possibility to use different kinds of text analysis techniques to assess risk in the text, the technologies cannot be tested accurately due to the lack of realistic data. Attacks carried out by lone offenders are rare, and it is not always the case that violent lone offenders publish a manifesto or written text that is available for research. Thus, we have based our research on a small data set consisting of written communication from violent lone offenders. The data consists of texts that have been made public, either by the offenders themselves or by law enforcement authorities.

Previous research has shown that violent lone offenders show signs of psychological warning behaviors that can be viewed as indicators of an increasing or accelerating risk of committing acts of violence. Warning behaviors are defined by Reid, Hoffman, Guldimann, and James in [66] as any behavior that "precedes an act of targeted violence, is related to it, and may, in certain cases, predict it". The eight different warning behaviors are described in Table 4.1. Some of Meloy's warning behaviors (leakage, fixation, and identification) have also been identified as potentially observable in social media communication [23].

An approach to assess written communication is described by Brian Van Brunt in [14]. Van Brunt presents a protocol for violence risk assessment of

written words. The protocol is used to assess written text such as emails, letters, or creative writing that contain direct threats or violent themes. Five different factors and corresponding sub-factors are used to assess written communication. The factors are fixation and focus, hierarchical thematic content, action and time imperative, pre-attack planning, and injustice collecting.

Table 4.1. *Different warning behaviors described in [66].*

Warning behaviour	Description
pathway	behavior that is part of research, planning, preparation, or implementation of an attack
fixation	behavior that indicates an increasingly pathological preoccupation with a person or a cause
identification	behaviour that indicates a psychological desire to be a "pseudo-commando" or have a "warrior mentality"
novel aggression	act of violence which is committed for the first time and is unrelated with previous attacks
energy burst	increase in the frequency of activities related to the target
leakage	the communication to a third party of an intent to do harm to a target through an attack
last resort	increasing desperation or distress through declaration in word or deed. The attacker feels that there is no alternative other than violence, and the consequences are justified
directly communicated threat	a written or oral communication of a direct threat to the target or law enforcement before attack

4.2 Violent Lone Offenders Written Communication

Several lone offenders have communicated with their surroundings using everything from lengthy manifestos to short posts on social media platforms. By studying their communication, we can learn more about the individuals behind the deeds: their mind-set, their emotional state and how they see themselves and others.

4.2.1 Characteristics of violent lone offenders manifestos

We have studied publicly available communication including emails and manifestos from violent lone offenders to find unique characteristics in their writing, compared to texts written by individuals from the general population. To this end, we have identified eight different LIWC-categories that we consider

to be in line with "a terrorist mind", as it is defined in the literature. Examples of common traits of terrorists are low cognitive flexibility and low tolerance for ambiguity [77], a trait that we hypothesize can be reflected in high use of words in the category "certainty". Additionally, an "extraordinary need for identity, glory, or vengeance; or a drive for expression of intrinsic aggression", is also considered common among terrorists [77], something that we suggest can be reflected in language as elevated frequencies of power and anger-words.

The LIWC categories that we focus on are:

- Use of big words
- The use of personal pronouns (third person plural)
- Expressions of emotion (positive, negative and anger)
- Social Processes - friends
- Cognitive processes - certainty
- Drives - power

The subjects that we included in our analyses comprise ten different lone offenders. The subjects are listed in Table 4.2. These individuals have conducted targeted violence and communicated information before their acts. For comparison, we used a baseline sample of 714,000 texts from LiveJournal.com and Blogs.com. The baseline is described in [64]. Statistical t-tests were performed to assess to what extent the lone offenders differ from the baseline sample on the individual level. The results showed significant differences between the two baseline samples and lone offenders for some categories. In particular, texts written by lone offenders had lower frequencies of positive emotion and friends and significantly higher frequencies of negative emotion, anger, power, certainty, third-person plural, and big words.

Table 4.2. *Lone offenders and statistics on their communicated text.*

Name	No. of words	Publication year	Age when writing
Nidal Malik Hasan	3555	2008	38
James von Brunn	47178	1999	79
Anders Behring Breivik	807712	2011	32
Dylan Roof	2446	2015	21
Elliot Rodger	108206	2014	22
Christopher Dorner	11489	2013	34
Jim David Adkisson	1056	2008	58
Ted Kaczynski	34719	1995	52
Lucas Helder	3296	2002	21
Andrew Joseph Stack III	3236	2010	54

4.2.2 Detecting violent lone offenders manifestos

Motivated by the identified differences in the communication written by lone offenders compared to the general population sample we decided to investigate

the possibility to use machine learning to automatically detect texts written by lone offenders.

Before conducting these analyses we extended the data set of lone offenders to include a total of 32 subjects that have written a total of 46 different texts. The subjects are a combination of school shooters, ideologically motivated offenders, and mass murderers. As a comparison/control sample, we collected data from a number of different digital environments. Specifically, we used text from Google blogs, where most of the blogs are about personal interests, news, fashion, and photography. We also used text from the white supremacy discussion forum Stormfront. Stormfront was founded in 1995 and is commonly described as the leading white supremacist web forum. The forum has grown into what may be the Western world's most popular forum for racists, Holocaust deniers, and criminals to post articles and engage in discussions and share news of upcoming racist events [10]. Finally, we included data from the Irish discussion board Boards.ie. Boards is a public forum where hobbies, politics, and sports are discussed. The data sets are listed in Table 4.3. The reason for choosing three such different data sets to represent texts that are not written by violent lone offenders is that we want to capture a fairly representative sample of social media instances where violent lone offenders might publish texts prior to an attack.

Table 4.3. *The different data sets used in the experiments.*

Data set	**No. of written communication**	**Description**
LO	46	Texts written by violent lone offenders
Blogs	54	Blogs about personal interests, news, fashion and photography
Stormfront	108	Posts written by a set of Stormfront users
Boards	108	Posts written by users from the Irish forum Boards.ie

The features used to conduct our classification task were the different LIWC categories; however, not all features necessarily contributed to the classification. Feature selection was done to get the optimal subset of features that contribute most to the classification with minimal error rate. Feature selection is the process of selecting a subset of maximally informative features and removing redundant or irrelevant features that do not provide any information [11]. One of the benefits of using LIWC categories as features is that they assure features that are data independent. This is particularly important in cases where data sets are small and unbalanced (as in this case). Since the size of the data set is small, we used a Leave-One-Out Cross-Validation (LOOCV) [67] approach while building a machine learning model. The model was built

leaving out a single sample, which was later used to derive a prediction for the left-out sample.

