Ole Jonas Liahagen
Martin Johannes Nilsen

Literature study on detecting traits of potential lone offenders using NLP techniques on social media data

TDT4501 Computer Science, Specialization Project
Autumn 2022
Supervised by Björn Gambäck

Data and Artificial Intelligence Group
Department of Computer Science
Faculty of Information Technology and Electrical Engineering
Norwegian University of Science and Technology

NTNU | Norwegian University of Science and Technology

Abstract

Social media continue to play a huge role in society and people's lives. With the reported amount of daily users of social media being in the billions, the platforms are host to all kinds of people. Allowing for free speech and in some cases anonymity, users are allowed to express themselves in any way they want. Investigations after lone wolf attack incidents have shown that a vast majority have used social media before and leading up to the incident, sometimes even posting material that could serve as warning signs of an upcoming attack. Previous work on sentiment analysis has shown that a significant amount of information about a person can be retrieved from their social media presence. As school shooting incidents and lone wolf attacks continue to increase in frequency, it is evident that today's methods are not enough. Utilizing advances in machine learning to profile lone wolf attackers via social media could be another tool to assist in intervening before a potential attack is carried out.

This preliminary project explores the possibilities of performing personality profiling on social media users using their social media presence as data. This profiling is then put into the context of profiling potential lone wolf attackers and school shooters to evaluate the feasibility of detecting said individuals. The project is structured as a literature review aimed at serving as a solid foundation for an upcoming Master's Thesis. The literature review is conducted to identify the state-of-the-art in personality prediction and how this is applied to the screening of social media users. After the identified studies are discussed, the report ends with a suggestion for future work for the said Masters's Thesis.

Sammendrag

De siste årene har utviklingen og tilgjengeligheten av internett og sosiale medier ført til en endring av folks liv. En stadig større del av verdens befolkning tar i bruk disse plattformene, og ettersom antallet månedlige brukere rapporteres i milliarder, vil disse tjenestene inneholde et vidt spekter av mennesker. For a tillate ytringsfrihet og relativ anonymitet, har brukere i noen tilfeller lov til a uttrykke seg uten sensurering. Forskning på individuelt utførte voldelige angrep har vist at en stor andel av nevnte utøvere har brukt sosiale medier eller forum i forkant av handlingen deres, og noe materiale kan kategoriseres som en advarsel for et kommende angrep. Samtidig viser relevant arbeid innenfor sentimentanalyse at en betydelig mengde informasjon om en person kan tolkes utifra deres tilstedeværelse på disse mediene. Ettersom skoleskytinger og andre individuelt utførte angrep fortsetter å øke i frekvens, er det tydelig at dagens observasjonsmetoder ikke er nok. Maskinlæringsmodeller kan bli trent på data fra sosiale medier med den hensikt å profilere og detektere personlighetstrekk som kan knyttes mot utøvere av slike handlinger. Dette har potensiale for å være et verktøy, som i kombinasjon med de på mange måter manuelle og menneskelige vurderingene som gjøres i dag, kan bli tatt i bruk for å gripe inn før et potensielt angrep utføres.

Dette prosjektet har som formål å utforske mulighetene for å utføre personlighetsprofilering på brukere av sosiale medier, ved å ta i bruk deres publiserte innlegg som data. Denne profileringen kan så bli satt opp mot trekk som i litteraturen knyttes opp mot de tilknyttet blant annet skoleskytere, for å utforske muligheten til å oppdage nevnte individer. Rapporten er strukturert som en litteraturgjennomgang med den hensikt å fungere som en base for en kommende masteroppgave. Arbeidet er utført for å identifisere relevant arbeid innen fagfelt som personlighetsmodellering, automatisk deteksjon med maskinlæring, bruken av data fra sosiale medier, samt allerede eksisterende arbeid innen identifisering av voldelige atferdstrekk. Relevant bakgrunnsteori innen nevnte områder vil også bli presentert. I tillegg, avsluttes rapporten med et forslag til fremtidig arbeid for den nevnte masteroppgaven.

Preface

The purpose of this report is to serve as a preparation for the upcoming Master's Thesis at the Norwegian University of Science and Technology. The project was carried out by two students in the Computer Science Department's Data and Artificial Intelligence Group at NTNU Trondheim, as part of the subject TDT4501 Computer Science, Specialization Project. For providing relevant work, as well as for his guidance and support during the drafting of this report, we would like to thank our supervisor Björn Gambäck.

Ole Jonas Liahagen & Martin Johannes Nilsen
Trondheim, 14th December 2022

Contents

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence 13

BOW Bag of Words 11, 16, 30

CBOW Continous Bag of Words 12

CNN Convolutional Neural Network 31

DF Document Frequency 11

FN False Negative 19, 20

FNN Feedforward Neural Network vii, 17, 18

FP False Positive 19

GNB Gaussian Naive Bayes 16

GRU Gated Recurrent Unit 31

IDF Inverse Document Frequency 11

KNN K-Nearest Neighbors 31

LDA Linear Discriminant Analysis 31

LIWC Linguistic Inquiry and Word Count 11, 29–31, 33, 35

LSTM Long Short-Term Memory 12, 31

MAE Mean Absolute Error 20

ML Machine Learning 5

MLP Multilayer Perceptron 31

MNB Multinomial Naive Bayes 16

MSE Mean Square Error 13, 20

NLP Natural Language Processing 5, 10, 16, 27, 34

RMSE Root Mean Square Error 20

SLR Structured Literature Review 3, 27, 36

SNA Social Network Analysis 28, 33

SPJ Structured Professional Judgment 8

SVM Support Vector Machine 14, 31

TF Term Frequency 11

TF-IDF Term Frequency-Inverse Document Frequency 11, 30

TN True Negative 19

TP True Positive 19, 20

Glossary

artificial intelligence A field of study focused on computer systems' ability to simulate human intelligence processes 10

feedforward neural network An artificial neural network without cycles in the connections between the nodes 17

information retrieval The process of searching for and retrieving recorded data and information from a file or database 11, 19

machine learning A subfield of artificial intelligence, devoted to understanding and building methods that leverage data to improve performance, thus 'learning' based on given data 11, 19

natural language processing A subfield of artificial intelligence that aims to enable computers to understand spoken and written communication in a manner similar to that of humans 10, 11

one-hot encoding A binary vector representation of a sentence or categorical variable 11

pearson correlation coefficient A measure of linear correlation between two sets of data, in the form of a normalized value between -1 and 1 30

reinforcement learning An area of machine learning concerned with how intelligent agents ought to take actions in an environment, in order to maximize a cumulative reward 13

supervised learning A subcategory of machine learning for problems where the available data consists of labeled examples 13

unsupervised learning A subcategory of machine learning for problems where the available data consists of unlabelled examples 13

1. Introduction

Prior work in the field of automatic personality detection has suggested that it might be possible to detect common traits in radicalized individuals using their own social media presence as data (Brynielsson et al., 2013; Barhamgi et al., 2018). However, little work has been done on applying this methodology to the group of perpetrators acting alone such as school shooters. This leaves an opportunity for a study into the possibility of leveraging this technique in a fresh environment. The hope is that the use of said technique may aid in the fight against lone-wolf attacks in the form of e.g. pre-emptive warning.

With this in mind, this chapter introduces the background and motivation for the project as well as the overall research goal. Further, a suggested structure for the literature review is presented before a list of contributions and an overview of the chapters to come.

1.1 Background and Motivation

For half a century, the proportion of attacks committed by a lone wolf attacker, that is, an individual with no connection to an organized group, has increased dramatically. According to the Global Terrorism Index, the proportion of unaffiliated lone wolf terrorists has seen a rise from 5% in 1970, to above 70% for the period between 2014 and 2018 (Vision of Humanity, 2019). This has led to a series of analyses on the online behavior of lone actor terrorists. One of these, a study commissioned by the U.S. Department of Justice, presents a sequence of common features associated with the pathways of lone wolf terrorists. These features include personal and/or political grievances forming the basis for an affinity with online sympathizers, followed by the identification of an enabler, broadcast of intent, and a triggering event (Hamm & Spaaij, 2015). As stated in the report, the ability to detect and prevent lone wolf attacks demands a clear understanding of the process and forms a potential for detection if the signs are clear. Further specifying the case of school shootings, the number of yearly incidents has increased drastically for the past ten years. As reported by the research center of The Violence Project, which is responsible for publishing the K-12 School Shooting Database, the numbers have gone from 20 school shooting incidents in 2012 to 296 occurrences in 2022 at the time of writing, in the U.S. alone (Riedman, 2022).

The rise of the internet, and social networking platforms, has enabled lone wolf actors to take part in virtual communities of like-minded people, radicalizing and educating one another on planning and executing attacks (Ganor, 2021). With social media platforms' immense growth, there is reason to believe that also lone wolf actors use these platforms, possibly years before an attack takes place.

With the number of school shootings steadily on the rise, there has been considerable effort invested in studying the psychological aspects of a school shooter. These studies are mainly concerned with pedagogic approaches to prevent someone from ever becoming a school shooter. However, there are times when a person has already gone over the edge and intervention is needed. Previous studies, although over ten years old, revealed that every prominent school shooter between 2005 and 2010 had a presence on social media, with some leaving clues hinting to a potential future attack (Semenov et al., 2010). Traditionally the screening of social media posts is done manually or through tips from students, and little work has been done to attempt to leverage machine learning to combat lone wolf attacks (Neuman et al., 2020). As social media continue to grow, the data needed to be screened is increasing at a fast pace, thus making automatic methods of detection more relevant than ever before.

1.2 Research Goal and Research Questions

This project will serve as a preliminary study of the interdisciplinary field of pre-emptively detecting potential lone wolves such as school shooters. The study aims to serve as an overview of previous work done on the topic as well as describe future work suggested by preceding authors. The goal is to create a solid starting point for an upcoming Master's Thesis that will base itself on the work done in this report. To help concretize the direction of this study and help structure research, the motivation and goal of the project have been synthesized into an overall research goal below.

Research goal *An exhaustive overview of the field of using social media to identify users with traits similar to known lone wolf perpetrators for use in a future Master's Thesis*

Four research questions have been constructed to distill the research goal into more concrete areas of interest that are to be explored. The purpose of these research questions is to help get an overview of the state-of-the-art concerning the automatic detection of lone wolf perpetrators using social media.

Research question 1 *What work has already been done to identify personality traits based on social media presence?*

The first research question aims to explore the approach of using personality traits to screen for potential lone wolf perpetrators. This also includes a study into the different methods of extracting said personality traits from texts.

Research question 2 *What sources of data documenting previous incidents of lone wolf attacks are already available?*

The second research question tackles the issue of the availability of datasets. It explores the accessibility of databases linked to e.g. school shootings and aims to determine if there are previous data collected that could be useful for future experimentation.

Research question 3 *What are proposed as distinguishing traits of a lone wolf perpetrator?*

Drawing inspiration from research done on detecting potential terrorists using personality traits, the third research question aims to investigate whether lone wolf perpetrators such as school shooters have distinguishing personality traits in common, and if so, what they are.

Research question 4 *What can be done as further work to screen for lone wolf perpetrators using social media?*

The fourth and final research question concerns the culmination of the research already performed in the problem domain. This question should serve as a means to get an overview of open problems concerning the automatic detection of lone wolf perpetrators as well as give a solid starting point for an eventual Master's Thesis.

1.3 Research Method

The research for this preliminary study was focused on finding literature pertaining to the act of using machine learning to detect potential perpetrators. This area of research encompasses the disciplines of both psychology and computer science, thus a robust and thorough method of the literature review was needed. In this study, the exploration of research is done by utilizing two main methods; snowballing and a structured literature review. The structured literature review (SLR) was conducted as described by Kofod-Petersen (2018). The queries used in the SLR were based on the research goal and questions presented in section 1.2. After the initial SLR, a snowballing approach was adopted. Snowballing was performed as described by Wohlin (2014). The documents retrieved from SLR were used as the starting set for the snowballing approach. Finally, documents retrieved outside of SLR and snowballing, including work suggested by our supervisor Björn Gambäck, can be found in the list of references.

1.4 Contributions

Through the preliminary study conducted in this paper, different topics will be presented and discussed regarding lone actor trait detection. Hence, as this is a literature review, the contributions involve discovery and utilization of the literature that already exists on the matter. No actual implementation is provided, as this will be a task for the following Master's Thesis. Below is the full list of contributions.

