UTRECHT UNIVERSITY

Department of Information and Computing Science

---

**Applied Data Science master thesis**

# Detecting and Classifying Online Radicalisation Phases: A Study of Dutch Social Media Discourse

**First examiner:**
Mirko Schaefer

**Second examiner:**
Dennis Nguyen

**Candidate:**
Dean Newar

**In cooperation with:**
Data School

July 26, 2024

**Abstract**

The increase of social media platforms such as Twitter and Telegram has transformed communication, enhanced free speech, but also facilitated the spread of harmful content. This rise in online threats against Dutch politicians highlights the need to understand online radicalisation processes to protect democratic values. This study combines Data Science and Media Studies to develop a multi-class text classifier aimed at identifying and classifying phases of radicalisation, based on the staircase to terrorism by Moghaddam (2005), in Dutch social media messages. The classifier achieved an accuracy of 98.91%, surpassing previous models. Precision varied, with high scores for non-radicalised messages and death threats, but lower for intermediate phases like dehumanising, demonising, and violent threats. Here we show that while the classifier excels in identifying non-radicalised messages and death threats, it struggles with nuanced phases of radicalisation, showed by lower precision scores for intermediate phases. Our case studies revealed no significant correlation between tried threats and the distribution of radicalisation phases, although peaks in messaging corresponded with public events like demonstrations and COVID-19 press conferences. Future research should explore additional social media platforms and features like retweets and emoticons to better understand the dynamics of online radicalisation.

Understanding these mechanisms is crucial for developing strategies to counteract online threats and safeguard democratic processes. By enhancing classifier accuracy and exploring broader social media data, we can better monitor the spread of radicalising content, ensuring that political discourse remains open and free from intimidation. This research provides a foundation for advancing and exploring tools to detect and analyse online radicalisation and its different phases. To refine these tools and improve their capability to identify nuanced patterns in online discourse, constant efforts have to be made.

# Contents

# 1. Introduction

In recent years, the emergence of social media platforms has transformed how we communicate and exercise free speech, for better and for worse (van Dijck et al., 2018). Social media platforms, such as Twitter [1] and Telegram, which have simplified information sharing, allowing individuals to express their opinions and participate in public discourse, often taking over the role of traditional media. These platforms' use increased in the Netherlands, as seen in Figure 1.1. While this has increased human connectivity, it has also created new ways to spread harmful and threatening content.

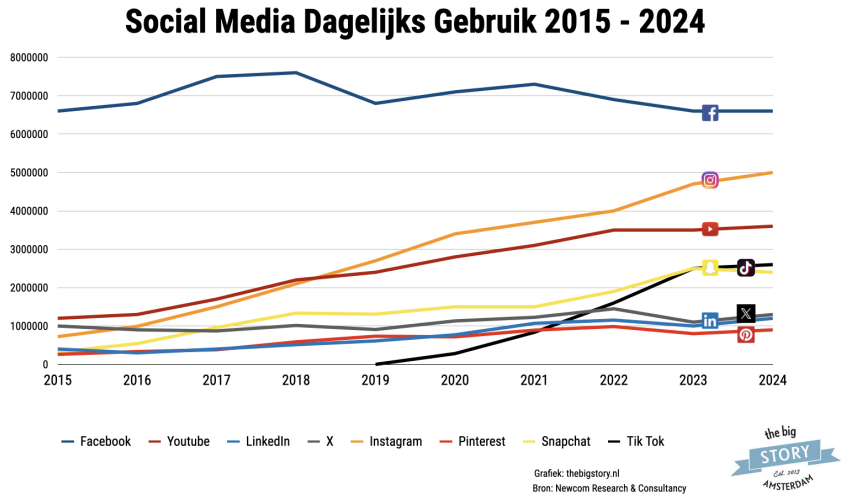


**Figure 1.1:** Social Media Usage between 2015 and 2024 by Vader (2024)

The rise in aggression, presented in the *Monitor Integrity & Safety* report by Kranenburg (2022), shows that the threshold has become lower to share disagreement and frustration towards Dutch politicians. Especially incidents on social media show a significant rise with respect to other com-

[1] Since July 2023, Twitter has been renamed as X. Since the original dataset obtained by DataSchool sourced the data from Twitter. In this research Twitter will be used instead of X

munication options, which can be seen in Figure 1.2. In the Netherlands, the Dutch *National Coordinator for Security and Counterterrorism* (NCTV) has reported a concerning increase in online threats directed at Dutch politicians. According to the *Terrorism Threat Assessment Netherlands 58* report published by the NCTV (2023), the year 2022 saw a record number of instances of threats against politicians. More than a thousand reports were received, almost double the number from previous years. Moreover, the severity of the threats has increased. This resulted in threatened politicians, sometimes keeping their opinions to themselves out of fear of threats, or even discontinuing their political career (Nieuwsradio, 2023). When participation and free political debate can no longer be practised by politicians due to intimidations and threats, it suppresses the values of democracy (NCTV, 2023).

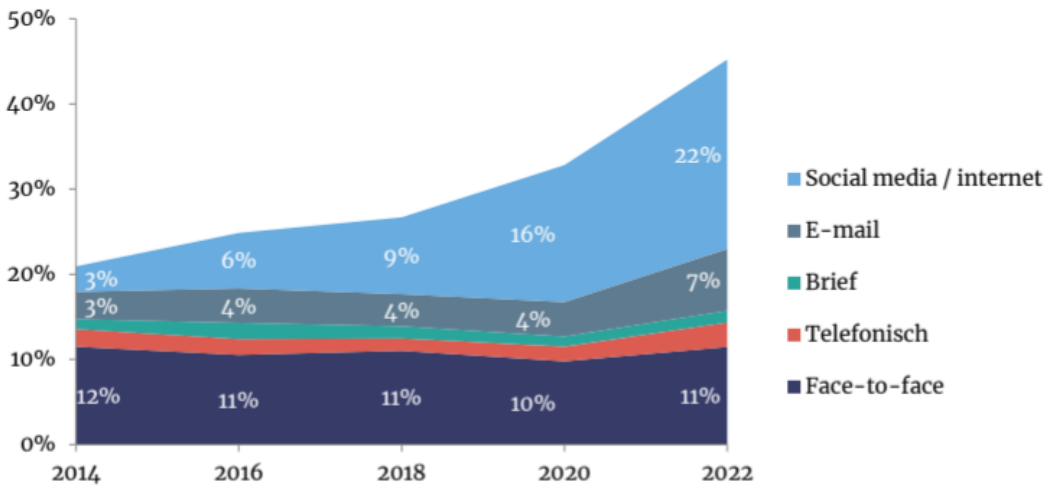


**Figure 1.2:** Distribution of threats across different (Kranenburg, 2022)

A recent example of this occurred in 2023 when a 22-year-old man was arrested for threatening to kill Dutch Prime Minister Rutte. The suspect used a channel on the platform Telegram to post motivations for violence, leading to his arrest and was sentenced by the court. Threats made by the suspect include:

> *"I'm not looking for protesters, I'm looking for revolutionaries. Shooters/hitters/armed/violence. Everything allowed'. And: 'The entire royal family can also eat bullets"* (translated from Dutch to English) [2]

This is one of many incidents that stresses the real and present danger posed by online threats.

## 1.1 Situating the research within Applied Data Science and Media Studies

This research employs an interdisciplinary approach combining Data Science and Media Studies to investigate the complex issue of online threats and radicalisation on social media. Quantitative methods and techniques, offered by data science, are used to analyse significant quantities of textual data extracted from social media. Furthermore, natural language processing (NLP) techniques were used to capture the complexities of textual data for classification. Media studies contribute to the provision of theoretical frameworks for explaining how radicalisation arises and how it appears in online spheres. Furthermore, media studies are researching the phenomenon of misinformation, hate speech and extremism and its distribution on social media (Das et al., 2020; Walther & Mccoy, 2021).

## 1.2 Ethical Considerations

When researching radicalization and online threats on social media, several ethical and societal considerations must be taken into account. This study utilizes the same dataset as *Playing with Fire* (Bakker et al., 2023), ensuring that its use is compliant with GDPR regulations.

By using the dataset from *Playing with Fire* (Bakker et al., 2023), this research integrates the same preconditions that were established in their study. Furthermore, key considerations from NLP papers, as outlined by

[2] https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBDHA:2022:14261)

Benotti and Blackburn (2022), have been carefully incorporated into the methodology.

### 1.2.1 Data Minimisation and Protection

Sensitive data is used in this research including usernames, and other personally identifiable information (PII). To ensure the protection of the PII, the data minimisation principle was used in the collection and storage of the data. For the research objectives, only necessary data was collected and stored locally and in the secure environment Yoda at Utrecht University and SURF Cloud which is in line with Utrecht University's guideline for data collection (Gerritsen, 2021). Since the dataset has been obtained by prior research, a data protection impact assessment (DPIA) has already been conducted. Furthermore, the researchers of the report *Playing with Fire* (Bakker et al., 2023), were advised by a privacy lawyer and reviewed each data point for GDPR compliance. To almost all Twitter accounts, data minimisation was applied with the exception of media accounts, public figures, or large accounts that qualify as such. The storage of the data follows the rules of the Universities of the Netherlands and further research activities were done by following the responsible research code of conduct (van Nederland, 2018).

### 1.2.2 Fairness and Bias

Several studies showed that biases extist in LLM models (e.g. McMilin, 2022). The multi-class text classifier is developed with an effort to reduce biases in the data and the model. To prevent any discriminatory implications, a particular focus is maintained on the data sampling, processing, and classification techniques. Since the model in this research is trained on large datasets, this cannot be ensured.

### 1.2.3 Transparency and Accountability

Throughout the research, transparency is maintained in explaining the data collection, analysis and interpretations clearly. Local copies of the data used in the research will be deleted, ensuring the data will not be kept longer than

necessary and preventing potential data breaches or unauthorised access.

### 1.2.4 Minimising Harm

The ethical challenges posed by potentially assigning harmful labels to messages using a classifier based on the escalation ladder proposed in *Playing with Fire* (Bakker et al., 2023) are acknowledged. By following the data minimisation rules, individual users are only highlighted if they are public figures (Bakker et al., 2023). This research does not aim for the identification of users but explores the possibilities of automatically labelling messages on social media using a classifier. Furthermore, it seeks to find a correlation between the messages and significant peaks in the summed label distribution.

In addition, the false labelling of high-level phases as low-level phases leads to missed opportunities. However, falsely labelling low-level phases as high-level phases can lead to more damaging consequences.

This research aims to explore data science techniques for the classification of messages, not the identification of authors of radicalisation.

### 1.2.5 Compliance with Regulations

This research follows the regulations stated in *Playing with Fire* (Bakker et al., 2023) for responsible data research, collection and storage, and ethical considerations by Benotti and Blackburn (2022). Furthermore, this study ensures GPDR compliance and follows the Code of Conduct of Dutch Universities (van Nederland, 2018)(Universiteiten van Nederland, 2018). In collaboration with the team from "Playing with Fire," these protocols are rigorously followed.

These ethical considerations highlight the commitment to conducting responsible research that protects people's privacy. Furthermore, it promotes equity in data handling and ethically develops an understanding of societal challenges associated with online radicalisation.

## 1.3 Research Problem

The issue of identifying and analysing radicalisation phases (explained in Section 2.1.1) in classified messages on Telegram and Twitter is addressed in this study. The main goal of this research is to develop a multi-class text classifier that identifies and classifies radicalisation phases of social media messages. This classifier will be trained and validated using a dataset of Dutch messages from Twitter and Telegram, sourced by Data School (Discussed in Section 3.1). The main research question that this thesis will answer is:

> *Can a classifier accurately identify and classify the different phases of radicalization?*

Furthermore, the application of the classifier will be researched through a case study by examining the distribution of classified radicalisation phases in a period before threats directed at Dutch politicians. Analysing this distribution can help to understand the progression and escalation of impactful events, which is especially interesting given the increasing number of online threats directed at Dutch Politicians. Therefore, the thesis will answer the question:

> *Is there a correlation between threats against politicians and the distribution and trends of Twitter and Telegram messages?*

## 1.4 Thesis Structure

The answer to these research questions will be obtained by first discussing related works about radicalisation on social media, text classification, and NLP techniques, as well as the classification of radicalisation and the gaps in the literature in Section 2. In Section 3, the methods that will be used for this research will be extensively described and how these methods will be applied. The results of the classifier and the case studies will be presented in Section 4. The research will conclude with a discussion of the findings, including limitations in Section 5 and concluding remarks (Section 6).