The data set is not only small; it is also imbalanced. An imbalanced data set means that there is an uneven distribution of the training set over the classes. In the case of classification, a data set is imbalanced if the presence of one class is considerably higher (majority class) than another class (minority class). One of the complications that arise due to imbalance is the effectiveness of accuracy while determining the performance of a classifier. Usually, the classification models built with balanced data underperform when the test data is unbalanced [39].

To overcome the imbalance, we used Synthetic Minority Over-sampling Technique (SMOTE) [18]. SMOTE is an over-sampling approach where the minority classes are over-sampled by creating synthetic examples rather than by over-sampling with replacement. The algorithm selects two or more similar instances using a distance measure and re-samples an instance's attribute individually by a random amount within the difference to its neighbor. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. We conducted four different experiments to get an understanding of the possibility to use machine learning to identify texts written by violent lone offenders from a normal population.

Three experiments were conducted on three different data sources, and one separate experiment was done where all the data are combined. One of the goals was to use different data sources to identify factors that separate the texts of violent lone offenders from other populations. The first experiment attempted to separate texts written by violent lone offenders from blogs, texts written by Stormfront forum users, and texts written by Boards.ie users. The results showed that using only the 11 most important features we correctly classified 35 out of 46 texts written by lone offenders and 242 correctly classified texts from the blogs and forums out of 270. The features that were used for the classification shows that the use of pronouns and personal pronouns played an important role in the classification. Two other linguistic dimensions played an important role in the classification, namely articles (e.g., *a*, *an*, & *the*) and prepositions (e.g., *to*, *with*, & *above*). Another feature that played an important role in the classification is negative emotions, in particular, anger that is a subcategory to negative emotions in LIWC. Other features that played an important role in the classification are psychological factors (corresponding to categories in LIWC) like Perceptual processes (e.g., $look, heard, feeling$), See (e.g., *view*, *saw*, *seen*), Differentiation (e.g., $hasn't$, *but*, and *else*), Affective processes (e.g., *happy*, *cried*) and Biological processes (e.g., *eat*, *blood*, *pain*).

In the second experiment, we aimed to separate texts written by violent lone offenders from bloggers. The (LIWC) linguistic dimensions that played an important role in the classification were Prepositions, Function words, Impersonal pronouns, Personal pronouns, Third person plural, Time orientations

(Time & Relativity) and Personal concerns (Leisure, Death, & Informal language). Psychological factors that played an important role were Perceptual processes (e.g., *look*, *heard*, *feeling*), Tentative (e.g., *maybe*, *perhaps*), Certainty (e.g., *always*, *never*), along with negative emotions and anger. Further investigation on how these psychological factors can be related to previous theories on warning behavior still needs to be conducted.

In experiment 3, we attempted to separate texts written by violent lone offenders from texts written by Stormfront forum users. The (LIWC) linguistic dimensions that played an important role in the classification were Article, Quantifiers, Prepositions and Personal pronouns. The psychological factors that played an important role in the classification were Differentiation, See, Biological processes, Cognitive processes and Negative emotion.

In experiment 4, we aimed to separate texts written by violent lone offenders from texts written by Boards.ie users. We found the linguistic dimensions that played an important role in the classification were: Article, Auxiliary verbs, and Personal pronouns. Personal concerns in the form of Assent and psychological factors such as Differentiation, See, Affective processes, Social processes, Negative emotions, and Anger also played an important role in the classification.

The linguistic dimensions that play an important role in all experiments are Personal pronouns and Prepositions. The use of pronouns in natural language has been linked to different aspects of personality and emotion [60]. For example, frequent use of third person plural (*they*, *them*, etc.) in a group suggests that the group is defining itself to a large degree by the existence of an oppositional group [61]. Why prepositions (e.g., words like *to*, *with*, & *above*) are important features needs to be investigated further.

Negative emotions also played an important role in the classification, in particular, anger (experiment 1, 2, and 4). When Pennebaker and Chung [61] studied al'Qaida-texts, they discovered that texts were relatively high in emotion and that the relation between positive and negative emotion differed from what is common in natural conversation. The natural conversation usually contains almost twice as many positive than negative emotion words while the al'Qaida-texts had a much higher relative degree of negative emotion words, mostly anger words [61]. One important feature when separating texts from Blogs and Stormfront from the texts written by the violent lone offenders is anger (with anger words such as *hate*, *kill* & *annoyed*).

Features about personal concerns only mattered when separating blogs and Boards.ie texts. This might reflect the fact that personal concerns are expressed more in blogs and in the discussion forum Boards.ie than on Stormfront forums where ideology is discussed more.

4.2.3 Summary of the results

The experimental results show that some LIWC categories are used significantly differently by lone actors in their communication compared to a baseline sample. When comparing the texts written by the 10 lone offenders with a set of blog texts, we found that the texts written by the lone offenders had significantly lower frequencies of positive emotion and friends and significantly higher frequencies of negative emotion, anger, power, certainty, third person plural, and big words. The identified categories, or linguistic indicators, seems to be in line with previous research [61] on what characterize violent extremist and lone offenders.

When using a machine learning approach, we were able to separate a set of lone offenders from a normal population (combination of Blogs, Stormfront, and Boards) with an accuracy of 87%, a specificity of 76% and a sensitivity of 89%. The results are promising, but it should be noted that the data set is very small and therefore we should be careful and await further research before drawing any conclusions. The data set used in the experiment is balanced in term of a number of lone offenders and normal population instances. It is a fact that the magnitude of positive cases (normal population) will be larger in the wild than our experiment scenario. So, the specificity is particularly unlikely to be able to be maintained at this level in real life applications.

4.3 PRAT - Profile Risk Assessment Tool

To be able to test if our findings hold in practice we have created a tool called PRAT (Profile Risk Assessment Tool). PRAT is used for risk assessment of written communication. PRAT extracts a profile consisting of 30 personality and risk-behavior related variables from a text. The extracted profile is compared to 27,834 profiles including known cases of violent lone offenders, school shooters, and social media users from various sources (Google blogs, Stormfront, Reddit, Islamic Awakening, and Boards). The choice of variables is based on previous research and theory regarding risk behavior but also core personality information that constitutes the individual's psychological fingerprint.

Figure 4.1 shows a sample of screenshots from PRAT. As can be seen, PRAT provides a word cloud for quick summarization of the text content. PRAT also presents values for each of the 30 variables that are measured (e.g., the emotionality variable anger) followed by percentile score. The percentile in the case in Figure 4.1 means that the expression of anger in the text (or for that individual) is higher than 95% of all other texts (individuals) in the entire sample.

Figure 4.1. Screenshots of the profile risk assessment tool (PRAT).