1. *A presentation of challenges and subjects of debate associated with automatic detection of traits similar to those of a known lone wolf perpetrator using social media data*
2. *A presentation of publically available data regarding the incidents of lone wolf attacks and/or school shootings*
3. *A presentation of potential future work for said automatic identification of lone wolf perpetrators*
4. *A presentation of relevant background theory and state-of-the-art in modeling personality, natural language processing, and machine learning in general*

1.5 Document Structure

The following sections of this paper serve as a preliminary (mainly literature) study for our Master's Thesis. For the purpose of gaining an overview of the rest of the paper to come, each of the sections could be shortly described as follows:

Chapter 2 presents the relevant background theory, that is, the methods, models, and metrics presented in the project. Additionally, the field of machine learning, natural language processing, and relevant psychological elements will be introduced.

Chapter 3 presents the previously performed related work done by others in the field.

Chapter 4 will discuss and evaluate the findings in Chapter 3.

Chapter 5 includes a conclusion of key findings and a proposal of potential future work.

2. Background Theory

This chapter gives an overview of the relevant theory that forms the basis of the work to be done in this literature review, as well as the following Master's Thesis. Starting with the introduction of a selection of the largest relevant social media platforms as of today, follows an overview of both general and widely used personality models, in addition to models being more closely connected to the purpose of identifying a lone wolf perpetrator. Furthermore, this chapter introduces the most relevant aspects of Natural Language Processing, accompanied by Machine Learning models used for classification through text.

2.1 Social Media Platforms

The widespread adoption of social media platforms has brought billions of people together, enabling them to connect and share information. These platforms have become a fundamental part of modern society, as well as a rich source of data and information about their users. In this section, the most relevant platforms for the sake of lone wolf detection will be described.

2.1.1 Twitter

The microblogging platform Twitter joins the list of the largest social networks, with its reported number of monthly active users in the first quarter of 2022 hitting 436 million (Statista, 2022). Content-wise, the user is presented with a feed of posts, or "tweets", related to followed accounts, along with a thread dedicated to discussing or commenting on the content of the post. Each post is strictly limited to 280 characters, which increased from 140 characters in 2017 and can include a variety of content such as plain text, media, hyperlinks, user mentions, and hashtags. With a real-time nature, the users can see each other's posts in a near-instantaneous fashion. The content on the platform can be viewed by anyone, even without creating an account. However, only registered users can post, comment, like, or reshare posts. With the ability to interact directly with public figures, businesses, and news organizations, there is no wonder why people are drawn to the platform.

2.1.2 Facebook

Facebook was launched in 2004 and has since then become the most popular social media platform by users (Statista, 2022), with a reported number of close to 2 billion daily active users in September 2022 (Meta, 2022). Each user creates their own profile, including information such as occupation, education, and demographic information, in addition to

hobbies, interests, and possibly a biography. By adding other users as their ”friends”, the user is met with a timeline of all their friends' posts, with the ability to react, comment, and share it to either their own list of friends or make it publically available at their own user page. The posts can include a wide range of content, ranging from plain or formatted text, images, videos, hashtags, and user mentions. In addition to interacting with friends through the feed, the social networking service has the option of creating groups, events, and private messaging threads, which further serves the purpose of facilitating communication between both friends and strangers. As the platform has become more and more popular over time, also news outlets and organizations have taken to the platform, letting each user follow them, resulting in their published content also populating each follower's feed. Finally, the use of targeted advertisement has been implemented, utilizing massive amounts of collected data with the purpose of more efficient and personalized marketing.

2.1.3 Stormfront

Stormfront is a white supremacist and neo-nazi internet forum, popularly known to be the first major racial hate site on the web. Founded in 1995 by former Ku Klux Klan leader Don Black, it has faced significant backlash and criticism for its extremist views and promotion of hate speech. In 2017, the website was shut down by its domain registrar because of insulting remarks being published on the forum in the aftermath of a violent rally in Charlottesville, Virginia (Hatewatch, 2017). However, the website has continued to operate on different domain names and hosts. This is not the only action that has been taken against the website, as the forum has been removed from the search engine results of Google in several countries (Zittrain & Edelman, 2002). Although attracting criticism, the website claimed in 2015 to have more than 300000 registered members (Walter et al., 2022), and advocates for white separatism and the creation of a white homeland, with the intention of being the voice of the ”new, embattled White minority” (Stormfront, 2022). Stormfront has been associated with a number of violent crimes, including the mass shooting at a Sikh temple in Wisconsin in 2012 (Elias, 2012) and the murder of British member of Parliament Jo Cox in 2016 (Potok, 2016).

2.2 Modeling Personality

For being able to identify specific personality traits of the ones connected to known lone wolf attackers, there is a need for a framework or model which picks up on this information. A substantial portion of psychology research is targeted against personality modeling, aiming to capture the traits and underlying factors that contribute to a person's behavior. Both general models such as the Myers-Briggs Type Indicator and Big 5 will be described in this section, in addition to more specific models for the special case of identifying personality related to violent behavior.

2.2.1 Myers-Briggs Type Indicator

Based on the theories of psychologist Carl Jung, the Myers-Briggs Type Indicator (Myers et al., 1985) is a self-report questionnaire for capturing the personality of the respondent. There are 4 distinct personality traits the type indicator focus on, where each person is said to have one preferred quality of each (binary) category, thus producing 16 unique types. The four categories are (1) introversion or extraversion, (2) sensing or intuition, (3) thinking or feeling, and (4) judging or perceiving. Furthermore, the letter from each category is used to produce a four-letter test result, or personality type, such as "ENTP" (extraverted, intuitive, thinking, perceiving) or "ENFJ" (extraverted, intuitive, feeling, judging). The full list of possible combinations is presented in Table 2.2.1.

ISTJ	ISFJ	INFJ	INTJ
ISTP	ISFP	INFP	INTP
ESTJ	ESFJ	ENFJ	ENTJ
ESTP	ESFP	ENFP	ENTP

Table 2.2.1: The 16 types of the Myers-Briggs Type Indicator

2.2.2 The Big 5

The Big 5 is a widely applied model for the task of grouping personality traits. Developed through years of research in the field of trait theory, the model, also known as the Five-Factor Model, is according to Power and Pluess (2015) one of the most recognized approaches in personality modeling. The model consists of five traits, as the name suggests, which are Openness, Extraversion, Conscientiousness, Neuroticism, and Agreeableness. Through standardized questionnaires, Big Five Inventory (John & Srivastava, 1999), NEO PI-R (Costa & McCrae, 2008), and The International Personality Item Pool (Goldberg, 1999) to name a few, the respondent is assigned scores on a continuous scale for each of the traits. This way, individuals can self-report and learn more about themselves, and potentially explain how they think, feel, and tend to behave in certain situations.

Further elaborating on each of the 5 traits, Openness refers to how open-minded an individual is to new experiences, perspectives and ideas. On one side of the scale, you may find those who are inventive and curious, thus scoring high on Openness, whereas those who score low tend to be more cautious to change and prefer things to stay consistent. The next trait, Conscientiousness, aims to describe an individual's level of self-discipline, responsibility, and organization. On one side, the high scorers on Conscientiousness tend to be more optimistic, well-organized, goal-driven, and detail-oriented. While on the other

side, scoring low in Conscientiousness means you may be more impulsive, have trouble focusing on goals, be less structured, procrastinate more, and have more difficulty staying organized. Thirdly, Extraversion describes how socially confident, assertive, and energetic someone is. People who are high in Extraversion tend to be outgoing, sociable, and talkative, while those who are low in Extraversion are more reserved, independent, and introverted. The fourth trait is Agreeableness, which seeks to be a measure of empathy, compassion, and cooperativeness. Highly agreeable people tend to be kind to others, helpful, caring, trustworthy, and compassionate, while people with lower scores in this trait might be more selfish, stubborn, and manipulative, and less likely to help others. The fifth and final trait, Neuroticism or emotional stability, is characterized by upsetting thoughts and feelings of moodiness or sadness. People scoring low on this trait are usually more optimistic, tend to easily manage stress, and do not worry a lot while scoring high in Neuroticism suggests that you often feel insecure, get stressed easily, and may appear irritable, worried, or moody to others.

2.2.3 Dark Triad

As illustrated by Figure 2.2.1, and as the name suggests, The Dark Triad (Paulhus & Williams, 2002) consists of three personality traits related to malicious human behavior, namely Narcissism, Machiavellianism, and Psychopathy. The first of the traits, Narcissism, is characterized by said lack of empathy, hypersensitivity to criticism, and often selfish, boastful and arrogant behavior. Attributes associated with the next trait, Machiavellianism, include self-interest, manipulation, and a lack of morality and emotion. Psychopathy, the last of the traits in the triad, is often connected with impulsive, volatile, remorseless and antisocial behavior. In summary, people exhibiting these traits are often described as lacking empathy, being manipulative, and having no conscience.

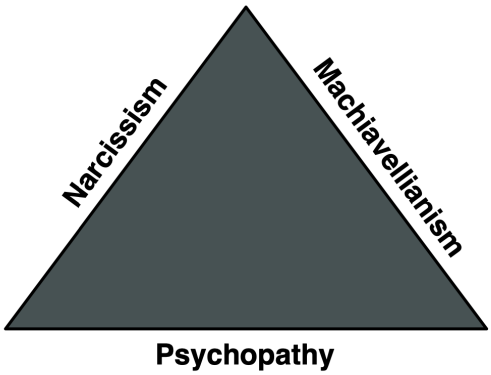

Figure 2.2.1: Illustration of the dark triad

2.2.4 TRAP-18

As presented by Meloy and Gill (2016), TRAP-18 is a Structured Professional Judgment (SPJ) tool for evaluating individuals who may potentially engage in lone actor terrorism. The Terrorist Radicalization Assessment Protocol (TRAP-18) is two-fold in nature, consisting of eight nearby warning behaviors, and ten distant characteristics which can be

related to lone actor terrorists. Table 2.2.2 lists both the proximal and distal features of the tool, whereas the proximal features are described as dynamic and based on patterns of behavior, meanwhile certain of the distal characteristics, e.g. a history of mental disorder, are static risk factors. It should be mentioned that the creators of the framework do emphasize the fact that the framework is neither a psychological test nor an actual risk assessment instrument, but have shown promising results in their presented work (Guldimann & Meloy, 2020).

Proximal attributes	**Distal attributes**
Pathway	Personal grievance and moral outrage
Fixation	Framed by an ideology
Identification	Failure to affiliate with extremist or another group
Novel aggression	Dependence on the virtual community
Energy burst	Thwarting of occupational goals
Leakage	Changing in thinking and emotion
Directly communicated threat	Failure of sexual intimate pair bonding
Last resort behavior	Mental disorder
	Creativity and innovation
	History of criminal violence

Table 2.2.2: The proximal and distal attributes of the TRAP-18 tool

2.3 Natural Language Processing

Natural language processing is a field combining linguistics, computer science and artificial intelligence, where the goal is to be able to process and interpret natural language. With an often ambiguous, unstructured, creative, and redundant nature, it differs from logical and structured languages such as those used in mathematics and programming. For being able to use the vast amounts of data that are collected every day, every hour and every minute, this ambiguous language needs to be translated into something that a computer can understand and utilize. This is where text processing, text representation and feature selection come in. In this section, some of the most relevant techniques related to the processing of text within a document, and the representation of a collection of documents, called a corpus, will be described before elaborating on the task of feature selection.

2.3.1 Preprocessing of Textual Information

Preprocessing of textual data is the act of structuring and sanitizing textual data prior to use in downstream tasks. Depending on the task at hand, the steps for preprocessing vary. Common preprocessing steps for most NLP tasks are:

- *Segmentation* - extracting sentences from the document
- *Tokenization* - dividing the sentences into meaningful semantic units, often words or sub-words, called tokens
- *Normalization* - the act of converting terms into a common canonical representation. In its simplest form, the process can include replacing non-alphabetical units such as numbers with the textual alternative or transforming the casing to e.g. lowercase or title case
- *Stemming* - another form of text normalization, involving removal of word affixes, leaving only the word-stem. This technique should be used with caution, as some words such as ”work” have both a verb and noun form, and will be truncated into the same stem.
- *Lemmatization* - a more complex type of normalization, trying to group together the inflected forms of a word. Using the word’s lemma or dictionary form, lemmatization enables each word, despite its inflection, to be analyzed as a single item. This is especially useful for cases where words have different stems in singular and plural form, e.g. ”foot” and ”feet”, where stemming will take an unsatisfactory decision of not combining these into the same unit.
- *Stop word elimination* - removal of frequently used words such as ”a”, ”for” and ”the”. These words tend not to contribute to the overall textual meaning, and can therefore be removed. Although, this should be performed with caution, as stop words are domain-specific and some words often being listed as a stop word, such as ”not”, can change the entire meaning of the sentence if removed.
- *Noise removal* - in certain domains, the use of elongated words, hyperlinks, user mentions, hashtags, misspellings, emojis, and other unconventional characters can be classified as noise. Because of the highly domain-specific nature of these instances, there might be some cases where it is seen as beneficial to remove these, but be aware of the possible semantic meaning being lost.