# 2. Literature Review

In this section, the literature relevant to radicalisation and its implications on social media will be reviewed. Furthermore, the current computational methods to classify textual data and method gaps are discussed. The aim is to discuss the origins of radicalisation and its manifestation within social media platforms, using computational methods to analyse these phenomena. Due to time constraints, the review will focus primarily on two main areas: the definition and context of radicalisation, and the role of social media platforms in political discourse. Some aspects, such as an exhaustive review of all social media platforms or a detailed examination of non-computational methods, will not be covered in depth.

## 2.1 Defining Radicalisation

The definition of radicalisation by McCurdy (2021): 'Radicalization is the transition into acceptance and approval of extremist beliefs and actions, including condoning or committing acts of violence', is one of the many definitions of the term. The complex and multifaceted nature of radicalisation is a concept that varies across academic disciplines and contexts. By examining the differences and similarities between different contexts, a concise conceptualisation of radicalisation can be found. Radicalisation can manifest in political, social, and religious contexts, each with distinct characteristics but also sharing commonalities. In the political context, as discussed by Muxel (2020), radicalisation is defined as the process by which individuals or groups adopt increasingly extreme political ideas, beliefs, feelings, and behaviours. These often reject democratic norms, deny political institutions, and support conflict and violence between groups. An example is anti-institutional extremism, which, according to the AIVD (2023), involves incidents targeting democratic institutions and processes, including govern-

ments, police, and media, through (non)violent acts that undermine the rule of law. From a social perspective, radicalisation is considered a social process involving a shift in the identity of individuals or groups towards more extreme actions and views, influenced by their social environment (Bott et al., 2006). This collective phenomenon is facilitated by networks and interactions in social spheres, spreading radicalising ideas (Wadhwa & Bhatia, 2015). In the religious context, radicalisation is linked to the adoption of extreme religious ideologies, where individuals embrace beliefs that justify violence towards perceived enemies (de Graaf & van den Bos, 2021). Perliger and Pedahzur (2016) note that such individuals become deeply involved in belief systems that promote violence. Despite differences in definitions across contexts, there are similarities, such as the view of radicalisation as a dynamic process (Borum, 2011) where shifts in ideology and behaviour lead to more radical beliefs and actions. In all contexts, social networks significantly influence the radicalisation process. However, the motivations and objectives in political and social radicalisation aim to transform society to align with their ideologies, while religious radicalisation focuses on theological goals and targets groups with differing beliefs.

In recent years, radicalisation obtained a new context where it can exist, online (media) platforms. Due to social, cognitive, and economic influences, online radicalisation has evolved as a primary platform for forming political ideologies, organising acts of violence, and promoting social movements (Erdogan, 2023). In the report by Nabilah Risky (2023), the author mentioned that the usage of social media facilitates the spread of radicalising ideas. A clear correlation exists between radical conversations on social media platforms, like Twitter, Telegram, Facebook, and YouTube, the distribution of messages in and between social media platforms, and the transfer of this sentiment to actions in the offline public sphere (Bakker et al., 2021).

### 2.1.1 Escalation Ladder

Multiple models for radicalisation processes have been developed, as well as the staircase to terrorism, developed by Moghaddam (2005). It was cre-

ated to help better understand and possibly intervene in the radicalisation process. This can be done by outlining the progressing phases that individuals may go through before switching to extreme acts of violence. In the report *Playing with Fire* (Bakker et al., 2023), the authors name the staircase to terrorism model the escalation ladder obtained from the paper by Moghaddam with an online context in the Dutch political sphere. For the rest of this paper, the term escalation ladder will be used to address the staircase to terrorism model. According to the paper, the process of (online) radicalisation following the escalation ladder occurs in five steps.

1. Dehumanisation: *"A person is stripped of their human dignity and reduced to one aspect. E.g.: rat, puppet, Nazi, dog, clone, slave, sheep".*
2. Demonisation: *"Similar to dehumanisation, but a more urgent sense of danger is expressed. E.g.: satanist, demon, witch, devil".*
3. (Violent) Threat: *"Direct or indirect threat of violence towards a person or group of persons. E.g.: we're coming to get you, you're going to be punished, I'll kick your ass".*
4. Death threat: *"Direct or indirect threat of death towards a person or group of persons. E.g.: hang that business, you get the bullet, I hope someone kills you".*
5. Kinetic action: *The last escalation step towards the execution of threats.*

This escalation ladder will be the foundation of the labelling of data, where its usage in this research will be explained in Section 3.1

## 2.2 Role of Social Media Platforms in Political Discourse

Due to their extensive usage and significance, social media platforms such as Telegram and Twitter have become popular elements of modern political discourse (Nguyen et al., 2022). In this section, we explore the role of Twitter and Telegram in political discourse and how these platforms contribute to the process of radicalisation.

Twitter's real-time updates and broad public have made it a vital platform for global political communication. It is a platform that politicians, activists and regular users use to express their opinions, gather support, and participate in public discussions. This open and accessible nature of Twitter enables all users to directly interact with each other (Wieringa et al., 2018). According to the paper by Agarwal and Sureka (2015), the platform is a place which is exploited for radicalisation which violates the community guidelines of Twitter.

Telegram distinguishes itself through its focus on security and privacy. Discrete communication among political activists and groups is one major reason for users to use the platform. Without the same moderation as other platforms, channels and groups on the platform allow users to organise activities and spread information (Simon et al., 2022). This information can include the ideologies of radicalising/radicalised users.