5. Identifying Internet Users with Multiple Aliases

In this chapter, we present research done with the aim to identify users with multiple aliases on social media. There are many different reasons for a user to use multiple aliases, for example, it could be the case that an old alias has been deleted due to inactivity or the password has been forgotten. It could also be the case that a moderator has banned an old alias. Another reason could be that a user wants an extra alias to support his or her arguments to cause debate or controversy. The use of multiple aliases can be divided into two main cases that are particularly interesting for using multiple aliases in public fora such as discussion boards, web blogs, social media, or web pages. The first one is the "non-concealed" case where a user creates a new alias without any attempt to disguise the fact that multiple aliases have been created by the user. The second case is "concealed" case where the user does not want to reveal that several aliases belong to the same individual and wants to hide his/her real identity. The approaches for addressing the two different cases differ. In the non-concealed case, a similar username and profile picture might be a strong indicator that two accounts belong to the same individual while in the concealed case this is most likely not the case.

When trying to detect individuals who are using multiple aliases, several kinds of approaches and techniques may be considered. Our research focuses on solving two different problems: **author identification** and **alias matching** (or similarity detection). In the author identification problem, we compare an anonymous user to a fixed set of candidates. We assume that the anonymous user is in the list of candidates. In an alias matching setting, we cannot assume that we have knowledge of all candidates. Instead, we are interested in finding the candidate that is most similar to the anonymous user. Alias matching can be used to group all candidates that are more similar than a certain threshold value. Hence, while author identification can be seen as a supervised machine learning problem, alias matching is an unsupervised problem where the same supervised algorithms cannot be used directly. We have used a number of different approaches for solving the author identification problem and the alias matching problem. The approaches that we purpose use different features such as stylometric features, time features, and expression of emotions. We have implemented and tested our approaches, and the experimental results indicate that there is a possibility to identify users with multiple aliases - at least in some cases.

In the rest of the chapter, we introduce the concept authorship analysis and explain its use in forensic linguistics, i.e., the linguistic investigation of authorship for forensic purposes and some related work done in the field of authorship analysis. We describe the advantages of *time-based feature* over *stylometric feature* for authorship identification and alias matching. Finally, we summarize techniques and experiments that we have done in the area of author identification and alias matching along with a discussion of the experiment results. The results summarized here are based on the publications VIII, IX, and X (listed in Section **List of papers**).

5.1 Authorship Analysis

Authorship analysis is a field of research which focuses on examining the characteristics of a text in order to draw conclusions on its authorship as well as characteristics of the author. Authorship analysis can be divided into three different categories: authorship identification, similarity detection, and authorship profiling [12]. Authorship identification or authorship attribution determines the author of an unknown text by comparing it with previous texts produced only by the same author. In machine learning, this is viewed as a text classification problem, and it can be considered as a multiclass single-label text classification problem where the author is a class of a given text. Alias matching or similarity detection measures the degree of similarity between two texts rather than identifying the actual authors of a text. The goal is to determine whether two texts are written by the same author or not - without knowing the real author. Similarity detection is often used in the context of plagiarism detection. Authorship profiling or characterization identifies the characteristics of an author that produced a text. The author's characteristics could be gender, age, occupation, educational and cultural backgrounds, native language, etc.

Authorship analysis is a well-studied problem, where algorithms and various features have been extensively described in [3] [30] [43] [58] [83]. The practical applications of authorship analysis could be identifying the authorship of emails and electronic messages [2] [6], plagiarism detection in student essays [75], and forensic cases [17]. In our work, we have focused on authorship identification and alias matching.

5.2 Related Work

In the past, various approaches to authorship analysis have been used. During the 18th century, the English logician Augustus de Morgan suggested that authorship can be identified by determining the use of word length between the documents [82]. In 1964 Mosteller and Wallace [57] used a Bayesian ap-

proach and stylometric features to 11 of the 84 Federalist Papers to identify the actual author of the papers. The Federalist Papers were anonymously published by Alexander Hamilton, John Jay, and James Madison in 1787-1788. By comparing the use of function words in the papers, they concluded that all of them were written by one particular author who had written the majority of the other Federalist Papers. In the conviction of Ted Kaczynski, also known as the Unabomber, authorship profiling was used [69]. Kaczynski was responsible for a nationwide bombing campaign in America between 1978 and 1995 that killed three people and injured 23 others. In 1995 Kaczynski released a manifesto entitled *Industrial Society and Its Future*. The writing style was recognized by Kaczynski's brother. The linguistic analysis of the manifesto and a comparison with some of the letters that Ted sent to his brother eventually resulted in an arrest of Kaczynski.

Over the year the approaches and quality in authorship identification have improved significantly. Researchers use machine learning algorithms and complex artificial neural network models for author identification. A scientific event called PAN [65] has shared tasks on digital text forensics and stylometry and invites researchers and practitioners to work on a specific problem of interest. Authorship identification is usually one of the tasks at PAN and researchers use state-of-the-art techniques and algorithms to solve the problem and to get the best results. Ahmed et al. [56] used deep learning for solving the author identification problem. A stacked denoising Auto-Encoder was used to extract features, and then a support vector machine classifier was used for classification. In [27] an ensemble approach is used for authorship identification. The approach combines predictions made by three independent classifiers built using char n-grams, char n-grams with non-diacritic distortion [73] and word n-grams. The authors have applied text distortion and then extract character n-grams to highlight the use of punctuation marks, numbers, and characters with diacritics (e.g., ó, é, etc.). Halvani and Graner [34] used a compression-based cosine similarity distance measure for the author identification task. The compression-based algorithm estimates the amount of information shared by any two objects. With this approach, there is no need for defining features; instead, the compression algorithm performs the feature engineering procedure.

Narayanan et al. [58] did an experiment on Internet-scale authorship identification. By using a small sample of blog posts, they tried to identify the rest of the posts written by the same author, mixed in with 100,000 other blog posts. Their algorithms ranked the possible authors in descending order of probability, and the top guess was correct about 20% of the time. If the classifier was given the option of not making a guess, the precision of the top guess was increased to over 80%. Authorship attribution can be applied to languages other than English. Abbasi and Chen [2] used linguistic features and a support vector machine classifier for authorship identification on both Arabic and English web forum messages.

The problem of "anti-aliasing" in the context of authorship analysis is studied in [59]. Anti-aliasing refers to linking multiple aliases to known individuals based on their posts in public fora. The matching is based on the used vocabulary (i.e., what different words that are used) which make it relies heavily on the topic. Anti-aliasing is therefore not as suitable for heterogeneous topics.

Although there are many different approaches to author identification and similarity detection suggested in the research literature, to the best of our knowledge there are no previous attempts to make use of publishing times of posts for author identification and similarity detection.