2.3.2 Text Representations

One of the fundamental problems in text mining, information retrieval and natural language processing, is how to numerically represent the unstructured content in documents to make them mathematically computable. These representations can further be fed into machine learning algorithms as features. In this section, some of the commonly used textual representations will be presented.

TF-IDF

Term Frequency-Inverse Document Frequency (Jones, 2004) is a numerical measure of a term's importance within a corpus. The first part of the measure, the Term Frequency (TF), measures the number of occurrences of a term within a single document. Multiple versions of the weight exist, but the most commonly used is either the count alone (raw frequency) or the log normalized version $1 + log(tf_{d,t})$. The latter tends to be preferred because it makes them directly comparable to the IDF. Given the fact that an infrequent term is more selective than frequent terms, it can be beneficial to highly rank the rarest terms, that is, those occurring in the least amount of documents. This can be achieved by taking the inverse of the Document Frequency (DF), representing the number of documents in which a term occurs. With the Inverse Document Frequency (IDF) commonly represented as $log(\frac{N}{df_t})$, the components can be combined into the equation below.

$$\text{TF-IDF} = (1 + log(tf_{d,t})) * log(\frac{N}{df_t})$$

Bag of Words (BOW)

If the order or relationship between words has no significance, the Bag of Words (Harris, 1954) representation can be a decent choice. The method keeps track of the words that appear in a document and expresses them as a vector with the same size as the vocabulary. Using either one-hot encoding or the term frequency, the Bag of Words approach performs effectively in fields where simply the presence of words represents the content of a document.

N-Grams

N-grams are a way of representing text as a sequence of words of size n. By iterating word by word, combining n words together, the n-grams preserve the textual order to some degree for values higher than 1. Commonly applied representations are unigrams of one word, bigrams of two, or trigrams consisting of three words each.

Linguistic Inquiry and Word Count (LIWC)

Linguistic Inquiry and Word Count (Pennebaker et al., 2001) is a language analysis tool that places each word in a given text into one of more than 70 predefined categories. The application then calculates the percentage of words in each category, before utilizing the percentage distribution for deriving semantic meaning from texts. This can further

be used for classification and detection, as an attempt at personality modeling can be made based on the distribution of each category. Affect, occupation, pronouns, positive emotions, and negative emotions are a few examples of specified categories.

Word Embeddings

Another technique for representing text is word embeddings, which aim to group together vocabulary words with semantic similarity. Furthermore, embeddings address the problems associated with simpler methods' sparse vectors, inability to handle unidentified words, and lack of context-holding. Several implementations exist, such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), fastText (Bojanowski et al., 2016), and Elmo (Peters et al., 2018).

The first of these, Word2Vec, converts words to a vector representation using either Continous Bag of Words (CBOW) or Skip-Grams. Sliding a window over the sentences, CBOW uses the two surrounding words (context) to learn what word embedding should be given as output, while Skip-grams learns to predict the surrounding words given a word as input. In other words, these are opposite approaches to performing the same task. GloVe is an unsupervised approach that builds a co-occurrence matrix across the corpus, utilizing the term frequency. The fastText implementation differs from the former ones by using subwords of length n or single characters (n=1) as a base. As the subword might have been observed even though the full word was not present in the training data, this enables the model to correctly predict rare or even unseen words. Finally, the Elmo embedder assigns each token a function of the entire input sentence. Elmo is trained with a bi-directional LSTM network, which again is trained on a language model task for predicting the previous and next words. Elmo can distinguish between different meanings of the same word by using context. Additionally, Elmo can handle misspellings since the first layer deals with characters rather than words.

Regardless of using general pre-trained word embeddings or domain-specifically trained ones, their size is determined by the vector space's dimensions. Simply put, a larger vector space can hold more data. However, the performance is not proportionate to the size of the word embeddings, and a larger vector space does not always result in better performance. Hence, the challenge at hand should be taken into consideration when choosing the architecture, training method, and size of the word embeddings.

2.4 Machine Learning

Machine learning is a field within Artificial Intelligence (AI), which gives systems the ability to automatically learn and improve based on experience that has not been explicitly programmed. To train and adapt models for the generalization of data, which is the main essence of machine learning, one must first have a set of observations to explore potential underlying patterns. What is done next to find these patterns depends on the type of problem and data available, and hence which subcategory of machine learning one is dealing with. It is common to divide machine learning into three categories: supervised learning, unsupervised learning and reinforcement learning.

The goal of supervised learning is to create a model that is able to provide the correct label on unseen data. To achieve this, one must create a function (model), trained using pre-labeled training instances, for mapping given data to a target label. Typical examples of supervised learning are classification, where the computer program is tasked with mapping input to a discrete set of pre-defined classes, or regression, where the input is mapped to continuous numerical values (Goodfellow et al., 2016, p. 100–101).

Unsupervised learning is more of an exploratory approach. This subcategory of machine learning involves working with unlabeled data instances, that is, the data does not include a target class. Therefore, the aim becomes learning the structures of the data, in addition to finding hidden patterns. It is common for the models to group data based on given features, thus often used in cluster analysis with large amounts of data, where you want to group similar data points that at first glance do not seem to possess clear connections.

The third and final subcategory, being reinforcement learning, involves an agent trying to perform the correct actions in a given environment or *action space*. For such a model, it is common for the agent to be rewarded for correct choices, and punished for wrong ones. The aim of the actor would therefore be to increase its score, bringing the agent closer to the goal.

As we have a clear goal of identifying a set of certain characteristics, this would classify as a supervised learning problem. In this section, the most common algorithms and approaches from the field of supervised machine learning will be described.

2.4.1 Linear Regression

Linear regression is a supervised machine learning algorithm, with the aim of fitting a line to a set of data points. Hence, a linear regression model is given by the equation of a straight line:

$$y = ax + b$$

Linear regression can be modified to fit different kinds of problems depending on the data provided. This modification is done by choosing a suitable cost function. The most common cost function, Mean Square Error (MSE), squares the sum of errors taking the difference between the predicted value given by y and an actual data point. By minimizing such a cost function, the model will be able to place each data point closer to the ideal value. Some other commonly used forms of linear regression are ridge regression and LASSO. When the dataset is small, linear regression will quickly start to overfit the model leading to high variance. Ridge regression combats this by introducing a slight bias by penalizing to reduce variance, eventually leading to a model that is more generalized than the overfitted linear regression model.

2.4.2 Support Vector Machine

The method of Support Vector Machine (Hearst et al., 1998) builds on basic maximal margin classifiers. A maximal margin classifier is a classifier that tries to fit a hyperplane in such a way that the data points on either side and closest to the hyperplane have the largest possible distance to the hyperplane (Figure 2.4.1).

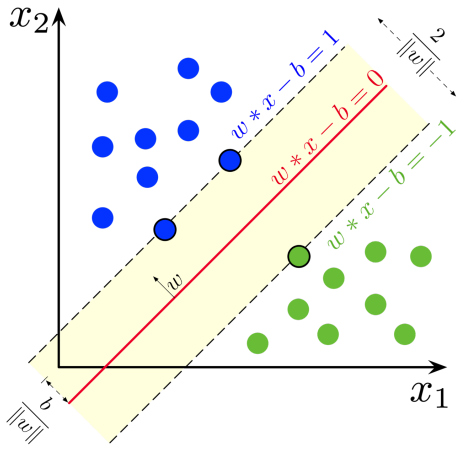

Figure 2.4.1: Support Vector Machine[1]

However, this classifier falls short when encountering outlier data. Imagine a case in 1 dimension where data points of class 1 usually reside around the values 0-2 and data points belonging to class 2 resides around the values 6-8. At this point, a maximal margin classifier will function well, since the classes are well separated. The problems start if we introduce data points belonging to class 1 that are much closer in value to the data points of class 2 than 1 (Figure 2.4.2). This outlier will force the maximal margin classifier into creating a separating threshold much closer to the values of class 2 than 1, leading to newly sampled points that are closer in value to class 2, possibly being classified as 1 instead!

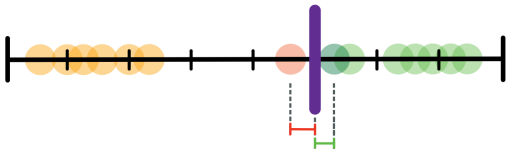

Figure 2.4.2: Maximal margin model weakness

To solve this problem Support Vector Machine step in. Two main differences are introduced. These are support vectors that allow for misclassification to benefit the larger amount of classifications (soft margin). And kernels that transform data points to higher dimensions to allow for the classification of data not clearly split into two separate parts. Support vectors are vectors that run along the extreme data points closest to the maximal margin vector in parallel. The actual maximal margin vector is now seen as more of a soft margin since support vectors allow for some misclassifications to benefit the larger classification problem as a whole. The final vector is found by solving an optimization problem.

[1]By Larhmam - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=73710028

2.4.3 Logistic Regression

The approach of Logistic regression fits a curve to a set of data, separating the data into classes, thus often used for binary classification problems. An example could be to guess whether or not a person is a man, given a certain height. The height would be the entered parameter, and the curve would give a probability of the person being a man based on the height provided.

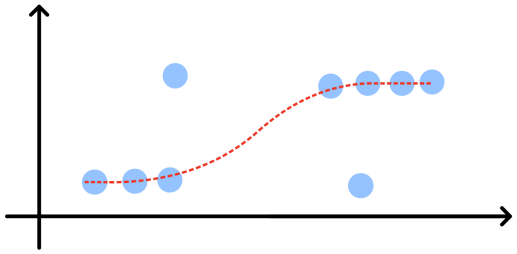

Figure 2.4.3: Logistic regression

If the x-axis in Figure 2.4.3 is a given person's height, the curve is interpreted as the probability from 0-1 of a person being male. The blue dots are the data points used to fit the logistic regression curve.

2.4.4 Decision Trees

Decision trees are a very common machine learning algorithm. Some of their main advantages are the simplicity of calculation and the ease of interpretation. They can be tweaked to be used for both classification and regression problems.

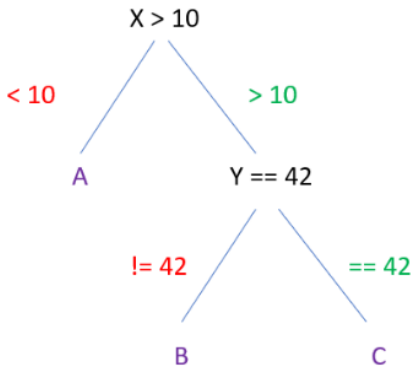

Figure 2.4.4: Decision tree

Decision trees consist of a set of nodes. Each node can either be an end node, a branch node, a leaf, or a splitting node. At each splitting node, there is a comparison being done that splits the data further. What variable or value to split on is determined by what split yields the highest information gain. This information gain is calculated by comparing the current node's *gini impurity* to the potential total impurity of the nodes resulting from an eventual split. The Gini impurity is given as

$$Gini = \sum_{i=1}^{J} p(i) * (1 - p(i)),$$

where $p(i)$ is the likelihood of picking a data point from a class in the given node. Now, let U be the total information gain of a split, G be the Gini impurity score at a node, N be the total amount of data points and n be a subset of the total data points. We can then express the information Gain of a split as

$$U = G_{start} - (G_{left} * (\frac{n_{left}}{N})) - (G_{right} * (\frac{n_{right}}{N}))$$

This splitting continues down the tree until a set stop condition is met. Common stop conditions are no longer being able to achieve a satisfactory amount of information Gain from a split or that a sufficient degree of certainty in classification has been reached in a given node.

2.4.5 Naive Bayes

Naive Bayes is a probabilistic machine learning classifier. It is a simple implementation of Bayes' theorem applied to statistical data. The most commonly used forms of Naive Bayes are Multinomial Naive Bayes (MNB) and Gaussian Naive Bayes (GNB). For the implementations covered in this project, the Multinomial Naive Bayes is relevant. In Natural Language Processing, MNB is often used to classify terms or passages of text based on a Bag of Words. The classifier calculates the probability of the data provided belonging to each class included in the classification problem. The final result is the class to which the data is most likely to belong.