Several factors ease the spread of extreme content and ideologies on these platforms, as well as other platforms like YouTube and Facebook, which are presented in the paper by Berjawi et al. (2023). One factor is the algorithm-based ground for these platforms, which analyses the user's interests, behaviours, and interactions and uses them for personalised content. This way filter bubbles, and echo chambers arise which could create the idea of a tunnel vision towards extremist content. A second factor is the possibility of only communicating with users who share the same ideologies, possibly leading to radicalisation.

According to the report *Playing with Fire* (Bakker et al., 2023), there is a difference in the severity of radicalisation between the two platforms. It stated that Twitter is a less radicalised platform than Telegram, which can be seen from the "strong presence of conspiracy theories, anti-institutional thinking, and language that is dehumanising, demonising and threatening towards individuals perceived as opponents" (Bakker et al., 2023).

## 2.3 Classification and Imbalanced Data Handling Techniques in Data Science

Extensive research has been done on the classification of text in various contexts. In this section, multiple machine learning, deep learning and NLP techniques are being discussed for the use of (multi-class) text classification. Furthermore, imbalanced data is a continuing problem in the realm of data science, which has resulted in various techniques to handle it. These techniques, related to textual data, are discussed. Lastly, an overview of studies that researched radicalisation classification using NLP techniques is presented.

### 2.3.1 Machine and Deep Learning approaches in Multi-class text classification

In the paper by Parmar et al. (2018), the authors compared five different machine learning techniques for multiclass text classification. Multinomial Naïve Bayes (MBN), Support Vector Machine (SVM), Decision Tree, Random Forest, and K-Nearest Neighbours (KNN), were trained for the classification of text messages from customer support into 12 predefined categories. These categories represented technical system defects. The data was split into a train and test set, and pre-processed by handling missing values, encoding categorical variables, data scaling and converting text to numerical vectors using a TF-IDF Vectorizer. After training, the model was tested on the test set which presented the following results: the SVM classifier achieved the highest accuracy (63.02%) outperforming the other algorithms, while the lowest accuracy was obtained by the Decision Tree model (48.97%). The authors conclude that the SVM model is the most suitable model for classifying multiclass customer support messages.

A comparative study, by Kamath et al. (2018), was conducted to research the difference between traditional machine learning and deep learning approaches for multi-class text classification. The authors used two image datasets: a health insurance dataset (containing images of invoices received

by the insurance company) and the publicly available Tobacco-3482 dataset (containing images related to tobacco from the media). Both datasets contained multiple classes, 18 and 9, respectively. After converting the images to text using the 'tesseract' module (text character recognition in image software with the support of a language model), the data was split into train and test sets. The data is processed by punctuation removal, stop word removal, and stemming. The authors employed four traditional machine learning (Logistic Regression, SVM, Naïve Bayes, and Random Forest) and two deep learning models (Multi-Layer Perceptron (MLP) and a Convolutional Neural Network (CNN)). The models were trained on the datasets and the results showed that the deep learning approach CNN outperformed the other models. On the health and tobacco datasets, it achieved an accuracy score of 96% and 89.27%, respectively.

In addition to the results of Kamath et al. (2018), several other studies present more promising results for the use of deep learning techniques over machine learning techniques.

The authors of the paper by Hasib et al. (2023), proposed an MCNN-LSTM (Multi-class CNN – LSTM) model for multi-class text classification on imbalanced news data from HuffPost (containing six years of news headlines and a short description). A combination of CNN and Long Short-term Memory models was developed to improve accuracy while identifying minority classes among 12 classes in the data. The CNN part of the proposed model is used to extract local features from the textual data, while the LSTM captures the long-term dependencies. A comparison between the proposed model and traditional machine learning techniques like logistic regression, Naïve Bayes, decision trees and SVM. The MCNN-LSTM model outperformed the other models by achieving an F1-score and accuracy of 98% and 99.71%, respectively. However, the authors discuss one limitation of the research as not applying transfer learning.

### 2.3.2 Transfer Learning and Transformers in Multi-class text classification

According to Dhyani (2021), transfer learning involves optimising pre-trained models for a specific use case in NLP. The models have already been trained for certain target tasks on large datasets, which means that a smaller amount of labelled data can be used to modify the weights of the model during optimisation. This makes sure that the model learns the task-specific features and modifies its settings to the specific problem that is researched. These settings refer to the representations which represent how textual data is transformed into numerical data so that it can be processed by models.

Revolutionising NLP, transformer models use a self-attention mechanism to capture the local and global contexts from text data (further explained in Section 3.2.1). As proposed by Vaswani et al. (2017), transformers create the baseline for various benchmarking NLP models, such as BERT, GPT, and RoBERTa (Dhyani, 2021).

In the paper by Devlin et al. (2019), the authors introduce the BERT model, an acronym for Bidirectional Encoder Representations from Transformers. The employment of unsupervised bidirectional training allows the model to grasp the meaning of context in both directions. This extends the original unidirectional training used in Transformers, as well as pre-training the model on Masked Language Modelling and Next Sentence Prediction. The pre-trained BERT model can be used in a broad range of NLP tasks after optimisation with only one output layer added.

The last discussed pre-trained transformer model by Liu et al. (2019) is RoBERTa (Robustly optimised BERT approach). This model is an improved differentiation of the BERT model by differences in the pre-training process, such as "increased batch sises, more training data, and a new training target" (Dhyani, 2021). Due to this improvement, the RoBERTa model showed more promising results than the original BERT model on various NLP benchmarks. A study by Wang and Banko (2021), the authors researched the possibilities of this BERT family in a classification task on the

Cross Lingual Sentiment dataset and the Hateval (hate detection) dataset. The XLM-Roberta model outperformed the BERT and RoBERTa model in the Hate detection task with an F1 score of 72.6. The XLM-RoBERTa model did not outperform the other models on the Cross-Lingual Sentiment classification task where it obtained an F1-score of 93.3 and the other models obtained a score of 93.5.

### 2.3.3 Handling Imbalanced Text Data in NLP

Different techniques have been developed in balancing imbalanced data for the improvement of NLP models. Several approaches are discussed, including oversampling using SMOTE, and paraphrasing.