5.3 Features for Author Identification and Similarity Detection

We have used several sets of feature for both author identification and similarity detection. The features used are stylometric, time-based, and emotion-based.

5.3.1 Stylometric features

Stylometric techniques draw conclusions of text authorship by examining different properties of a text like linguistic features, patterns of language preferences or most common words used, and structural characteristics. Previous researches [3] [30] [58] [83] have shown that stylometric features can be used for authorship identification as well as for similarity detection. When stylometric features are used for author identification, there is an underlying assumption that an author tends to write in a relatively consistent, recognizable and unique way. The author's writing style is analyzed by constructing a "writeprint", which can be used like a fingerprint.

Several algorithms and different features for stylometry-based author identification have been proposed in the literature [3] [72]. In addition to using existing features from previous work, we have also included various emoticons as features (the relative frequency of various smilies). Other features could have been used, including lexical features such as vocabulary richness (e.g., using frequency of hapax legomena i.e., once-occurring words or Yule's K measure [80]), syntactic features such as part-of-speech tags, n-grams, and idiosyncratic features such as misspelled words. We are not arguing that we have used the richest set of features possible, but rather that we have incorporated a lot of useful features that can be extracted from text reasonably quickly. The stylometric features used for authorship attribution and similarity detecting are listed in Table 2.1.

5.3.2 Time-based features

Most existing author analysis approaches rely on linguistic features. However, authorship analysis can be circumvented if the author is changing writing style. There are methods similar to those suggested in [5] [13] [48] [54], which can be used to obfuscate regular writing style intentionally or to imitate someone else. These methods facilitate an individual to change his or her writing style. Similarly, there are techniques like Round-trip machine-translation [13] and special software like Anonymouth [54] that is used to detect stylometric patterns that the user should remove/add to help obscure their style and identity. Syntactic features of a text, like function words and content specific features, are language dependent. This could be a problem in cases when several different languages are mixed. In addition to the language problem, the size of the text could be an issue in author identification since stylometric techniques are computationally slow (as found in our experiment). To address the writing style obfuscation problem, the language problem and the efficiency problem, we have studied how time based features can be used in author identification and similarity detection. In some cases, the publishing time of a post is the only thing that is present, in particular in cases where users post images or videos with illegal content. Another reason is that an individual's timeprint is most likely uncorrelated to his or her writeprint, making it possible to combine stylometric and time-based feature into a more powerful feature set.

Looking at the point in time when an individual is active and inactive in digital environments can give important clues to identify them. A user's activity profile is based on his or her activity peak(s) and period(s) of inactivity. An activity peak is a time period when the user is most likely to create a post, and a time period when the user is least likely to create a post is a period of inactivity. Research [36] has shown that individuals act according to their chronotypes as either 'morning types' or 'evening types'. 'Morning types' tend to get up early in the morning and be more active early in the day, while 'evening types' tend to sleep late and become active in the evening. A morning type person is typically most active early in the day and gradually grow more tired, whereas an evening type person will have his/her peak time sometime during the afternoon or early evening. An individual's chronotype tends to be stable over time [50]; i.e., it does not change significantly in a short period of time. The fact that a person's chronotype does not change significantly over time makes it reasonable to consider the observed activity of a user as a stable measure over time for a given individual.

As mentioned before, stylometric features can be referred to as a user's "writeprint" in the sense that it can be compared to a fingerprint when it comes to identifying a person. Similarly, we use the term "timeprint" when referring to how various time features can be used to identify a person. A timeprint can be seen as a property that reflects something about the characteristics of an

individual's activity and habits. The time-based features are listed in Section 2.2.2.

5.3.3 Emotion-based and Twitter specific features

Other feature that we have used for similarity detection is emotion and Twitter specific features. A reason for incorporating emotion and Twitter specific features is if traditional stylometric features are expected to yield a bad performance when applied to tweets. The emotion and Twitter specific features are listed in Table 2.2.

5.4 Techniques for Author Identification

When solving the author identification problem, an anonymous user is compared to a fixed set of pre-defined known entities. In this way, we assume that the anonymous user is one of the candidate authors present in the exhaustive lists of candidates. In our research, we have used time-based features in combination with supervised learning algorithms to solve the author identification problem. In supervised learning, the algorithm builds a mathematical model based on training data that contains both the input (social media communication with time and date) and the desired output (the author of the text).

5.4.1 Data set and experimental setup

To get an understanding of the performance when using time features, we have conducted a number of experiments using data from a discussion forum. The data was obtained from ICWSM[1] scientific conference. While we would have preferred to conduct experiments on real-world data in which a subset of the users was known to have multiple accounts, we face the fact that such data is very hard to obtain. Instead, we simulate a scenario with multiple users by dividing posts of a user into five separate "sub-users" i.e. each user u_i has been split into five "sub-users" $u_{i1}, u_{i2}, \ldots, u_{i5}$. Each post contains the publishing time that can be used to create a timeprint for each user. The reason for using five sub-users is that we will construct several training instances for each user to facilitate the learning phase in our supervised learning experiments. The userID u_i is the target class for each feature vector, and there are five different instances for each userID. Based on the publishing time a scaled and centered time-based feature vector is constructed for each sub-user. A total of n feature vectors for n sub-users where n is the total number of sub-users in the simulated data set.

[1] https://www.icwsm.org/

It is hard to know exactly how someone would make use of several accounts in social media. Would they first make a post using one of their accounts, then switch to a second, and so on? Would they first write a large number of posts using one account, then switch to the next, and so on? We have investigated the problem by simulating a scenario. We did two experiments using different approaches to divide the posts among the sub-users. In one experiment the posts were divided randomly among the sub-users and in another experiment the posts were divided sequentially rather than randomly (i.e., where the user's first post has been assigned to u_{i1}, the second post to u_{i2}, etc.).

In the experiments, we select 4000 users who have been the most frequent posters in the discussion forum. Next, we randomly picked 1000 users out of 4000 to introduce randomness in our data set. In the experiment, we varied the number of potential users from 200 to 1000 in steps of 200. We have used two different supervised learning classification algorithms: Näive Bayes (NB) classifier [81] and a support vector machine (SVM) classifier [76]. For the SVM, we have used the nu-SVC classifier from the libsvm package [16]. A linear kernel with default parameter settings was used since this has shown to give better results than a radial basis function in experiments. The experiments were done using 10-fold cross-validation. The data was randomly partitioned into 10 equal-sized parts. The classification task is performed 10 times, each time a different part is used as test data while the remaining 9 as training data. The results from these ten folds have been averaged into a single accuracy value.