The probability of an element A belonging to class B is given by Bayes' theorem:

$$P(B|A) = \frac{P(A|B) * P(B)}{P(A)},$$

where $P(A|B)$ is the probability of observing A in class B. This metric is learned from training data.

2.4.6 Neural Networks

In the most recent years, the best-performing artificial intelligence systems have resulted from a technique called deep learning. Though the name *deep learning* is of newer existence, the name is in fact a new name for an approach called neural networks, which have been going in and out of fashion for almost 80 years. The first mathematical model of an artificial neuron, built upon the ideas of Alan Turing, was proposed by McCulloch and Pitts (1943). The *McCulloch-Pitts Neuron* though has some limitations, as it can only represent non-weighted boolean functions with a hand-coded threshold. Overcoming these limitations, the first real implementation saw the light of day fifteen years later in the form of a machine by F. Rosenblatt, called the Mark I Perceptron. One may argue that the whole area of deep learning and neural networks build on the suggested model by Rosenblatt (1958), called the *Perceptron model*.

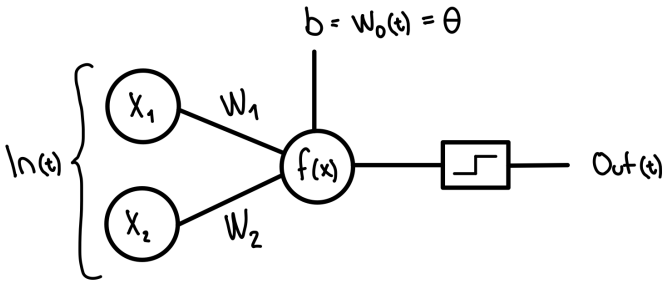

Figure 2.4.5: The perceptron model

Rosenblatt's perceptron model is illustrated in its most simple nature in Figure 2.4.5. Provided with two input variables, here noted as x_1 and x_2, each entry is multiplied with a weight. This yields the main improvement, being the introduction of a measure of importance, seeing as some of the inputs may weigh more than others on the decision of the desired outcome. The node then proceeds to aggregate all the inputs and their belonging weights, i.e. taking the weighted sum, before adding the bias. Instead of coding a threshold for the final step function, Rosenblatt adds a new weight with the same intention of shifting the activation function, but rather as an adjustable (and learnable) value. The function $f(x)$ in Figure 2.4.5 can therefore be denoted by the equation

$$f(x) = \sum_{i=1}^{n} x_i w_i + b = \sum_{i=0}^{n} x_i w_i,$$

where the bias is included as an additional weight w_0, attached to a dummy input x_0 having assigned the value of 1. The output of this summation is then passed to a (Heaviside) step function, for deciding whether or not the neuron should fire.

In addition to improving the perceptron model with the addition of weights, a major achievement of Rosenblatt was to come up with a fairly simple and yet relatively efficient algorithm, enabling the perceptron to learn the correct synaptic weights from examples. This lays the foundation of Feature Learning, also known as Representation Learning, where the model itself can learn from a set of examples given as raw data.

Still, one limitation remains, as pointed out by Minsky and Papert (1969), being the impossibility of computing XOR due to it not being linearly separable. The paper argued that the proposed algorithm would not work as a model would need to have multiple layers, which the authors deemed too computationally expensive given the available hardware at the time. Although the potential is toned down, which is widely believed to have led the way to what is known as the AI Winter, the later work on multilayer perceptrons, combined with the immense rise in computer power and available data, makes the way for what is known as deep learning today.

Feedforward Neural Network (FNN)

A feedforward neural network is in the simplest form a composition of multiple layers of artificial neurons. Two distinctions to the formerly defined perceptrons could be made. True perceptrons use the threshold step function, being a special case of the artificial neurons, which in fact can employ any arbitrary activation function. In addition, a true perceptron performs binary classification, whereas an artificial neuron in general is free

to either perform classification or regression. The neurons, also known as nodes, are then layered into three or more layers; with a single input- and output-layer, and one or more hidden layers. For each layer, the resulting output is passed as input to the subsequent layer, normally being a product of the propagated information and the weight of the corresponding edge, in a feedforward manner. The process of sending the data through the pipeline is called forward propagation or forward pass.

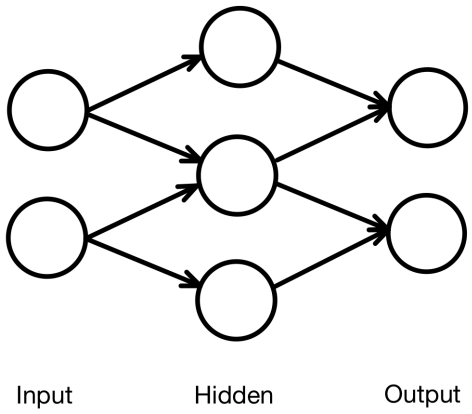

Figure 2.4.6: Feedforward Neural Network

For learning to occur, that is, the adjustment of weights, one performs a process called backward propagation or backward pass. This optimization technique compares the model's output to the correct output on labeled data, yielding an error, that is, the distance between the correct and predicted value. Going backward, each weight is adjusted based on the derivative of the error with respect to the weight, also called the gradient. The process also goes by the name of gradient descent, trying to decrease the error/loss. The training, being one cycle of the forward and backward pass, continues on new data until the network converges.

It turns out that a neural network consisting of just one hidden layer is enough to approximate any function, although the layer has to be unreasonably large and does not necessarily learn or generalize properly (Goodfellow et al., 2016). By *approximation of any function* it is meant that by using enough nodes one can always find a neural network that has an output, $g(x)$ that satisfies $|g(x) - f(x)| < \epsilon$ for all inputs, x. In other words, the approximation will be satisfactory for every possible input. Although the theorem says that one layer is enough, it is common to use several layers in practice to avoid problems during training such as under-/overfitting and exploding-/vanishing gradients. The optimal number of nodes and hidden layers are often found through experimentation, being heavily dependent on the available data and problem at hand.

2.5 Evaluation Metrics

For being able to determine whether machine learning models actually reach their goal of either regression or classification, one needs to have a measure of performance. This section aims to describe some of the relevant evaluation metrics for said task. For classification problems, this includes the metrics of Accuracy, Precision, Recall and F-Score, while different forms of error are used in problems of regression.

2.5.1 Accuracy

The first performance-related metric, Accuracy, is used in a wide variety of machine learning tasks. By quantifying the number of correctly classified instances, divided by the total number of instances, one can get a sense of how many correct classifications the model makes, thus also knowing the number of misclassifications. An equation for the metric is presented below.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

By looking at the equation, one can observe that it is made up of four distinct values, namely TP, TN, FP, and FN. True Positive (TP) is the number of correctly classified positive instances, while True Negative (TN) represents the instances that are correctly predicted as negative. Thus, the same logic applies to the other two, yielding that False Positive (FP) represents the incorrectly classified positive instances, and that False Negative (FN) is the number of negative instances being misclassified. These four values can be better visualized in a tabular form, as seen in Table 2.5.1.

		Ground Truth	
		Positive	Negative
Prediction	Positive	TP	FP
	Negative	FN	TN

Table 2.5.1: Matrix illustrating the four outcomes in binary classification

2.5.2 Precision and Recall

Another performance metric commonly used in classification problems in both information retrieval and machine learning is precision and recall. Precision, or the positive predictive value, is the portion of relevant instances among the retrieved instances. To elaborate, the measure tells us *how many of the returned items are relevant*. The equation for precision can therefore be presented as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

where the elements True Positive (TP) and False Positive (FP) are the same as presented in the former subsection. Recall, also known as sensitivity, represents the portion of relevant instances that were retrieved. This can be seen as the opposite of precision, where one rather focuses on *how many relevant instances are retrieved*, considering the

entire set of elements/documents, instead of only looking at correctly classified instances in the retrieved subset. The equation for the recall is presented below, with both the True Positive (TP) and False Negative (FN) being equal to the definitions already described.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

2.5.3 F-score

The F-score attempts to be a measure of accuracy, including both precision and recall. In its simplest form, the F_1 score is the harmonic mean of the two given metrics. Furthermore, the more generic F_β score implements an additional weight, β, valuing either precision or recall more than the other. The output is a score on a scale between 0 and 1, where 1 indicates the ideal precision and recall. In its general form, the equation for the F-score can be introduced:

$$F_\beta = \frac{\text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}} \times (1 + \beta^2)$$

2.5.4 Error

As mentioned in the introduction, the former metrics are used in classification tasks, while not being an applicable measure for regression problems. Instead, different forms of measuring error can be used for evaluating the performance of a regression model. The three most commonly used measures are the Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE). The first of these, Mean Absolute Error, is a measure of the absolute average distance between the expected value, and the value given by the model. The next one, Mean Square Error, takes a similar route but measures the average square distance between the predicted value and the expected value instead. Finally, the Root Mean Square Error builds upon Mean Square Error, taking the square root of MSE, thus penalizing larger errors more than smaller ones, which is a neat feature.

3. Related Work

In this chapter, the research method used during the project will be described, securing the reproducibility of our work. Firstly, the structure of the conducted literature review will be presented, followed by our findings and evaluation of the resulting literature. Secondly, this section will include research on preprocessing and feature extraction. Finally, methods used for identification are presented.

3.1 Literature Review

During this project, two methods of literature review have been used. The methods used were a structured literature review and the snowballing approach. In this chapter, the use of both methods will be presented.

3.1.1 Structured Literature Review

The structured literature review was conducted based on a suggested outline by Kofod-Petersen (2018). This subsection will present the literature retrieved from the structured literature review as well as the five proposed phases.

The first phase is the planning phase, which involves specifying the research questions and developing a review protocol. The second phase is where one conducts the actual review, and includes identification of research, selection of primary studies, quality assessment, data extraction, monitoring, and data synthesis. In the third and final phase, one is tasked with communicating the newly acquired knowledge, and should therefore include a specification of a dissemination strategy, a formulation in a report, and an evaluation of the said report. Each step and the results of the conducting phase are described in the following subsections.

Step 1: Identification of Research

The collection and exploration of literature were done using Google Scholar. To identify literature relevant to the project, two queries were constructed.

Q1: `(school shooting OR gun violence OR shooting OR violent act) AND (machine learning OR computational OR ai OR prediction OR classification) AND (social media OR 4chan OR twitter OR reddit OR facebook OR gab OR weibo)`

The first query focuses on literature concerning the automatic prediction of individuals predisposed to violence who could potentially become lone wolf perpetrators.

Q2: `(machine learning OR computational OR ai OR prediction OR classification) AND (social media OR 4chan OR twitter OR reddit OR facebook OR gab OR weibo) AND (personality OR personality traits OR personality profiling)`

The second query tries to extract literature concerned with automatic personality profiling and prediction using data mined from social media. Both queries were constructed based on a set of key terms deemed highly relevant to the problem domain. The key terms are presented in Table 3.1.1.

Group 1: Violence	**Group 2: Machine Learning**	**Group 3: Social media forums**	**Group 4: Personality prediction**
School shooting	Machine learning	Social media	Personality
Gun violence	Computation	4chan	Personality traits
Shooting	AI	Twitter	Profiling
Violent act	Prediction	Reddit	
	Classification	Facebook	
		Gab	
		Weibo	

Table 3.1.1: Search terms for the conducting phase described in Kofod-Petersen (2018)

Step 2 + 3: Selection of Primary Studies

To keep the literature review feasible, the articles retrieved from the queries were limited to the first 50. This was done after removal criteria were applied to the retrieved set. The criteria for removal were duplicate entries and if a study was published before 2013.

The studies retrieved from the initial search were finally subject to a study quality assessment to assure only relevant papers were included. The inclusion criteria and quality assessment criteria used can be found in Appendix A and Appendix B. The final set of included studies consists of 12 papers. The studies retrieved from the structured literature review are presented in Appendix C.

Step 4 + 5: Data Extraction, Monitoring, and Synthesis

Finally, the data retrieved from the corpus resulting from the SLR is summarised in Table 3.1.2.