In the paper by Mujahid et al. (2024), the authors studied the comparative results of employing different SMOTE methods techniques on two highly imbalanced Twitter datasets. These methods were compared after training six different machine learning models. SMOTE (Synthetic Minority Oversampling Technique) is a common method to use for handling imbalanced data and has multiple differentiations including SVM-SMOTE, K-Means SMOTE, ADASYN, and Border-Line SMOTE, which are used in the research. The results show that the highest accuracy was achieved by a SVM model in combination with the ADASYN method (99.67%). While the authors are highly promoting the use of SMOTE on imbalanced text data, the paper by Glazkova (2008) mentions the caution that SMOTE (and derivatives) are hardly applicable to textual data. This can be deducted from the fact that text is high-dimensional and sparse, and numerical data does not represent contextual relations in text.

The handling of imbalanced data by using paraphrasing has not been extensively researched. However, paraphrasing using large language models (LLMs) like GPT-2, have proven to be able to achieve results of high quality in various NLP tasks (Witteveen & Andrews, 2019). In the paper by Witteveen and Andrews (2019), the mentioned GPT-2 model was optimised on various paraphrase datasets which includes sentence- and paragraph level examples. The optimisation of the model resulted in the ability to generate

paraphrased sentences that have similar semantics but a unique word order and phrasing in comparison to the original text.

### 2.3.4 Specific Studies for Radicalisation Detection on Social Media

Several studies have been conducted on the detection and classification of radicalisation on social media. However, these studies prominently research binary instead of multi-class classification of messages. This gap in the method will be discussed in Section 2.4.

In the paper by Agarwal and Sureka (2015), the authors propose a one-class SVM and KNN classifier for the classification of hate-promoting tweets. The study uses two publicly available Twitter datasets and combines them into one dataset for the training of the classifiers. According to the authors, hashtags were important for identifying hate-promoting tweets and the list of hashtags grew while manually labeling the tweets. The results show that the SVM classifier (97% accuracy) outperformed the KNN classifier (90% accuracy), with the most important tags for classifying tweets being the presence of religion, war-related terms, bad words and negative emotions.

To explore the capabilities of deep learning models to classify extremism in text, Rajendran et al. (2022) conducted research regarding the detection of extremism on Twitter during the U.S. Capital Riot on 6 January 2021. The authors collected tweets and developed an extremism dataset which is used for "classifying extremism texts as propaganda, recruitment, radicalisation, and non-extremism". Several deep learning models were employed including Bi-LSTM, BERT, RoBERTa, and Distill-BERT. From the models, the RoBERTa model outperformed the other models with 95% accuracy in classifying the texts.

## 2.4 Gaps in the Current Methods

Several gaps in the current methods are discussed, including the under-representation of the radicalisation staircase (Section 2.1.1) for quantitative

research, especially the task of identifying and classifying these different phases. Furthermore, the lacking focus on quantitative research of the online Dutch political sphere, which often only considers Twitter data. Lastly, the efforts for using XLM-RoBERTa-large in Dutch multi-class text classification are lacking in current research.

The radicalisation staircase, proposed by Moghaddam (2005), has been extensively researched and compared in several papers (Lygre et al., 2011; Stafford et al., 2019). However, these papers only focused on qualitative methods while limited to no research was done to study the quantitative properties of this staircase on online political data.

This lack of focus of online Dutch political data is discussed in the paper by Rajendran et al. (2022). The authors mention that English messages on social media are extensively researched and that further contributions to the field would be to include other languages. Furthermore, the paper mentioned the research on social media platforms other than Twitter.

The classification of radicalisation, whether binary or multi-class, within the Dutch political sphere, remains a clear gap in the literature. To date, no substantial attempts have been made to utilise new NLP techniques, such as XLM-RoBERTa, for classifying messages within this context.

Based on these identified gaps, this research aims to fill the described gaps by quantitatively analysing the radicalisation phases using messages from online Dutch political discourse. Furthermore, expanding the scope of research to include Telegram messages and implementing a relatively new model (XLM-RoBERTa-large) to train a multi-class text classifier for messages in the Dutch political discourse.

# 3. Methods

In this section, we discuss the methodologies that are used to develop and implement the classifier for detecting the distinct phases of radicalisation. The process includes data collection, preprocessing, handling class imbalance with paraphrasing, classifier development, classification process, and validation and testing.

## 3.1 Data

The dataset utilised in this research focuses on the classification of different radicalisation phases within Dutch online discourse targeting Dutch political actors. Provided by the Data School, the dataset was sourced as part of a research project investigating the dynamics between debates in the Dutch House of Representatives and radicalisation in the online sphere. The data consists of Dutch Twitter and Telegram messages spanning from January 1, 2021, to October 1, 2022. According to the report *Playing With Fire* (Bakker et al., 2023), these platforms were selected due to their accessibility via APIs during the data collection. Furthermore, Twitter is used by a large number of Dutch (speaking) users including Members of the House of Representatives, journalists, scientists, and politically involved citizens. The report presents that Telegram often holds more anti-institutional users and radicalized critics than Twitter (Bakker et al., 2023).

### 3.1.1 Twitter Data

A total of 16,241,779 unique messages (excluding reposts) were collected. According to the *Playing With Fire* (Bakker et al., 2023), Twitter messages tend to contain more subjective content (opinions, viewpoints, etc.) than Telegram messages, which contributes to more diverse information. However, at the time of the data collection, messages were manually and au-

tomatically moderated, which resulted in the quick deletion of threatening content (Bakker et al., 2023). The collection of the data was performed by inserting queries to only include Dutch messages that contained the (user)name (full name or last name) of Members of the House of Representatives or the (user)name of national political parties. This was done to minimise the number of unrelated messages in the data. The report by DataSchool states that by performing these queries, some messages will not be found that should be in the data, such as messages that are indirectly related to the House of Representatives.

### 3.1.2 Telegram Data

The Telegram part of the data consists of 10,179,216 messages, where approximately half of the messages are considered unique. Compared to Twitter, Telegram has a higher volume of messages, however, due to the chat service structure their level of informational content varies. Numerous Telegram messages are part of continuous conversations, which is why they often contain short, contextually limited sentences like: *"Is goed"*, *"Hoezo"*, or *"Haha inderdaad"*. According to the report by Data School, Telegram does not have a function to insert queries. Therefore, the data is collected by the "snowball-method" (Peeters & Willaert, 2022), where links in publicly available Telegram channels are linked to other channels.