5.4.2 Summary of the results

A number of experiments were conducted to understand to what extent time-based features can be used for author identification. The experiments show that it is more difficult to solve the problem of author identification when the posts are divided randomly compared to when posts are distributed sequentially. This is expected since features such as *Month* will have very similar relative frequencies among sub-users corresponding to the same user when dividing the posts sequentially.

When using sequentially distributed posts for training data and a set of 1000 users the SVM classifier achieved over 90% accuracy. The experiments showed that the SVM classifier outperformed the NB classifier with approximately 5-20% higher accuracy. Nevertheless, the training phase of the NB classifier took a few minutes while it took days for the SVM classifier on a standard computer. Using the built SVM model in actual usage should not be an issue, even it takes time while training.

5.5 Techniques for Alias Matching

The problem of alias matching (or similarity detection) can be defined as, given two user accounts $\mathbf{a_1}$ and $\mathbf{a_2}$, determine whether they belong to the same individual I, i.e., we would like to learn an identification function $f(\mathbf{a_1}, \mathbf{a_2})$ where:

$$f(\mathbf{a_1}, \mathbf{a_2}) = \begin{cases} 1 \text{ if } a_1 \text{ and } a_2 \text{ belong to the same } I, \\ 0 \text{ otherwise.} \end{cases} \tag{5.1}$$

The aliases $\mathbf{a_1}$ and $\mathbf{a_2}$ could be from the same social media site (intra-platform alias matching), or from different social media sites (cross-platform/ inter-platform user linkage).

In the author identification problem, we compare an anonymous user to a fixed set of pre-defined known entities with the assumption that the anonymous user is one of the candidates. In an alias matching setting, we cannot assume that we have knowledge of all potential authors. The problem is instead to compare each anonymous identity to all other identities and group together users (aliases) which are more similar than a certain threshold.

We have investigated the possibility of solving alias matching problem using two different approaches: i) distance-based measures and ii) supervised learning.

5.5.1 Distance-based measures

In the first approach, we used unsupervised distance-based techniques for solving the alias matching problem. We used stylometric and time-based features and compared the similarity between two vectors.

To test the distance-based approach we selected the most active 4000 users from the discussion forum data set. Each selected user has posted at least 60 postings. To simulate a scenario with multiple aliases, we randomly divided posts of a single user u into two separate users u_a and u_b. The process was repeated for all users, and two different sets S_a and S_b are created. The set S_a includes all u_a users and set S_b includes all u_b users. For each user a set of three feature vectors was created: 1) a time-based vector 2) a stylometric vector, and 3) a combination of the time-based feature vector and the stylometric-based feature vector.

In the experiments, we varied the number of users from 500 to 4000 in steps of 500. Each user in the set S_a was compared (using a distance measure) one at a time with all the users in the set S_b. Based on the results from the stylometric matching and time-based matching we rank the users in set S_b according to how similar they are to the selected user in set S_a. The similarity measure used to compare the vectors is cosine similarity. Cosine similarity

between two vectors p and q is given by:

$$\cos(p,q) = \frac{p \cdot q}{\|p\|\|q\|} = \frac{\sum_{i=1}^{n} p_i \times q_i}{\sqrt{\sum_{i=1}^{n} (p_i)^2} \times \sqrt{\sum_{i=1}^{n} (q_i)^2}} \tag{5.2}$$

Time-based and stylometric-based features work very well for a limited number of users. When we have 4000 users, we achieved 70% accuracy using just time-based feature whereas stylometric feature yields around 44%. For the same number of users, the combination of time-based and stylometric-based features was around 75%. The experiments indicate that time-based features are powerful on their own for alias matching and that combining time-based features with stylometric-based features allow for even (statistically significantly) better results.

5.5.2 Supervised learning

We have also experimented with the use of supervised learning algorithms to solve the alias matching problem. In our supervised learning experiment we used stylometric, time-based and emotion-based and Twitter specific features.

We used three different data sets to evaluate our approach: a discussion forum, a set of tweets, and a set of blogs. The tweets were downloaded from Twitter and the blogs from Blogger[2]. Only users that posted at least 60 posts in total were selected. To create a data set that could be used to test our approach, the posts from one user were randomly divided among five sub-users. For each sub-user, we create a feature vector and a user ID, where the user IDs are unique so that all sub-users created from a user share the same user ID, but is different among all sub-users who were not created from the same user. In the next step, we calculate the pairwise absolute difference of all the resulting feature vectors. After taking the pairwise difference, the user ID is now specifying whether sub-users were taken from the same user (0, altered to 1), or not ($\neq 0$, altered to 0 in order to create a binary classification problem). By using this transformation, we alter the problem from an unsupervised learning task of alias matching to a classification problem of classifying matching and unmatching pairwise-difference vectors. We call this approach the *diff-vector* approach.

We tested different classifiers, and the performance of the classifiers showed that AdaBoost performs significantly better than NB and SVM classifiers. Therefore, the AdaBoost classifier [26] was used. The underlying idea of boosting algorithms is to combine simple classification rules (the base classifiers) to form an ensemble, whose performance can be significantly improved. An ensemble is based on the idea that the average of a group of judgments is better than individual judgment. We have used classification trees as base

[2]https://www.blogger.com/

classifiers, wherein the feature space was divided into regions by recursive partitioning. We set the maximum number of iterations to 200 in the experiments. The default parameters were used for the rest of the AdaBoost parameter settings.

The experiments were done on both the intra-platform and cross-platform/ inter-platform scenario. In the case of intra-platform, the classification models were trained, evaluated and tested on the discussion forum and the Twitter data set. While evaluating a cross-platform scenario, the model was trained on the discussion forum data set and tested on the Twitter data set and vice-versa. This is of interest for cases such as where we would like to determine the (unknown) author of a number of tweets given the forum posts written by a set of known set of candidate authors.

Our results showed that both stylometric and time-based features work well when detecting multiple aliases and that the combination of stylometric and time-based features is even more powerful. In the case of the intra-platform experiment when we tested on the discussion forum data set for 2000 test cases (1000 matched pairs and 1000 unmatched pairs), we achieved around 92% accuracy when time-based features were used and 99% accuracy when stylometric features were used. Similarly, for the Twitter data set with 2000 test cases (1000 matched pairs and 1000 unmatched pairs), we obtained 96% accuracy when time-based features were used and 99% accuracy when stylometric features were used. While using all features, we achieved 99% accuracy on both the discussion forum and Twitter.

The cross-platform classification performed poorly when trained on one type of data and tested on others. But a classifier with the mixture of both data sets and with all features performed very well. We achieved 96% accuracy when testing on 2000 test cases (1000 matched pairs and 1000 unmatched pairs). AdaBoost's use of boosting is likely involved in this since it permits different base classifiers in the ensemble to concentrate on learning to classify different subsets of the training data. The good results of the combined classification indicate that a direction for future work is to examine whether training classifiers on data drawn from large numbers of varied data sources permit generalization to new data sources.