Table 3.1.2: Data synthesized from SLR

Id	Title	Author(s)	Year	Algorithm	Features	Dataset(s)	Summary
1	Identifying Warning Behaviors of Violent Lone Offenders in Written Communication	L. Kaati, A. Shrestha & T. Sardella	2016	Adaboost w. classification trees	LIWC	Texts from schoolshooters.info, publicly available written comm. from mass shooters, posts from the white-supremacist forum Stormfront + more forums	Found differences in psychological features retrieved from LIWC for non-violent offenders vs. violent offenders. The anger feature and psychological process feature differentiation were important in separating offenders and non-offenders.
2	Profiling school shooters: automatic text-based analysis	Y. Neuman, D. Assaf, Y. Cohen & J. L. Knoll	2015	KNN, Tree classification (CHAID), binary logistic regression	Vectorial semantics approach used as features	Blogs Authorship Corpus, Texts written by school shooters gathered from schoolshooters.info	Using a more statistics-based approach, the authors constructed a ranking of texts most likely to have been written by a school shooter. Using their method, the school shooters' texts were all contained in the first 210 texts of the ranking = 3% of their total corpus.
3	A multi-label, semi-supervised classification approach applied to personality prediction in social media	A. C. E. S. Lima & L. N. Castro	2014	Naive Bayes, SVM, MLP	LIWC, MRC	Obama-McCain Debate, ++	Accomplished an accuracy of around 83% on Big 5 Personality traits

4	Personality Predictions Based on User Behavior on the Facebook Social Media Platform	M. M. Tadesse, H. Lin, B. Xu & L. Yang	2018	XGBoost, SVM, Log-Reg, Gradient Boosting	LIWC, SPLICE, SNA (Social Network Analysis)	myPersonality	Found that extroverted users tend to use fewer, but more positive words. XGBoost outperformed other models, most notably on the extroversion feature (78.6%). The SNA feature set outperformed traditional linguistic feature sets such as LIWC.
5	Personality classification based on Twitter text using Naive Bayes, KNN and SVM	B. Y. Pratama & R. Sarno	2015	MNB, KNN, SVM	Vector space model	User data and posts retrieved from Twitter, myPersonality (All translated to Indonesian)	Used a binary classifier for each class in the Big 5 Model. After tokenization, stemming, filtering of stop-words, and weighting, they achieved a max accuracy of 60% with MNB, ultimately failing to improve on previous work.
6	Recognising Personality Traits Using Facebook Status Updates	G. Farnadi, S. Zoghbi, M.-F. Moens & M. De Cock	2013	KNN, SVM, NB	LIWC, users' friend network, time stamps (SNA)	myPersonality and Facebook status updates	Used binary classifiers for multi-class classification of Big 5 Model. They found that there is no single kind of feature that give the best results for all Big 5 personality traits. They argue ML approaches to personality recognition are generalizable across domains.
7	Predicting personality traits of Chinese users based on Facebook wall posts	K.-H. Peng, L.-H. Liou, C.-S. Chang & D.-S. Lee	2018	SVM, NB, Log-Reg	Used BOW approach for features / TF-IDF	Online survey created by authors, Facebook user data/posts	Used Big 5 with SMOTE to overcome class imbalances. Achieves surprisingly high accuracies with around 95% + being common for all algorithms used.

8	Personality Prediction System from Facebook Users	T. Tandera, H. Derwin, S. Rini, W. Yen & L. Prasetio	2017	NB, SVM, Log-Reg, Gradient Boosting, LDA, MLP, LSTM, GRU, CNN, LSTM + CNN	LIWC2015, SPLICE, SNA features for old methods, deep learning with open vocab (word embeddings - GLOVE)	myPersonality + 150 manually sampled Facebook users	Authors compared deep learning approach with an open vocabulary approach using GLOVE to traditional machine learning algorithms. Study shows that given large enough datasets deep learning approaches can outperform traditional ml approaches.
9	Mining Facebook Data for Predictive Personality Modeling	D.Markovikj, S. Gievska, M. Kosinski & D. Stillwell	2013	SVM, Simple Minimal Optimization, MultiBoostAB / AdaBoostM1	Social Network features (friends, likes ...), demographic (age, gender...), LIWC, POS Tag, Afinn, H4Lvd	myPersonality	Confirmed previous results, no best set of features for all classes. Selecting more discriminative features improves accuracy.
10	Automatic Personality Assessment Through Social Media Language	G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar & M. E. P. Seligman	2014	Ridge Regression	Open vocab. features (words + phrases, topics). Dim. reduction: Univariate feat. selection, Rand. principal comp. analysis	77,556 users of myPersonality	Used NEO-PI-R five factor model. The study applied ridge regression with the mentioned features to attain state-of-the-art performance. Concludes that open vocab. captures more semantic meaning.

11	Personality Traits on Twitter or How to Get 1,500 Personality Tests in a Week	B. Plank & D. Hovy	2015	Log-Reg classifier	Binary word n-grams, metadata (followers, n tweets, statuses...)	Tweets from Twitter users self-identifying with an MBTI	Study found that linguistic cues are strongest indicators of personality. Meta-features can help add to accuracy.
12	25 Tweets to Know You: A New Model to Predict Personality with Social Media	P.-H. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju & V. Sinha	2017	Gaussian Processes	Word Embeddings (GloVe)	Tweets from 1.3k Twitter users surveyed via web-app. IPIP format	Word embeddings and Gaussian Processes approach outperforms previous state-of-the-art (LIWC + Ridge Regression). Dense word embeddings make performance better on short texts.

3.1.2 The Snowballing Method

As described in Section 1.3, the snowballing approach was used for finding additional relevant literature. Following the procedure described in Wohlin (2014), an initial start set of three documents were made using Google Scholar. The first of these was the paper by Kaati et al. (2016), an article put forward in Table 3.1.2. Furthermore, the work of Neuman et al. (2020) was found in an author search from one of the initial literature review findings (Neuman et al., 2015). Finally, the paper by Simons and Meloy (2017) was included for trying to supplement with documents regarding threat assessment, adding a level of diversity as mentioned by Wohlin (2014). This particular paper was selected due to being a reference in Neuman et al. (2020), in addition to being a seemingly popular article and part of a highly rated book associated with the field. Using this relatively small start set, justified by the fact that the snowballing approach was intended to be supplementary to the SLR, both the backward and forward snowballing approaches were followed for each of the articles. The resulting papers, presented in Appendix D, can serve as additional literature to the work already presented.

In the following sections to come, the findings of performing both SLR and the snowballing method will be presented.

3.2 Datasets

The internet and social media platforms are huge sources of data that can be used in different NLP tasks. There has already been done substantial work on collecting datasets from social media for use in personality profiling, however, there seems to be little to no datasets directly linked to profiling school shooters. In this chapter, the datasets deemed to be relevant to further work will be presented.

3.2.1 Personality Profiling Datasets

The large amount of data gathered from social media allows for an insight into people's lives and habits. There have been constructed several datasets pertaining to personality profiling based on social media activity. The use and creation of these datasets will be covered in the following subsections.

Facebook Datasets

Facebook is a platform that offers its users the ability to join groups, chat and post messages to different forums, leaving behind traces of the user for everyone to see. The myPersonality dataset allowed users to take a short questionnaire resulting in a Big 5 personality score upon completion. This is perhaps the most used dataset linked to Facebook. Although it is now discontinued, the dataset contained user posts, their own personality score, and user data (friends, groups, likes, etc.). According to the dataset's authors, the application had at least 6 million users complete the questionnaire, with about 40% of them agreeing to their data being used for research.

Tadesse et al. (2018) used a subset of 250 users including 9917 status updates of the myPersonality dataset. Their final dataset contained user information, status updates,

Big 5 personality labels and Social Network Analysis (SNA) features. They attempted to improve on previous results by feeding SNA data into an XGBoost architecture. Pratama and Sarno (2015) retrieves a subset of 250 myPersonality users accompanied by 10000 user posts in total. These posts were then translated into Indonesian and appended with Twitter posts from the corresponding person. The authors assume that not too much meaning is lost in translation. The study done by Farnadi et al. (2013) opts to use the same amount of users as both Pratama and Sarno (2015) and Tadesse et al. (2018). 9917 posts are extracted and used in combination with a corpus of essays annotated with personality traits collected by Mairesse et al. (2007). Their approach attempts to incorporate spatial data as well as user metadata to predict users' Big 5 personality scores. Tandera et al. (2017) compares the efficacy of traditional machine learning algorithms such as SVMs and newer deep learning architectures. For their dataset they also retrieved 250 users, extracting 10000 status updates. Their final dataset was a combination of these 250 users as well as 150 manually collected users all annotated with Big 5 personality scores. These users were collected using the Facebook API. The largest study using myPersonality found in the literature review was by Park et al. (2014) using 71,556 users in the myPersonality dataset. External personality tests were used as validation for the ridge regression method used.

Twitter Datasets

Twitter is another social media giant with millions of users active every day. The social media platform is a microblogging site that allows users to write posts with at most 280 characters. The platform allows users to write what they want with little filter, making it easy to express their true feelings whatever they may be. There have been several datasets created to take advantage of this quality of Twitter to use for personality prediction. Plank and Hovy (2015) uses a novel dataset of over 1.2 million tweets annotated with Myers-Briggs personality types and gender to attempt to predict personality dimensions. The dataset was constructed by the authors and made available on a public code repository for others to use. It is worth noting that the MBTI types annotated are taken from the users' own Twitter account description. Arnoux et al. (2017) constructs a novel corpus from a self-made application. This dataset contains tweets from 1323 unique users annotated with their Big 5 personality traits. Lima and de Castro (2014) proposes a multi-label classifier approach using the Big 5 Model as class labels. For performance evaluation, they utilize different sentiment analysis datasets. The Obama-McCain Debate (OMD) contains 3238 tweets from the 2008 presidential campaign annotated with sentiment scores. The tweets are accompanied by date and user identification. The SemEval2013 and Sanders Analytics datasets were also used in their study. The latter seems to be discontinued but has been backed up on GitHub (Sanders Analytics, 2013).

Other authors using Twitter datasets have gone for an approach where they combine Twitter data with user profiles found on Facebook. Examples of this can be seen in the *Facebook datasets* chapter above.

3.2.2 Datasets on School Shooters and Lone Wolf Perpetrators

With such a large amount of users active on social media every day, it is only natural to assume that also potential lone wolves or school shooters utilize them. As of the time of writing, there are no publicly available datasets directly linking social media posts and accounts to known lone wolf perpetrators known to the authors. Work done in this area

has relied on the construction of novel datasets such as the work of Ekwunife (2022). Ekwunife constructed a dataset consisting of 500 tweets related to five different mass shooting incidents in the U.S. Neuman et al. (2020) and Kaati et al. (2016) attempt to overcome this scarcity of social media data by using established databases of known school shooters. The database used was schoolshooters.info (Langman, 2022). The database contains an overview of all registered school shooters in the U.S. along with documents relevant to each case. The documents used in both studies were written material produced by the school shooters prior to an incident.

3.3 Preprocessing and Feature Extraction

This section will present the techniques used for preprocessing and feature extraction for the purposes of detecting lone wolf perpetrators and personality prediction.

3.3.1 Preprocessing

Preprocessing is a crucial step of any NLP task, however, the amount and extent of the preprocessing vary depending on the task at hand. Due to the inherently noisy nature of social media posts, the preprocessing step usually has to weigh the amount of processing to be done up against the potential loss of semantic meaning. Park et al. (2014) chooses to perform a minimal amount of preprocessing for this reason. They utilize the emoticon-aware tokenizer made by Potts (2011) which preserves punctuation. Palomino and Aider (2022) argues that the removal of punctuation could hurt sentiment analysis, due to special characters or emoticons potentially conveying sentiment. The same paper stresses the importance of preprocessing and found that a combination of common preprocessing steps indeed does increase the accuracy of sentiment analysis on social media texts. Due to the concerns of over-processing listed above, most papers found in the literature review follow the norm of lowercasing, hyperlink elimination, special character removal, and finally removal of punctuation. Further substantiating this concern Plank and Hovy (2015) found that the removal of stop words harmed performance when trying to model Big 5 personality traits. This is however an evaluation that should be done depending on the dataset used.

3.3.2 Feature Extraction

Features used for automatic detection and personality prediction can be divided into two main categories; linguistic features and user-specific features. This section briefly presents the features used in previous work.