A subset of this extensive dataset was manually labelled, with the support of scientists with domain knowledge (Bakker et al., 2023), to classify messages into five distinct phases of radicalisation: nothing, dehumanising, demonising, violent threat, and death threat.

The labelled dataset, consisting of 10,830 instances ranging from June 4, 2021, to June 16, 2022, showed a significant class imbalance, as shown in Table 3.1.

| Label | Count |
|---|---|
| 0 (Niets) | 10153 |
| 1 (Dehumaniserend) | 360 |
| 2 (Demoniserend) | 176 |
| 3 ((Gewelds)bedreiging) | 100 |
| 4 (Doodsbedreiging) | 42 |

**Table 3.1:** Distribution of counts per label

### 3.1.3 Preprocessing data

The data was pre-processed to ensure the quality and relevance of the data used for classification. The preprocessing steps involve de-identification, cleaning the data, preparing it for synthetic data generation and model training.

The dataset consists of various types of messages, some of which mention usernames of politicians and normal users. To focus on messages that are relevant to Dutch politicians and to ensure the anonymity of users who are not politicians. The usernames belonging to the latter group were removed. A list of usernames referencing members of the Dutch House of Representatives (Tweede Kamer) that were active during the data collection period (Wikipedia contributors, n.d.), was used to filter out non-political user tags.

Before the process of data synthesis or model training can be started, it is necessary to clean the textual data. The messages were cleaned by removing hyperlinks, special characters, numbers, and extra white spaces. These steps were essential to reduce noise in the textual data. Double quotation marks were reduced to single quotation marks and the text was transformed to lowercase.

### 3.1.4 Imbalance Data Handling using Paraphrasing

To correctly handle the significant class imbalance which can be observed in the labelled dataset (Table 3.1), the AI21 Paraphrase API was employed to paraphrase Dutch text. The paraphrasing of text will be used on the split training set for training the model, this will be discussed in Section 3.3.

The API uses one of the AI21 Studio task-specific models, which are optimised to perform specific tasks with high efficiency and accuracy, in this case paraphrasing. The model is based on AI21's Jurassic-2 model [1], which is a Generative Pretrained Transformer (GPT) autoregressive language model with over 60 billion parameters. It is trained on a dataset that is composed of text posted or uploaded on the internet from sources including CommonCrawl, Wikipedia BookCorpus, arXiv, and Stack Exchange. Jurassic 2 was trained on approximately 1.2 trillion tokens. This model acts similar to GPT-3, however it is optimised and fine-tuned for specific tasks like paraphrasing. Due to this optimisation, it allows Paraphrase API to provide multiple phrasing options, which ensures that the meaning of the original text is preserved while offering varied expressions. ("Paraphrase", 2024).

The Paraphrase API takes a sentence and uses a request to return a list of paraphrases that convey the same meaning using different words. According to "Paraphrase" (2024), the model can handle various languages including Spanish, French, German, Portuguese, Italian and Dutch. However, when prompting Dutch sentences the model output Dutch and English mixed. For example, when prompting the sentence:

> *'e dag van vandaag en enkele ruzies later en mn zoon vorig jaar half jaar niet gezien het is in en in triest en misdadig wat die ratten hebben gedaan in den haag een kogel is nog een te simpele dood publiekelijk ophangen'*

The output of the list of paraphrases consisted of sentences like:

> *'In the day of vandaag, a few ruzies later, and my son half a year ago, I never noticed how in the world. I still think that what has been done in den haag a kogel is no bit more than a simple dood publically ophanged.'*

And also,

> *'The day of today, a few ruisses later, and my son half a year ago are not recognizing it is in dread or in a cloud of misdain.'*

[1] The Jurassic-2 model is currently changed to JAMBA

Therefore, the Dutch text was first translated to English using the Google-Translator from the deep-translator module, this text was paraphrased and translated back to Dutch. Prompting the same sentence as previously stated resulted in the following paraphrased sentences:

> *'Het is werkelijk tragisch en misdadig wat die ratten in Den Haag hebben gedaan. Een kogel is nog steeds een te simpele dood. Hang op in het openbaar.'*

And also:

> *'Vandaag overdag en daarna wat ruzies. Vorig jaar heb ik mijn zoon zes maanden niet gezien. Het is werkelijk treurig en misdadig wat die ratten in Den Haag hebben gedaan. Het is nog steeds een te simpele dood om door een kogel te worden veroorzaakt.'*

This proved to be the most effective solution, unlike other methods that are explained below. The goal of the process was to augment messages of minority classes by generating synthetic data while preserving the semantic and contextual meaning of the original messages.

In addition to paraphrasing, other imbalance techniques were explored, including Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002). While SMOTE is widely used for handling class imbalance by synthetically generating samples from the minority classes, it has its limitations when dealing with textual data. Since SMOTE generates samples by interpolating between minority samples in the feature space, for textual data this often results in new samples that do not have the meaning of natural language sentences. Since natural language is highly complex, it is difficult to capture through interpolation. The classification model utilises the RobertaForSequenceClassification model from HuggingFace. The model is a transformer-based model that understands and processes textual data by leveraging contextual embeddings, which is why SMOTE would not work optimally (resulting in lower scores) while using this model. The use of a random sampler would not work in the sampling process. Since the model must be trained on a diverse set of sentences, replicates would not be sufficient enough.

## 3.2 Classifier Development

For the identification and classification of distinct radicalisation phases of messages from Twitter and Telegram, a transformer-based model named XLM-RoBERTa-large is employed. The model, proposed by Conneau et al. (2019), is known for its excelling performance in multilingual Natural Language Processing (NLP) benchmarks, outperforming other models like Bangla BERT Base, Multilingual BERT (Mukherjee et al., 2023)(Mukherjee et.al, 2023) and the Dutch model RobBERT (de Vries et al., 2023).

Transformers, a special type of neural network that is based on encoder-decoder models proposed by Cho et al. (2014), have completely changed the field of natural language processing. The method was initially proposed by Vaswani et al. (2017) and is currently the basis of many complex language models. The following sections will briefly go over encoder-decoder models, how transformers work and the differentiation of the XLM-RoBERTa model.