Apart from evaluating the techniques in simulated data, we have evaluated the techniques on non-synthetic data consisting of a set of blogs consisting of 1414 distinct bloggers out of which 260 had written at least two separate blogs. When testing our approach on the non-synthetic blog data, we first trained a model on 4000 examples with 663 matched pairs of blogs and randomly selected unmatched pairs. The model was tested on 1000 examples where there are 150 matched pairs and rest are randomly selected unmatched pair. We achieved an accuracy of 93%, a recall on 56% and a precision of 92%.

6. Contribution and Future Perspectives

In this thesis, we have presented different methods and techniques that can be used to analyze text-based social media communication. The different techniques can be used to study environments, phenomena, and individuals. To assure that the suggested techniques can be applied in a real-world scenario, we have implemented and tested our approaches on real data.

Chapter 3 describes three different studies that focus on analyzing different aspects of digital communication. In the first study, we investigate the possibilities to detect propaganda from IS. When recognizing propaganda from IS, we used machine learning models to distinguish between Twitter accounts related to IS and normal users. We used both data dependent and data independent features to get an idea of how well the model would work in the real world. Our results are promising but more research needs to be done to gain a deeper understanding on the performance, in particular, more training data that discusses IS or similar topics as IS but are not pro-IS needs to be included. Our approach of using machine learning to detect IS propaganda on Twitter was one of the first attempts, and the data set of IS supporters that we collected was unique.

In the second study, we used text analysis to identify the use of a narrative containing xenophobic and conspiratorial stereotypes on immigration critic Swedish alternative media. We used a dictionary-based approach to identify references to the stereotypes that the narrative is based upon. We found that references to the stereotype occur in all alternative media in our study but that there is a variance in the frequency of references. The results indicate that all alternative media should not be classified as one type of media. Swedish alternative media have been studied before [35] but using manual approaches. We have not seen any similar studies using text analysis techniques. We acknowledge the fact that dictionary-based method usually performs worse than machine learning methods and for future work, it would be interesting to test different methods, for instance, machine learning and deep neural network models to detect the presence of narratives.

In the third study, we focused on a Swedish discussion forum. We singled out a group of so-called extreme adopters of community-specific jargon, persons that use substantially more jargon-words or expressions than the rest of the users. By studying the set of extreme adopters, we noticed certain linguistic features that we view as markers of a radicalized mind-set - a certain style of understanding and relating to the world that has often been observed among violent extremists. We noticed that the group of extreme adopters differs significantly from the rest of the forum and that it is possible to separate

the two groups using machine learning. One of the main contributions from this study is the definition of the set of linguistic features that indicate a radicalized mind-set. The identified linguistic markers can be used as a part of a risk assessment. For future research, it would be interesting to repeat the same experiment on other forums to get an idea of the generalizability of our results.

In Chapter 4, we studied the possibilities of assessing written communication using different technologies. We used text analysis to learn more about the author's psychological state and world view. The texts we have studied are written by violent lone offenders prior to their engagement in targeted violence. Machine learning experiments show that there seems to be a possibility to assess and use the assessment as a component in a more extensive risk assessment. By using psychological features, it seems to be the case that we can separate communication from violent lone offenders from a normal population of online communication. However, this is a result that needs to be investigated further. The psychological warning behaviors that we are trying to identify in this work relate to targeted violence, but it might be possible to use a similar approach to detect signs of other warning behaviors, such as suicidal behavior. In addition to the scientific publication, we have also created a prototype tool for law enforcement that can be used in risk assessment. The tool is described in a report published by Europol [4]. A natural direction for future work is to continue the development and testing of the tool, in particular, the technologies that are used to assess different aspects of the texts.

Our research contributions are in the field of risk assessment of written communication. In our study of communication from known violent lone offenders, we identified a set of psychological warning behaviors that are present in their communication, and that can be explained by previous psychological research. Another contribution is our approach of using machine learning to separate lone offenders from a normal population and the use of psychological categories (LIWC) as features.

Our research findings are implemented in the risk assessment tool PRAT. The tool can be used to analyze written communication from potential violent lone offenders, and it should be seen as a complement to a manual risk assessment. The tool is currently evaluated by Swedish law enforcement and psychologists.

In Chapter 5, we investigate the possibilities of identifying users with multiple aliases on social media. We propose different matching techniques and features for solving the alias matching problem and the author identification problem. We propose a time-based approach where we use posting times to identify a user. We are not aware of any previous research that uses time profile-based matching and a combination of different techniques for solving the alias matching problem. Our experiments show that the time-based matching technique provides promising results and that the combination of time with stylometric features performs better than the individual approaches alone. A common criticism in machine learning based solutions is that the model is bi-

ased towards the data used for training. We have investigated the possibility to use machine learning techniques to identify users in cross-domain conditions, where the training and test data comes from different platforms. In our work, we have transformed the multi-class author attribution problem to a binary classification problem where we use a machine learning approach to learn the difference between feature vectors. A natural direction for future work is to use deep neural networks or artificial neural networks for author identification rather than using just traditional machine learning algorithms like SVM and Random Forest.

The contribution to the field of research is the development of techniques for solving the alias matching problem. First of all, we have listed different reasons for using multiple aliases and what kind of techniques that can be used to identify users with multiple aliases depending on the reason. We have also introduced *time-based features* to identify users with multiple aliases. To the best of our knowledge, time-based features have not previously been used for solving alias matching problem. The experimental setup that we developed has inspired other researchers such as [70].

Another contribution is the formulation of the alias matching problem as a binary classification problem (as in authorship attribution) and to solve it using our *diff-vector* approach. The method was evaluated on non-synthetic blog data consisting of authors who own multiple blogs. In addition, we have systematically evaluated different classification algorithms on well-controlled data sets and the impact of using various feature sets for alias matching. Apart from this, we have done a cross-platform evaluation where a model is trained and tested on data sets from different platforms. The cross-platform approach was first used in our work.

The techniques suggested in this thesis have the potential to be used to solve problems in the area of *Intelligence and Security Informatics*. Given the challenges posed by the enormous amount of information on the Internet, a more automated approach to analyze digital communication has become a necessity. Automatic processing of massive digital data and utilizing machine learning technologies makes it feasible for a human observer to analyze digital communications quantitatively. We believe the work that we have done in this thesis is a step towards such a development.