Linguistic Features

Linguistic features are features that can be extracted from written text. Plank and Hovy (2015) found that in their study, the strongest indicator of personality was linguistic cues. There should thus be sufficient work done to extract the cues possible to ease the process of personality prediction. The following subsection explains the linguistic features authors found through the literature review utilized in their studies. Linguistic Inquiry and Word Count (LIWC) have been widely used for linguistic cue extraction. LIWC is a large framework with fine-grained features that can contribute to personality

prediction. However, not all features are discriminative enough, leading to the need for feature selection. Kaati et al. (2016) applied LIWC features to screen for lone offenders in written text. They found that using the Mahalanobis distance to rank each LIWC feature and selecting the 11 highest-ranking features resulted in an increase in accuracy. Important features were articles, prepositions and negative emotions. From the work of Tadesse et al. (2018), the correlation between LIWC features and personality is further analyzed using pearson correlation coefficient. An interesting finding is that LIWC features seem to be more discriminative or correlated with personality than the newer SPLICE dictionary.

Although older, the traditional statistical approach of Bag of Words and TF-IDF weighting is still used as an important tool in personality prediction, often combined with Naive Bayes classifiers. In the paper *Predicting personality traits of Chinese users based on Facebook wall posts*, Peng et al. (2015) applies a simple TF-IDF weighting process to represent each user's posts as a vector. It is worth noting that the words given to the BOW model were tokens constructed by the *Jieba Chinese text segmentation* tool (Junyi, 2020).

Another method of statistical analysis is the vectorial semantics approach of Neuman and Cohen (2014). They argue the features often used in other NLP tasks for personality profiling, namely LIWC and n-grams, are too inflexible to be able to be accurate in all situations. E.g. a person might behave and write differently in a school setting and a private setting. With this in mind Neuman et al. (2015) proposes vectorial semantics as an alternative to traditional machine learning to detect lone wolf perpetrators. Vectorial semantics assumes that a word is closely correlated to the words often co-occurring with it. This assumption allows one to model a word as an n-dimensional vector, which dimensions are determined by the amount of co-occurring words chosen. The direction and magnitude are determined by the number of times each co-occurring word appears in tandem with the modeled word. The words modeled can be chosen at will, from any corpus of text, making vectorial semantics more flexible than other commonly used alternatives. In their work, they propose 13 different vectors based on previous research.

The aforementioned vectorial semantics method is an example of an open vocabulary approach to NLP. Arnoux et al. (2017) uses an open-vocab approach utilizing the popular GloVe framework, as presented in Subsection 2.3.2. The GloVe framework uses the same steps as the vectorial semantics presented above, requiring a pass over the corpus to be used to learn word co-occurrences. However, the authors find that combining the GloVe vectors leads to an average 33% increase in accuracy when combined with gaussian processes as their ml model.

User Specific Features

User-specific features are features unique to a user profile on a social medium. These features can help provide further insight into a user's life and ultimately personality. Available features vary depending on the social media. The myPersonality dataset provided researchers with data on the number of friends, posts, and likes. The same type of data can be extracted from Twitter with the number of followers, amount of likes, and posts being publicly available data. Utilizing this data can further help improve accuracy in personality prediction tasks. Plank and Hovy (2015) found that the addition of user-specific data such as gender and number of followers improved accuracy when predicting personality traits. Further corroborating this, the work of Markovikj et al. (2013) shows that there exists a significant correlation between user-specific features and personality

traits. The Big 5 personality model's extroversion was the highest correlated trait, being tightly linked to amount of friends and the number of groups a user is a member of.

3.4 Models Used

The choice of model can in many cases be the most crucial factor in whether or not one manages to extract the latent information contained in the given features. In previous work, a wide variety of machine learning models have been utilized. As mentioned in Section 3.2, Tadesse et al. (2018) used an XGBoost model on labeled samples from the myPersonality dataset. This model outperformed the three baseline models used, which were Support Vector Machine (SVM), Logistic Regression and Gradient Boosting. Also using the same dataset, Tandera et al. (2017) sought to find the best performing model, comparing both traditional and newer machine learning models. The models categorized as traditional algorithms were Naive Bayes, Support Vector Machine (SVM), Logistic Regression, Gradient Boosting, and Linear Discriminant Analysis (LDA). They were all tested using a 10-fold cross-validation approach, dividing the dataset into a 90/10 train/test split. The newer models used for comparison were all deep neural networks. The architectures used were Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and a one-dimensional Convolutional Neural Network (CNN). Their study found that the deep learning networks, especially their implementation using an LSTM connected with the 1D CNN architecture, achieved the best results. Kaati et al. (2016) utilize an Adaboost model with classification trees on features from LIWC and undisclosed topic models, while Neuman et al. (2015) preferred the use of statistical models, namely a binary logistic regression model, a tree classification with CHAID and 10-fold cross-validation, and K-Nearest Neighbors with the same cross-validation. It is safe to say that previous research includes the exploration of both traditional and more modern deep learning techniques.

4. Discussion

This chapter presents a discussion of the research identified. It aims to serve as an explanation of the application of the techniques and data found through the literature review as well as the usefulness of these in the context of our own future research goals.

4.1 Datasets and Data Availability

After conducting a literature review, there does not seem to exist a dataset connecting known solo perpetrators to social media accounts. Previous studies on the topic of detecting school shooters and solo perpetrators have also stated that there does not seem to be much work done here (Neuman et al., 2020), making the absence of such datasets a fact. Although not the exact data that was desired for this study, the same article by Neuman presents a database of school shooters. As presented in Section 3.3 this dataset, schoolshooters.info, provides information on known school shooters and written documents produced by them before the incident. In Section 3.3 it is presented that the language used in social media may be different from the language used in other settings due to its informal nature. One could argue that this might affect the usability of the schoolshoters.info database for the purposes of this project. On the contrary, the documents presented are often diaries or written notes left behind, not intended for others to see, again prompting an informal use of language. There could therefore be merit in including this database for further research.

Other databases concerning school shooters and mass shootings do exist, however, there seems to be a limited amount of textual data accompanying these. Some examples of such databases are *The Washington Post's database of school shootings* (Cox et al., 2022) and the *K-12 school shooting database* (Riedman, 2022). These provide useful insights into where and when school shootings have taken place, but lack accompanying documents that could be analyzed for the purposes of feature extraction. Despite being void of textual data, one could theorize about using geospatial data as extra features. For example, if a given geolocation has a higher occurrence of solo perpetrator incidents this could weigh in on the final decision.

In contrast to the scarcity of datasets containing textual data from solo perpetrators, there is an abundance of datasets available to the public for the purposes of personality profiling. Since future work will look into the task of automatically detecting solo perpetrators, these datasets could be used as resources for texts generated from users of the opposing class, non-solo-perpetrators. They could also prove to be valuable resources for further study into the field of personality prediction should this be necessary.

With the lack of textual data produced by solo perpetrators in mind, an obvious solution may be the manual collection of data by the authors themselves. The work done during

this project has called the feasibility of collecting textual data directly linked to lone wolf perpetrators into question. The possibility of manually going through databases containing known offenders to then do a search for their social media accounts is there. It would however be likely that the time needed to collect enough data on the subject would be tremendous, making this a task that should be carefully considered before one decides whether or not to go ahead with it.

4.2 Preprocessing

After finding the data needed to go forth with personality prediction and trait detection, the data needs to be preprocessed for downstream tasks. The techniques used seem to vary from study to study due to the different models and features used. Depending on the model chosen, future work could perform as little preprocessing as possible on the textual data. With recent advances in large language models such as BERT or RoBERTa, comes the use of tokenizers that benefit from all parts of a document being present, including capital letters and punctuation due to these being of potential semantic significance. However, it is not a given that the use of such large transformer models is the best solution to the problem at hand due to their large size, driving up computational cost and time. With this in mind, one could look to utilize previous machine learning algorithms that require a more rigorous preprocessing pipeline beforehand. Considering the related work done on preprocessing it could be beneficial to do as much sanitization as possible, removing noise that is likely to appear in data collected from social media. The preprocessing steps presented in Subsection 2.3.1 seem like good options for this purpose. However, as previously stated, Plank and Hovy (2015) found that removing stop words hurt accuracy scores when using a logistic regression model. This leads to the belief that the preprocessing steps taken in future work should be decided experimentally due to the lack of literature directly concerning preprocessing for the purposes of automatic detection of lone wolf perpetrators.

4.3 Choice of Features

During the literature review, it was found that the most used features were linguistic features. Results showed that linguistic features were the best at determining personality traits when used in personality prediction models. However, there still is a choice to be made about what linguistic features to use. Several papers reported good results utilizing LIWC features, but still had to do feature selection to filter out indiscriminative features. Although well functioning, LIWC features seem to be very static in nature given that LIWC is a predefined dictionary. Some alternatives found to combat this is open vocabulary approaches using GloVe, n-grams, or vectorial semantics. If one were to proceed with further work using vectorial semantics, not utilizing GloVe, one should take great care in selecting the correct vectors for the problem domain. Although more involved, it seems this approach could yield positive results in the search for a method for lone-wolf screening (Neuman et al., 2015).

Further review of the literature collected suggests that there could also be merit in including non-linguistic features such as user-specific data (SNA). Including metrics along the lines of amount of friends and likes had shown positive results when combined with linguistic features. As stated earlier, user-specific features did not outperform the linguistic

features alone, but should instead be considered as an addition to further improve experiments. With this in mind, a natural conclusion would be to include both types of features in an eventual experiment. Unfortunately, this is not always feasible since the amount of data available about users varies widely from user to user. Some users may have made their account details or even their entire account private, making data extraction difficult. Therefore one should investigate how this might affect a data collection process before attempting to collect both types of data for all subjects.

4.4 Choice of Models

Judging from the results of previous work, there does not seem to be a clear-cut answer to what model one should choose for an NLP task. Several studies have concluded that a certain model performs better on e.g. the myPersonality dataset than previous ones. The performance difference is however quite minimal in most cases, with most approaches hovering around a 5% deviation from the observed norm of the results collected. This makes it hard to choose what model to focus on for further work. However, there have been promising studies showing that deep learning architectures may outperform previous traditional models, making this a potentially interesting area to point our attention to.

With recent advances in large language models, it could be beneficial to consider their usefulness for the purpose of personality prediction. This should however be after a consideration of the tradeoff that comes with the use of such models. These models tend to be quite slow both in training and inference, making their use on large datasets infeasible unless used on powerful machines or clusters. Depending on how the system developed in an eventual Masters's Thesis is desired to operate, the use of large transformer models could indeed have potential.

4.5 Difficulties Detecting School Shooters

Referring to the presented articles obtained from our literature review may leave one with the impression that the literature concerning this cross-section of the fields of psychology and machine learning is quite sparse. Indeed, this is what has been troubling us as well. A possible reason for this could be that a school shooting is a rather infrequent occurrence compared to other violent crimes (Neuman et al., 2020), making approaches needing large volumes of data hard to implement. This could make it difficult to implement traditional machine learning approaches to address the problem at hand. Furthermore, emphasis should be given to the fact that given the low prevalence of solo perpetrators, one could argue that a high rate of false positives is inevitable (Neuman et al., 2019). This statement builds upon the fact that some psychological characteristics often linked to shooters may not be discriminative enough to serve as an informative marker (Neuman et al., 2015). Due to these difficulties, this subsection is therefore dedicated to exploring the work already done on detecting school shooters with or without machine learning.

In the work of Kaati et al. (2016), they discuss the difficulties related to the fact that they are looking for a single person who could be classified as at risk of committing targeted violence. The first one is access to written communication. Some violent lone offenders publish a manifesto before their attack (Gill et al., 2014), but this is not the case for every perpetrator there is. Secondly, there does not exist a common profile, as a violent offender can represent any ideology, ethnicity, shape, or size (Kaati & Svenson, 2011). Thirdly, as

they act alone or with minimal help from others, relevant communication would typically be harder to retrieve, and would therefore be more difficult to identify (Bakker & de Graaf, 2011). Despite the difficulties mentioned, the authors refer to the work of Brynielsson et al. (2013), which found that there could exist *weak signals* in written communication preceding an attack. Using a set of features that are psychologically meaningful based on the LIWC analysis tool, they found that some psychological categories are different when comparing regular texts to the ones written by violent lone offenders (Kaati et al., 2016, p. 7). This supports the study's initial claim that violent lone offenders exhibit certain psychological warning behaviors that can be viewed as signs of an elevated or accelerating risk of committing targeted violence (Kaati et al., 2016, p. 1).