### 3.2.1 Encoder-decoder Models

Encoder-decoders, often used in Seq2Seq models (Sutskever et al., 2014) that are employed by tasks such as machine translation and text summarisation, primarily consist of two components: the encoder and decoder. The first component is responsible for understanding and extracting relevant information from the input sequence, often using a Recurrent Neural Network (RNN). In this research, the input sequence is a sentence or multiple sentences. The encoder transforms the processed input into a fixed-length continuous representation (context vector) that is passed to the second component. This component, the decoder (often an RNN), takes the context vector and generates the output sequence. For example, in machine translation and text summarisation, the encoder detects the context from the source language and outputs the input for the decoder. The decoder performs translation into the wished language or summarises text.

Transformer models are adapted encoder-decoder models, where they

differentiate by the introduction of the self-attention mechanism. Unlike encoder-decoders, transformers do not handle sequences sequentially but do process them parallel, which makes them faster than RNNs. By simultaneously processing entire sequences, transformers can model long-range dependencies and capture context more effectively. This is being handled by the self-attention mechanism. For each word in a sequence, the relative importance of other words is being calculated which allows the model to have more attention on distinct parts of the input sequence when generating the output sequence (Vaswani et al., 2017).

### 3.2.2 XLM-RoBERTa Model

An adaptation of the transformer model to enhance the multilingual performance is known as the XLM-RoBERTa-large model. The model is part of the BERT (Devlin et al., 2019) family developed by Huggingface, which is pre-trained on a large scale of multilingual data with 270 milllion parameters. This resulted in excelling achievements performing multilingual NLP tasks based on the DUtch Model Benchmark (DUMB) by de Vries et al. (2023), where it outperformed prominent models like RobBERT and other multilingual models. This is the primary reason why XLM-RoBERTa-large is chosen over any other model.

### 3.2.3 Application of Classifier

The RobertaForSequenceClassification framework is being used for the classification process. The framework adopts the capabilities of the XLM-RoBERTa-large model for the classification of radicalisation phases.

## 3.3 Classification Process

The classification process consists of several steps that are necessary for obtaining the final radicalisation phase classifier. These steps include loading and splitting the data, paraphrasing, weight computation, tokenisation, training of the model, and hyperparameter tuning.

### 3.3.1 Loading and Splitting Data

The first step of the process is loading the data using the pandas module (McKinney, 2010), which contains the original labelled data. This data is split into three different sets for training, validation, and testing with the following distribution: 60%, 20%, and 20%, respectively. The validation set is used to tune the hyperparameters and keep track of the performance of the model.

### 3.3.2 Paraphrasing Imbalanced Data

The imbalance of the class distribution across the dataset, as mentioned in Section 3.1, can also be observed in the training set as shown in Table 3.2.

| **Label** | **Count** |
|---|---|
| 0 (Niets) | 6093 |
| 1 (Dehumaniserend) | 213 |
| 2 (Demoniserend) | 112 |
| 3 ((Gewelds)bedreiging) | 58 |
| 4 (Doodsbedreiging) | 22 |

**Table 3.2:** Distribution of counts per label in the training set.

This process of paraphrasing the training set resulted in a more balanced class distribution of messages across the different labels. However, due to the boundaries of language, it is not possible to generate infinite paraphrases of a single sentence, therefore achieving the perfect balance in labels remains a challenge. The method of oversampling the minority classes is only employed in the training set. This is because the validation and test set should hold their original imbalanced distribution to provide an accurate representation of the real data. An updated distribution across the labels of the concatenated dataset (paraphrased data and original training data) can be seen in Table 3.3. The number of sentences that were paraphrased was chosen liberally, however the distribution of the original training data was kept. For example, label 4 has the least number of messages, which also appears in the original training data.

Despite the improvement of the class distribution after oversampling

| Label | Count |
|---|---|
| 0 (Niets) | 6093 |
| 1 (Dehumaniserend) | 426 |
| 2 (Demoniserend) | 335 |
| 3 ((Gewelds)bedreiging) | 290 |
| 4 (Doodsbedreiging) | 220 |

**Table 3.3:** Distribution of counts per label in the training set after paraphrasing.

the minority classes, down sampling of the majority class (label 0 (Nothing)) was performed to further reduce the imbalance of the training set. The number of instances with label 0 was reduced to 4000, which resulted in the distribution presented in Table 3.4.

| Label | Count |
|---|---|
| 0 (Niets) | 4000 |
| 1 (Dehumaniserend) | 426 |
| 2 (Demoniserend) | 335 |
| 3 ((Gewelds)bedreiging) | 290 |
| 4 (Doodsbedreiging) | 220 |

**Table 3.4:** Distribution of counts per label in the training set after paraphrasing and under sampling the majority label.

### 3.3.3 Dataset Dictionary

A new dataset dictionary for the training, validation, and test sets is created using the DatasetDict function of HuggingFace. This construction is useful for accessing the different sets used in training and evaluation.

### 3.3.4 Weight computation

Despite performing paraphrasing on the training data, the set is still imbalanced. Therefore, the class weights are being computed on the training set with the *compute_class_weight* function from sci-kit learn. The variable *class_weight* allows for the choice of "balanced" class weights, which is computed using the following formula:

$$classweights_{balanced} = \frac{N_{samples}}{N_{labels} \cdot bincount(labels)}$$

where:

- $N_{samples}$ = number of samples,
- $N_{labels}$ = number of labels,
- $bincount(labels)$ = array of classes occurring in the data.