When using machine learning, there is always the risk that the models that are built are applicable only on specific data set. We have included data independent features in our studies to increase the understanding of the performance of models in the wild. Our experiments indicate that machine learning and automated techniques can be used to solve many of the problems we have studied. However, it should be noted that the results have been obtained in a well-controlled experimental setting which does not necessarily hold in a real-world environment. While computerized text analysis and machine learning technologies will most likely serve as a component in the analysis process, it is important to stress that it might be hard for automated techniques to replace

human analysts. Such techniques will instead serve as screening tools as well as a provider of a complementary view. Even though the techniques suggested in this thesis are developed for the security of society, one cannot ignore the fact that there is always an ethical risk involved in developing such techniques.

References

[1] A. Abbasi. Affect intensity analysis of dark web forums. In *Intelligence and Security Informatics, 2007 IEEE*, pages 282–288, May 2007.

[2] A. Abbasi and H. Chen. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, Sept 2005.

[3] A. Abbasi and H. Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Trans. Inf. Syst.*, 26(2):7:1–7:29, 2008.

[4] N. Akrami, A. Shrestha, M. Berggren, L. Kaati, M. Obaidi, and K. Cohen. *Assessment of risk in written communication: Introducing the Profile Risk Assessment Tool (PRAT)*. The European Commission, Europol, 2018.

[5] M. Almishari, E. Oguz, and G. Tsudik. Fighting authorship linkability with crowdsourcing. In *Proceedings of the Second ACM Conference on Online Social Networks*, COSN '14, pages 69–82, New York, NY, USA, 2014. ACM.

[6] S. Argamon, M. Šarić, and S. S. Stein. Style mining of electronic messages for multiple authorship discrimination: First results. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 475–480, New York, NY, USA, 2003. ACM.

[7] M. Ashcroft, F. Johansson, L. Kaati, and A. Shrestha. Multi-domain alias matching using machine learning. In *Third European Network Intelligence Conference (ENIC)*, 2016.

[8] M. F. Atig, S. Cassel, L. Kaati, and A. Shrestha. Activity profiles in online social media. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2014.

[9] E. Baaker and B. de Graaf. Preventing lone wolf terrorism: Some CT approaches addressed. In *Perspectives on Terroris*, December 2011.

[10] H. Beirich. Intelligence report, 2015.

[11] M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro, and C. S. Haley. Application of high-dimensional feature selection: Evaluation for genomic prediction in man. *Scientific Reports*, 5:10312 EP –, 2015.

[12] S. El M. El Bouanani and I. Kassou. Authorship analysis studies: A survey. *International Journal of Computer Applications*, 86(12), 01 2014.

[13] M. Brennan, S. Afroz, and R. Greenstadt. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security*, 15(3), 2012.

[14] B. Brunt. Violence risk assessment of the written word (vraw2). *The Journal of Campus Behavioral Intervention*, 3:12–25, 11 2015.

[15] J. Brynielsson, A. Horndahl, F. Johansson, L. Kaati, C. Mårtenson, and P. Svenson. Harvesting and analysis of weak signals for detecting lone wolf terrorists. *Security Informatics*, 2(1):11, 07 2013.

[16] C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
[17] C. E. Chaski. Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8(1):1–65, 2001.
[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and P. W. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, June 2002.
[19] H. Chen. *Intelligence and Security Informatics for International Security*, volume 10. Springer US, 1st edition, 2006.
[20] H. Chen. Sentiment and affect analysis of dark web forums: Measuring radicalization on the internet. In *ISI*, pages 104–109. IEEE, 2008.
[21] B. M. Christopher. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.
[22] K. Cohen, T. Isbister, L. Kaati, and A. Shrestha. Linguistic markers of a radicalized mind-set among extreme adopters. In *1st International Workshop on Cyber Deviance Detection (CyberDD)*, 2017.
[23] K. Cohen, F. Johansson, L. Kaati, and J. C. Mork. Detecting linguistic markers for radical violence in social media. *Terrorism and Political Violence*, 2013.
[24] K. Cohen, L. Kaati, S. Lindquist, and A. Shrestha. Automatic detection of xenophobic narratives: A case study on Swedish alternative media. In *IEEE Conference on Intelligence and Security Informatics (ISI)*, 2016.
[25] K. Cohen, L. Kaati, and A. Shrestha. Linguistic analysis of lone offender manifestos. In *IEEE International Conference on Cybercrime and Computer Forensic (ICCCF)*, 2016.
[26] M. Culp, K. Johnson, and G. Michailides. ada: An r package for stochastic boosting. *Journal of Statistical Software, Articles*, 17(2):1–27, 2006.
[27] J. E. Custódio and I. Paraboni. Each-usp ensemble cross-domain authorship attribution: Notebook for pan at clef 2018. In *CLEF*, 2018.
[28] A. Dalgaard-Nielsen. Violent radicalization in Europe: What we know and what we do not know. *Studies in Conflict & Terrorism*, 33(9):797–814, 2010.
[29] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 307–318, New York, NY, USA, 2013. ACM.
[30] O. de Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *SIGMOD Rec.*, 30(4):55–64, December 2001.
[31] A. Fisher and N. Prucha. The call-up: The roots of a resilient and persistent jihadist presence on twitter. *CTC Sentinel*, 04, 2014.
[32] P. Gill, E. Corner, M. Conway, A. Thornton, M. Bloom, and J. Horgan. Terrorist use of the internet by the numbers: Quantifying behaviors, patterns, and processes. *Criminology & Public Policy*, 16(1):99–117, 2017.
[33] P. Gill, J. Horgan, and P. Deckert. Bombing alone: Tracing the motivations and antecedent behaviors of lone-actor terrorists. *Journal of Forensic Sciences*, 59(2):425–435, 2014.
[34] O. Halvani and L. Graner. Cross-domain authorship attribution based on compression: Notebook for pan at clef 2018. In *CLEF*, 2018.