Going more into the detail of the utilized methodology, one could take inspiration from the datasets collected for their experiments. The article describes four datasets used in particular. The first one is the collection of 46 texts written by 32 subjects, where each individual is categorized as either a school shooter, mass murderer, or ideologically motivated offender. As for the other three datasets, a collection of 54 personal blogs, 108 Stormfront posts, and 108 posts from an Irish blog site were collected. This selection is made as an attempt to capture a fair representation of social media, where violent lone offenders may submit posts prior to an attack (Kaati et al., 2016, p. 4). In other words, they assembled one collection of data from lone offenders, and three datasets from quite different sources for the sake of showing the broadness of social media, although they mention that more social media data should be included for the experiments to be more realistic. Worth mentioning is the use of Stormfront as a source, commonly described to be the leading white supremacist web forum (Kaati et al., 2016, p. 3), which was introduced in Section 2.1. Being a forum that one could argue would be natural for a lone offender to use, one has to believe that the authors did make sure that the posts were not written by lone offenders, as they describe it as part of the control group.

Although the use of machine learning and automatic methods to identify school shooters has not been extensively studied (Neuman et al., 2020), it appears that more research has been done on identifying lone wolf offenders. Due to the sparsity of information about approaches towards detecting school shooters, we have chosen to widen our scope to include information regarding the detection of lone wolf perpetrators in our literature review as well. Whether or not school shooters and lone wolf attackers have enough in common to use traits common for one for the detection of another, seems to be a point of contention (Leenaars, 2015). The choice to make this comparison and build on it for our future work could be seen as a flaw or a problem that has to be addressed further down the road. However, we choose to look to previous work in the area such as Kaati et al. (2016) which seemed to yield promising results with the use of this generalization. Doing so provides us with two main approaches to screen for school shooters we consider worth examining: machine learning and statistical analysis.

5. Conclusion and Future Work

This chapter serves as a conclusion to the specialization project. The completion of each research question will be judged. As a final section, future work for an eventual Master's Thesis on the topics presented in this study will be proposed.

5.1 Conclusion

The goal of this project is to create an exhaustive overview of the field of using social media to identify users with traits similar to known lone wolf perpetrators, as presented in Section 1.2. To help concretize the direction for the literature review, four research questions were formed. In this section, these will be picked up again for answering whether or not they are fulfilled.

RQ 1: What work has already been done to identify personality traits based on social media presence?

There has already been performed extensive work on personality prediction based on social media presence. A quick glance at the collected studies from the SLR shows that most are indeed based on activity on social media such as Twitter and Facebook. This should therefore serve as a good stepping stone and inspiration for potential future work on a Master s Thesis.

RQ 2: What sources of data documenting previous incidents of lone wolf attacks are already available?

The second research question aimed to direct the preliminary study toward the objective of finding relevant datasets already existing. After the conclusion of the literature review, the authors are left with very few datasets that can be used for the specific task of detecting lone-wolf perpetrators. The dataset deemed most relevant to the task, schoolshooters.info, does indeed contain information on school shooters and their written documents. However, there is little to no data linking school shooters and lone wolves to specific social media accounts allowing for textual analysis of data posted on social media. With that being said, the question reads as a question of whether or not there exists data pertaining to incidents of lone wolf attacks, which there indeed seems to do.

RQ 3: What is proposed as distinguishing traits of a lone wolf perpetrator?

As stated in the discussion, there seem to be no traits that will provide a definite prediction of whether or not an individual is or has the potential to become a lone wolf perpetrator. However, studies have found traits that may coincide with the behavior exhibited by said perpetrators. As presented by Neuman et al. (2015), traits coinciding with vengeful behavior or expression could be linked to people capable of executing a solo attack. Building

further on Neuman et. al's work, there looks to be promising potential in the vectors proposed in their vectorial semantics approach. The work of Kaati et al. (2016) also showed potential. A good starting point for a future investigation seems to be a combination of these two approaches.

RQ 4: What can be done as further work to screen for lone wolf perpetrators using social media?

This last research question is intended as a guideline for what can be done as work for the future Master's Thesis. The answer to this research question is found in the next section - future work.

5.1.1 Research Goal

Our research goal was to provide an exhaustive overview of the problem domain that would serve as a solid base for further work on a Masters's Thesis. To what degree this has been accomplished, should be evident from this concluding chapter. The areas of personality prediction and screening for lone wolf perpetrators and school shooters have been explored thoroughly. However, it is our belief that there is still some work to be done when it comes to studying the psychology of said offenders. There needs to be a solid understanding of what or why these people commit such acts or at least an understanding of what they have in common. If this is not present, any model constructed to screen for such perpetrators is not based on any solid truth. It would rather be a qualified guess at best. This problem is also presented as future work so that this shortcoming may be corrected. Despite this shortcoming, we would like to conclude that the study has managed to produce a solid overview of the areas of research that will be essential for a future Masters's Thesis.

5.2 Future Work

The coming sections will present areas of future work identified for the upcoming Master's Thesis. This involves the construction of a dataset, feature selection, the choice of machine learning models and choice of personality model for capturing the traits of lone wolf perpetrators.

5.2.1 Constructing a Dataset

Seeing as the literature study produced little information on a dataset satisfying the needs specified, a large part of a future project would have to be spent on the collection of relevant data. If one chooses to go ahead with social media data as the main source of textual data produced by lone wolf perpetrators and school shooters, this will be a large hurdle in and of itself. The data collection would have to be performed by identifying social media accounts owned by known lone wolf perpetrators, which additionally would have to be public. There also exists a possibility of the accounts of said perpetrators being removed due to the nature of some of their last posts. This is an issue that will have to be considered in the upcoming Master's Thesis.

5.2.2 Feature Selection and Machine Learning Models

As already discussed in section 3.3, the features used for the prediction of personality seem to vary based on the model and task at hand. Future studies should therefore aim to identify what features work the best for identifying our specific target group. It is also worth exploring the differences in performance linked to the use of closed or open vocabulary approaches. As presented earlier, closed vocabulary approaches could limit accuracy due to the difference in personality shown on social media versus the personality shown in other settings. Therefore, for now, it seems like the most desirable set of features would be a combination of an open vocabulary approach plus user specific features. This does however rely on the availability of these features in the first place.

The features found to be the best suited will also depend on the machine learning models used. The models presented in the literature review had their respective benefits, however, what seems to be the most promising is a move toward deep learning models. The best deep learning model identified was a combination between an LSTM and a CNN. Advances in machine learning have brought forth new models that aim to make up for the LSTM model's shortcomings, specifically the transformer models. They allow for greater retention of context and generally perform better on benchmarks across the board for text-based machine learning problems. These models are also usually accompanied by their own pre-trained tokenizer that requires very little pre-processing. This also enables the use of the desired type of features using an open vocabulary approach. With this in mind, future work should consider the possibility of employing these powerful models for the automatic detection of lone wolf perpetrators and school shooters.

5.2.3 Distinguishing Traits of Lone Wolf Perpetrators and School Shooters

Kaati et al. (2016) states that there is a lack of a common profile for lone offenders. This poses a significant challenge when confronted with the task of identifying them based on textual information. The desired course of action would be to be able to actually identify one such common profile and then use this profile as a guideline for traits to look for when screening social media data. Since this proves to be difficult, one has to look for another solution. One such solution may be to look further into the field of psychology and see what work has been done here to detect lone wolf perpetrators and school shooters. In hindsight, this report did regrettably not go into enough detail about this topic, leaving this as a very important open problem for future work. Potential places to look are psychological evaluations of school shooters or asking for expert opinions. There have also been found some common personality traits of lone wolf perpetrators, however, these are not traits exclusive to them. This is a very important fact that has to be kept in mind when attempting automatic detection. Someone exhibiting a trait linked to people having committed a school shooting is not necessarily a school shooter themself. This is what should be looked into at the very beginning of a future Master's Thesis, for having a basis for any assumptions made to make automatic detection possible.

Bibliography

Arnoux, P.-H., Xu, A., Boyette, N., Mahmud, J., Akkiraju, R., & Sinha, V. (2017). 25 tweets to know you: A new model to predict personality with social media. https://doi.org/10.48550/ARXIV.1704.05513

Bakker, E., & de Graaf, B. (2011). Preventing lone wolf terrorism: Some CT approaches addressed. *Perspectives on Terrorism*, *5*(5/6), 43–50. Retrieved 9th November 2022, from http://www.jstor.org/stable/26298538

Barhamgi, M., Masmoudi, A., Lara-Cabrera, R., & Camacho, D. (2018). Social networks data analysis with semantics: Application to the radicalization problem. *Journal of Ambient Intelligence and Humanized Computing*, 1–15. https://doi.org/10.1007/s12652-018-0968-z

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. https://doi.org/10.48550/ARXIV.1607.04606

Brynielsson, J., Horndahl, A., Johansson, F., Kaati, L., Mårtenson, C., & Svenson, P. (2013). Harvesting and analysis of weak signals for detecting lone wolf terrorists. *Security Informatics*, *2*(1), 1–15. https://doi.org/10.1186/2190-8532-2-11

Costa, P., & McCrae, R. (2008). The revised NEO personality inventory (NEO-PI-R). *The SAGE Handbook of Personality Theory and Assessment*, *2*, 179–198. https://doi.org/10.4135/9781849200479.n9

Cox, J. W., Rich, S., Chiu, A., Thacker, H., Chong, L., Muyskens, J., & Ulmanu, M. (2022). The Washington Post's database of school shootings. Retrieved 2nd December 2022, from https://www.washingtonpost.com/graphics/2018/local/school-shootings-database/

Ekwunife, N. E. (2022). *National security through social media intelligence: Domestic incident prediction* (Doctoral dissertation) [Access was given by Donna Schaeffer, Ph.D., committee chair]. Marymount University. Retrieved 10th November 2022, from https://www.proquest.com/openview/47a13c5fc4a34bc47135c2998cd7d94d/

Elias, M. (2012). Sikh temple killer Wade Michael Page radicalized in army. *Southern Poverty Law Center*. Retrieved 6th December 2022, from https://www.splcenter.org/fighting-hate/intelligence-report/2012/sikh-temple-killer-wade-michael-page-radicalized-army

Farnadi, G., Zoghbi, S., Moens, M.-F., & De Cock, M. (2013). Recognising personality traits using Facebook status updates. Retrieved 7th November 2022, from https://ojs.aaai.org/index.php/ICWSM/article/view/14470/14319

Ganor, B. (2021). Understanding the motivations of "lone wolf" terrorists: The "bathtub" model. *Perspectives on Terrorism*, *15*(2), 23–32. Retrieved 5th December 2022, from https://www.jstor.org/stable/27007294

Gill, P., Horgan, J., & Deckert, P. (2014). Bombing alone: Tracing the motivations and antecedent behaviors of lone-actor terrorists. *Journal of forensic sciences*, *59*(2), 425–435. https://doi.org/10.1111/1556-4029.12312

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, *7*, 7–28.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. Retrieved 5th November 2022, from http://www.deeplearningbook.org

Guldimann, A., & Meloy, J. (2020). Assessing the threat of lone-actor terrorism: The reliability and validity of the TRAP-18. *Forensische Psychiatrie, Psychologie, Kriminologie*, *14*. https://doi.org/10.1007/s11757-020-00596-y

Hamm, M., & Spaaij, R. (2015). Lone wolf terrorism in America: Using knowledge of radicalization pathways to forge prevention strategies. *Washington, DC: US Department of Justice*. Retrieved 28th November 2022, from https://www.ojp.gov/pdffiles1/nij/grants/248691.pdf

Harris, Z. S. (1954). Distributional structure. *WORD*, *10*(2-3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hatewatch. (2017). Waning storm: Stormfront.org loses its domain. *Southern Poverty Law Center*. Retrieved 6th December 2022, from https://www.splcenter.org/hatewatch/2017/08/29/waning-storm-stormfrontorg-loses-its-domain

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support Vector Machines. *IEEE Intelligent Systems and their applications*, *13*(4), 18–28. https://doi.org/10.1109/5254.708428

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. *2*, 102–138. Retrieved 1st December 2022, from https://pages.uoregon.edu/sanjay/pubs/bigfive.pdf

Jones, K. (2004). A statistical interpretation of term specificity in retrieval. *Journal of Documentation*, *60*, 493–502. https://doi.org/10.1108/00220410410560573

Junyi, S. (2020). Jieba Chinese text segmentation. Retrieved 5th December 2022, from https://github.com/fxsjy/jieba

Kaati, L., Shrestha, A., & Sardella, T. (2016). Identifying warning behaviors of violent lone offenders in written communication, 1053–1060. https://doi.org/10.1109/ICDMW.2016.0152

Kaati, L., & Svenson, P. (2011). Analysis of competing hypothesis for investigating lone wolf terrorist, 295–299. https://doi.org/10.1109/EISIC.2011.60

Kofod-Petersen, A. (2018). How to do a structured literature review in computer science. Retrieved 15th September 2022, from https://www.researchgate.net/publication/265158913_How_to_do_a_Structured_Literature_Review_in_computer_science

Langman, P. (2022). School shooters .info: Resources on school shootings, perpetrators, and prevention. Retrieved 14th December 2022, from https://schoolshooters.info

Leenaars, J. (2015). Lone perpetrators. to what extent are school shooters and lone wolf terrorists comparable? Retrieved 4th November 2022, from https://www.leidensecurityandglobalaffairs.nl/articles/lone-perpetrators-to-what-extent-are-school-shooters-and-lone-wolf-terrori

Lima, A. C. E., & de Castro, L. N. (2014). A multi-label, semi-supervised classification approach applied to personality prediction in social media. *Neural Networks*, *58*, 122–130. Retrieved 29th November 2022, from https://www.sciencedirect.com/science/article/pii/S0893608014001282

Mairesse, F., Walker, M., Mehl, M., & Moore, R. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res. (JAIR)*, *30*, 457–500. https://doi.org/10.1613/jair.2349

Markovikj, D., Gievska, S., Kosinski, M., & Stillwell, D. (2013). Mining Facebook data for predictive personality modeling. https://doi.org/10.1609/icwsm.v7i2.14466

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

Meloy, J., & Gill, P. (2016). The lone-actor terrorist and the TRAP-18. *Journal of Threat Assessment and Management*, *3*, 37–52. https://doi.org/10.1037/tam0000061

Meta. (2022). Meta reports third quarter 2022 results. Retrieved 5th December 2022, from https://s21.q4cdn.com/399680738/files/doc_news/Meta-Reports-Third-Quarter-2022-Results-2022.pdf

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. https://doi.org/10.48550/ARXIV.1301.3781

Minsky, M., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry*. MIT Press. https://doi.org/10.7551/mitpress/11301.001.0001

Myers, I. B., McCaulley, M. H., & Most, R. (1985). *Manual, a guide to the development and use of the Myers-Briggs Type Indicator*. Consulting Psychologists Press.

Neuman, Y., Assaf, D., Cohen, Y., & Knoll, J. L. (2015). Profiling school shooters: Automatic text-based analysis. *Frontiers in Psychiatry*, 86. https://doi.org/10.3389/fpsyt.2015.00086

Neuman, Y., & Cohen, Y. (2014). A vectorial semantics approach to personality assessment. https://doi.org/10.1038/srep04761

Neuman, Y., Cohen, Y., & Neuman, Y. (2019). How to (better) find a perpetrator in a haystack. *Journal of Big Data*, *6*(1), 1–17. https://doi.org/10.1186/s40537-019-0172-9

Neuman, Y., Lev-Ran, Y., & Erez, E. S. (2020). Screening for potential school shooters through the Weight of Evidence. *Heliyon*, *6*. https://doi.org/10.1016/j.heliyon.2020.e05066

Palomino, M., & Aider, F. (2022). Evaluating the effectiveness of text pre-processing in sentiment analysis. *Applied Sciences*, *12*, 8765. https://doi.org/10.3390/app12178765

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. P. (2014). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*. https://doi.org/http://dx.doi.org/10.1037/pspp0000020

Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*(6), 556–563. https://doi.org/https://doi.org/10.1016/S0092-6566(02)00505-6

Peng, K.-H., Liou, L.-H., Chang, C.-S., & Lee, D.-S. (2015). Predicting personality traits of Chinese users based on Facebook wall posts. https://doi.org/10.1109/WOCC.2015.7346106

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic Inquiry and Word Count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*. Retrieved 2nd November 2022, from https://www.researchgate.net/publication/239667728_Linguistic_Inquiry_and_Word_Count_LIWC_LIWC2001

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. https://doi.org/10.48550/ARXIV.1802.05365

Plank, B., & Hovy, D. (2015). Personality traits on Twitter—or—How to get 1,500 personality tests in a week, 92–98. https://doi.org/10.18653/v1/W15-2913

Potok, M. (2016). Radicals react with delight to murder of British lawmaker. *Southern Poverty Law Center*. Retrieved 6th December 2022, from https://www.splcenter.org/hatewatch/2016/06/17/radicals-react-delight-murder-british-lawmaker

Potts, C. (2011). Happyfuntokenizing.py. Retrieved 9th December 2022, from http://sentiment.christopherpotts.net/code-data/happyfuntokenizing.py

Power, R. A., & Pluess, M. (2015). Heritability estimates of the Big Five personality traits based on common genetic variants. *Translational psychiatry*, *5*(7), e604–e604. https://doi.org/10.1038/tp.2015.96

Pratama, B. Y., & Sarno, R. (2015). Personality classification based on Twitter text using Naive Bayes, KNN and SVM, 170–174. https://doi.org/10.1109/ICODSE.2015.7436992

Riedman, D. (2022). K-12 school shooting database. Retrieved 10th November 2022, from https://k12ssdb.org/all-shootings

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386. https://doi.org/https://doi.org/10.1037/h0042519

Sanders Analytics. (2013). Twitter corpus. Retrieved 8th December 2022, from https://github.com/zfz/twitter_corpus

Semenov, A., Veijalainen, J., & Kyppo, J. (2010). Analysing the presence of school-shooting related communities at social media sites. *International Journal of Multimedia Intelligence and Security*, *1*(3). Retrieved 29th November 2022, from https://www.inderscienceonline.com/doi/abs/10.1504/IJMIS.2010.03754?journalCode=ijmis

Simons, A., & Meloy, J. R. (2017). Foundations of threat assessment and management, 627–644. Retrieved 10th November 2022, from https://drreidmeloy.com/wp-content/uploads/2017/12/2017_FoundationsOfThreat.pdf

Statista. (2022). Most popular social networks worldwide as of January 2022, ranked by number of monthly active users. Retrieved 5th December 2022, from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Stormfront. (2022). White nationalist community forum. Retrieved 2nd December 2022, from https://www.stormfront.org/forum/

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2018). Personality predictions based on user behavior on the Facebook social media platform. *IEEE Access*, *6*. https://doi.org/10.1109/ACCESS.2018.2876502

Tandera, T., Hendro, Suhartono, D., Wongso, R., & Prasetio, Y. L. (2017). Personality prediction system from Facebook users. *Procedia Computer Science*, *116*, 604–611. https://doi.org/10.1016/j.procs.2017.10.016

Vision of Humanity. (2019). Global Terrorism Index: The rise of the self-radicalised lone wolf terrorist. Retrieved 4th December 2022, from https://www.visionofhumanity.org/increase-in-self-radicalised-lone-wolf-attackers/

Walter, D., Ophir, Y., Lokmanoglu, A. D., & Pruden, M. L. (2022). Vaccine discourse in white nationalist online communication: A mixed-methods computational approach. *Social Science & Medicine*, *298*. https://doi.org/10.1016/j.socscimed.2022.114859

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering, 1–10. https://doi.org/10.1145/2601248.2601268

Zittrain, J., & Edelman, B. (2002). Localized Google search result exclusions [Berkman Center for Internet & Society, Harvard Law School]. Retrieved 1st December 2022, from https://cyber.harvard.edu/filtering/google/

Appendices

A Primary and Secondary Inclusion Criteria

Q1:
Primary Inclusion Criteria

- The study's main concern is predicting/screening people capable* of performing a school shooting.
- The study is a primary study presenting empirical results.

Secondary Inclusion Criteria

- The study focuses on finding people capable of performing a school shooting based on written social media activity

Q2:
Primary Inclusion Criteria

- The study's main concern is the prediction of personality.
- The study is a primary study presenting empirical results.

Secondary Inclusion Criteria

- The study focuses on predicting personality traits based on written social media activity.

B Quality Assessment Criteria

1. Is there a clear statement of the aim of the research?
2. Is the study put into the context of other studies and research?
3. Are system or algorithmic design decisions justified?
4. Is the test data set reproducible?
5. Is the study algorithm reproducible?
6. Is the experimental procedure thoroughly explained and reproducible?
7. Is it clearly stated in the study which other algorithms the study's algorithm(s) have been compared with?
8. Are the performance metrics used in the study explained and justified?
9. Are the test results thoroughly analyzed?
10. Does the test evidence support the findings presented?

C Primary Studies

Id	Title	Author(s)
1	Identifying Warning Behaviors of Violent Lone Offenders in Written Communication	L. Kaati, A. Shrestha & T. Sardella
2	Profiling School Shooters: Automatic Text-Based Analysis	Y. Neuman, D. Assaf, Y. Cohen & J. L. Knoll
3	A Multi-Label, Semi-Supervised Classification Approach Applied to Personality Prediction in Social Media	A. C. E. S. Lima & L. N. Castro
4	Personality Predictions Based on User Behavior on the Facebook Social Media Platform	M. M. Tadesse , H. Lin, B. Xu & L. Yang
5	Personality Classification Based on Twitter Text Using Naive Bayes, KNN and SVM	B. Y. Pratama & R. Sarno
6	Recognising Personality Traits Using Facebook Status Updates	G. Farnadi, S. Zoghbi, M.-F. Moens & M. De Cock
7	Predicting Personality Traits of Chinese Users Based on Facebook Wall Posts	K.-H. Peng, L.-H. Liou, C.-S. Chang & D.-S. Lee
8	Personality Prediction System from Facebook Users	T. Tandera, H. Derwin, S. Rini, W. Yen & L. Prasetio
9	Mining Facebook Data for Predictive Personality Modeling	D. Markovikj, S. Gievska, M. Kosinski & D. Stillwell
10	Automatic Personality Assessment Through Social Media Language	G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar & M. E. P. Seligman
11	Personality Traits on Twitter or How to Get 1,500 Personality Tests in a Week	B. Plank & D. Hovy
12	25 Tweets to Know You: A New Model to Predict Personality With Social Media	P.-H. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju & V. Sinha

Table C.1: List of studies returned from literature search before quality assessment

D Supplementary Papers Extracted During Snowballing

Title	**Author(s)**	**Year**
How to (Better) Find a Perpetrator in a Haystack	Y. Neuman, Y. Cohen & Y. Neuman	2019
Empath: Understanding Topic Signals in Large-Scale Text	E. Fast, B. Chen & M. S. Bernstein	2016
Themes of Revenge: Automatic Identification of Vengeful Content in Textual Data	Y. Neuman, E. S. Erez, J. Tschantret & H. Weiss	2022
Visualizing the Relationship Among Indicators for Lone Actor Terrorist Attacks: Multidimensional Scaling and the TRAP-18	A. Goodwill & J. R. Meloy	2019
TRAP-18 Indicators Validated Through the Forensic Linguistic Analysis of Targeted Violence Manifestos.	J. Kupper & J. R. Meloy	2021
Detecting Linguistic Markers of Violent Extremism in Online Environments	F. Johansson, L. Kaati & M. Sahlgren	2017
Detecting Linguistic Markers for Radical Violence in Social Media	K. Cohen, F. Johansson, L. Kaati & J. C. Mork	2014
Linguistic Analysis of Lone Offender Manifestos	L. Kaati, A. Shrestha & K. Cohen	2016
Assessing Violence Risk in Threatening Communications	K. Glasgow & R. Schouten	2014
The Role of Warning Behaviors in Threat Assessment: An Exploration and Suggested Typology	J. R. Meloy, J. Hoffmann, A. Guldimann & D. James	2012
A Linguistic Analysis of Mass Shooter Journals, Diaries, Correspondence, and Manifestos	H. Duong	2020
Advances in Violent Extremist Risk Analysis	P. Gill, Z. Marchment, S. Zolghadriha, N. Salman, B. Rottweiler, C. Clemmow & I. V. D. Vegt	2020
Analysis of Weak Signals for Detecting Lone Wolf Terrorists	J. Brynielsson, A. Horndahl, F. Johansson, L. Kaati, C. Mårtenson & P. Svenson	2012
Assessment of Risk in Written Communication Introducing the Profile Risk Assessment Tool (PRAT)	N. Akrami, A. Shrestha, M. Berggren, L. Kaati & M. Obaidi	2018

Table D.1: List of 14 supplementary papers after applying the snowballing method