[35] K. Holt. "Alternativemedier"? En intervjustudie om mediekritik och mediemisstro. In *Migrationen i medierna – men det får en väl inte prata om? Institutet för mediestudier*, 2016.
[36] J. A. Horne and O. Östberg. A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *Int J Chronobiol*, 4(2):97–110, 1976.
[37] Molly Ireland. Candidates' language in speeches and interviews: Summary comparisons, 2008.
[38] G. James, D. Witten, T. Hastie, and R Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.
[39] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, October 2002.
[40] F. Johansson, L. Kaati, and A. Shrestha. Detecting multiple aliases in social media. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2013.
[41] F. Johansson, L. Kaati, and A. Shrestha. Time profiles for identifying users in online environments. In *IEEE Joint Intelligence and Security Informatics Conference (JISIC)*, 2014.
[42] F. Johansson, L. Kaati, and A. Shrestha. Timeprints for identifying social media users with multiple aliases. *Security Informatics*, 4(1):7, 2015.
[43] P. Juola. Authorship attribution. *Found. Trends Inf. Retr.*, 1(3):233–334, 2006.
[44] L. Kaati, E. Lundeqvist, A. Shrestha, and M. Svensson. Author profiling in the wild. In *2017 European Intelligence and Security Informatics Conference (EISIC)*, 2017.
[45] L. Kaati, E. Omer, N. Prucha, and A. Shrestha. Detecting multipliers of jihadism on Twitter. In *IEEE 15th International Conference on Data Mining Workshops (ICDMW)*, 2016.
[46] L. Kaati, T. Sardella, and A. Shrestha. Identifying warning behaviors of violent lone offenders in written communication. In *IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2016.
[47] L. Kaati and P. Svenson. Analysis of competing hypothesis for investigating lone wolf terrorist. In *European Intelligence and Security Informatics Conference (EISIC)*, 2011.
[48] G. Kacmarcik and M. Gamon. Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the 2006 COLING/ACL*, 2006.
[49] L. Kelly and K. Karsna. Measuring the scale and changing nature of child sexual abuse and child sexual exploitation. Scoping report. Technical report, London Metropolitan University, July 2017.
[50] R. J. Larsen. Individual differences in circadian activity rhythm and personality. *Personality and Individual Differences*, 6(3):305 – 311, 1985.
[51] W. Magdy, K. Darwish, and I. Weber. #failedrevolutions: Using twitter to study the antecedents of ISIS support. *CoRR*, abs/1503.02401, 2015.
[52] R. Mankiewicz. *The Story of Mathematics*. Princeton University Press, Princeton, NJ, 2004.
[53] C. McCauley and S. Moskalenko. The psychology of lone-wolf terrorism. *Counseling Psychology Quarterly*, 24(1):115–126, 2011.

[54] A. W. E. McDonald, S. Afroz, A. Caliskan, A. Stolerman, and R. Greenstadt. Use fewer instances of the letter "i": Toward writing style anonymization. In *Privacy Enhancing Technologies*, pages 299–318, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[55] J. R. Meloy and M. E. O'Toole. The concept of leakage in threat assessment. *Behavioral Sciences and the Law*, 4(29):513–527, 2011.

[56] A. M. Mohsen, N. M. El-Makky, and N. Ghanem. Author identification using deep learning. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 898–903, Dec 2016.

[57] F. Mosteller and D. L. Wallace. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309, 1963.

[58] A. Narayanan, H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. C. R. Shin, and D. Song. On the feasibility of internet-scale author identification. In *2012 IEEE Symposium on Security and Privacy*, pages 300–314, May 2012.

[59] J. Novak, P. Raghavan, and A. Tomkins. Anti-aliasing on the web. In *Proceedings of the 13th international conference on World Wide Web*, pages 30–39, New York, NY, USA, 2004. ACM.

[60] J. W. Pennebaker. *The secret life of pronouns: What our words say about us.* CT New York: Bloomsbury Press., 2011.

[61] J. W. Pennebaker and C. K. Chung. Computerized text analysis of al-Qaeda transcripts. In K. Krippendorf and M. A. Bock, editors, *The Content Analysis Reader*. Sage, 2008.

[62] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count (LIWC): A text analysis program. 2001.

[63] J. W. Pennebaker, M. Mehl, and K. Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[64] J.W. Pennebaker, R.L. Boyd, K. Jordan, and K. Blackburn. The development and psychometric properties of liwc2015. In *Austin, TX: University of Texas at Austin*, 2015.

[65] M. Potthast, B. Stein, P. Rosso, and E. Stamatatos. PAN. https://pan.webis.de. [Online; accessed 10-March-2019].

[66] J. Reid Meloy, Jens Hoffmann, Angela Guldimann, and David James. The role of warning behaviors in threat assessment: An exploration and suggested typology. *Behavioral Sciences & the Law*, 30(3):256–279, 2012.

[67] C. Sammut and G. I. Webb. *Encyclopedia of Machine Learning*. Springer Publishing Company, Incorporated, 1st edition, 2011.

[68] A. Shrestha, L. Kaati, and K. Cohen. A machine learning approach towards detecting extreme adopters in digital communities. In *1st International Workshop on Advanced ICT Technologies for Secure Societies (AICTSS)*, 2017.

[69] R. W. Shuy. Linguistic profiling when there is no known murder suspect. In *The Language of Murder Cases: Intentionality, Predisposition, and Voluntariness*, chapter 4. Oxford Scholarship, Oxford, 2014.

[70] M. Spitters, F. Klaver, G. Koot, and M. Staalduinen. Authorship analysis on dark marketplace forums. In *European Intelligence and Security Informatics Conference (EISIC)*, pages 1–8, Los Alamitos, CA, USA, sep 2015. IEEE Computer Society.

[71] M. Spitters, S. Verbruggen, and M. V. Staalduinen. Towards a comprehensive insight into the thematic organization of the tor hidden services. In *2014 IEEE Joint Intelligence and Security Informatics Conference*, pages 220–223, Sep. 2014.

[72] E. Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

[73] E. Stamatatos. Authorship attribution using text distortion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1138–1149. Association for Computational Linguistics, 2017.

[74] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[75] H. van Halteren. Linguistic profiling for author recognition and verification. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[76] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY, USA, 1995.

[77] J. Victoroff. The mind of the terrorist a review and critique of psychological approaches. *Journal of Conflict resolution*, 49(1):3–42, 2005.

[78] P. Wadhwa and M. P. S. Bhatta. Classification of radical messages on Twitter using security associations. In Nauman Israr Biju Issac, editor, *Case Studies in Secure Computing*, chapter 14, page 22. Auerbach Publications, New York, 2014.

[79] N. Woolf. Twitter suspends 235,000 accounts in six months for promoting terrorism. https://www.theguardian.com/technology/2016/aug/18/twitter-suspends-accounts-terrorism-links-isis. [Online; accessed 3-January-2019].

[80] G. U. Yule. In *The Statistical Study of Literary Vocabulary*. Cambridge University Press, 1944.

[81] H. Zhang. The optimality of naïve bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*. AAAI Press, 2004.

[82] R. Zheng, J. Li, H. Chen, and Z. Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *J. Am. Soc. Inf. Sci. Technol.*, 57(3):378–393, 2006.

[83] R. Zheng, Y. Qin, Z. Huang, and H. Chen. Authorship analysis in cybercrime investigation. In *Intelligence and Security Informatics*, pages 59–73, Berlin, Heidelberg, 2003. Springer.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1786

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-379605

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2019