The function assigns higher weights to classes which represent less samples. This guarantees that the model pays more attention to the minority classes. The distribution of the class weights is presented in Table 3.5.

| **Label** | **Weights** |
|---|---|
| 0 (Niets) | 0.26355 |
| 1 (Dehumaniserend) | 2.47464789 |
| 2 (Demoniserend) | 3.14686567 |
| 3 ((Gewelds)bedreiging) | 3.63517241 |
| 4 (Doodsbedreiging) | 4.79181818 |

**Table 3.5:** Weight distribution across labels.

### 3.3.5 Tokenisation

An additional step in preprocessing has to be made before the data can be fed to the model. This step involves the tokenisation of the messages using the AutoTokenizer from HuggingFace, which initiates a tokenizer for the XLM-RoBERTa-large model. It converts the raw messages to a sequence of tokens that reprents words or characters, so that the model can understand. The tokenized dataset is preprocessed to remove hyperlinks, double empty quotes and replaces usernames with “@user” to follow the anonymisation explained in Section 1.2.1, which helps in improving the model’s performance.

### 3.3.6 Training of the model

The training process of the RobertaForSequenceClassification with the XLM-RoBERTa-large uses the Trainer function developed by HuggingFace. By using this function, the training process simplifies by using data loading, training loop, and evaluation functions.

### 3.3.7 Hyperparameter Tuning with Optuna

The process of evaluating different combinations of tuneable parameters of a model, hyperparameter tuning, searches for the best combination of parameters with the best classification performance. Optuna is an automatic hyperparameter optimiser developed by Huggingface, and is used for the hyperparameter tuning process. In this process, several parameters are researched:

The first hyperparameter is the learning rate which is used for controlling the size of each step that the optimiser takes to find the minimum of the loss function. This process of choosing a learning rate differs per objective. Therefore, no theoretical source can be found that provides a learning rate specific for this objective. The second parameter that is tuned by Optuna is weight decay, which is a regularisation technique commonly known as L2 regularisation. It penalises larger weights by adding a term to the loss function which makes sure the model does not assign more importance to a single feature than necessary. When the weight decay is too low, the model might overfit the data since it almost does not penalise the weights. However, when the value of the weight decay is set too high, the model would receive too many penalties and would underfit. Lastly, the optimiser is tuned by Optuna, extending the hyperparameter space while exploring the AdamW and SGD optimisers.

Some hyperparameters were not researched in the hyperparameter space, including batch size. This hyperparameter defines the number of training examples that are being used in one iteration of training the model. Due to the fact that the dataset is heavily imbalanced, the batch size cannot be too low since that would present an incorrect representation since every

label must be present in the batch to make the model learn. However, during training the limited computational power and time resulted in an established batch size of 8. A second hyperparameter that is not tuned by Optuna, is the number of epochs (one epoch is one pass through the entire dataset). The number of epochs was set to 10 due to the time constraints. Furthermore, while training the model, the loss per epoch did not decrease significantly after 10 epochs.

## 3.4 Evaluation and Metrics

The traditional metrics of a multiclass classification task with imbalanced data including accuracy, recall, precision, and f1 score, are discussed below. The metric precision is carefully chosen to evaluate the classifier regarding the minimisation of false positives.

### 3.4.1 Accuracy

$$accuracy = \frac{correct predictions}{all predictions}$$

Accuracy calculates the number of correctly classified messages out of all the messages. This metric is not suitable for the multi-class classification task with imbalanced data, since it disregards the class balance. A high score is obtained by often predicting the majority class while ignoring the minority classes.

### 3.4.2 Recall

$$recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Recall measures the number of correctly classified positive instances out of all positive instances. It is not a suitable metric for this problem since a high

recall score may lead to the misclassification of messages belonging to the majority class (0) as classifying them to a minority class (1, 2, 3, or 4). In the context of this research, as discussed in Section 1.2.1, the objective of the metric must be to minimise the misclassification of the majority class.

### 3.4.3 Precision

$$recall = \frac{TruePositives}{TruePositives + FalsePositives}$$

Precision measures the number of instances predicted as positive that are indeed positive. The metric is used for objectives where the minimisation of false positives is crucial. A low precision score is obtained when the model predicts messages as a minority class (1, 2, 3, or 4) that belong to the majority class (0), resulting in a high number of false positives. Since a normal precision score does not account for data imbalance, the objective of the macro-average precision was set to be maximised. This score divides the sum of all individual calculated precision scores per label, by the number of labels. This makes sure that all the labels have the same weight.

### 3.4.4 F1

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

The F1 score uses precision and recall to compute the harmonic mean between the two metrics. The metric considers both false positives and false negatives equally resulting in a balanced view of the problem. Although F1 is a suitable option for the problem in this research, with effective handling of imbalanced data and minimising misclassifications, the importance of minimising false positives weighs heavier than the objective of the F1 score.

## 3.5 Case Study Methods

By investigating individual cases regarding threats against a political figure using social media, the aim is to identify patterns in the distribution of distinct classified radicalisation phases before a threat and after, directed at a Dutch politician. This research does not aim to help in identifying and targeting users who are threatening Dutch politicians. However, this research can aid in predicting and preventing death threats, as well as enhance our understanding of when specific labels emerge.

Two distinct groups of four threats are being researched: physical and online threats. All eight cases were tried by the Dutch court. The cases are sourced from Rechtspraak.nl (an online website where tried cases are publicly available) and news reports.

To analyse the distributions of classified messages before a threat, several steps have to be taken. The obtained Telegram and Twitter datasets from Dataschool are used to filter out the dates of the incidents, specifically filtering on the incident date and a look forward of 5 days and a lookback period of 20 days. This range is chosen to fully capture the distribution and to see how the distribution acts after the incident. While considering the computational and time limitations, this range was chosen. Filtering between the specified time frames ensures researching only necessary messages. The time range-specific datasets are each labelled independently with the classifier. The messages, including the political usernames/names relating to the case study, are filtered. Due to computational and time limitations, the filtering of the online threats only contained the usernames. After filtering the Twitter and Telegram data, Table 3.6 is obtained which represents the counts of messages per label for both platforms.

In Table 3.6, the eight different cases each split into counts for Twitter and Telegram data are presented. Label 0 consistently has the highest count across all cases and sources, whereas label 1 to label 4 show significantly lower counts compared to label 0. In most cases, the count of messages on Twitter is much higher than that for Telegram. Specifically, label 0 shows a

| | **Case 1** | | **Case 2** | | **Case 3** | | **Case 4** | |
|---|---|---|---|---|---|---|---|---|
| | Twitter | Telegram | Twitter | Telegram | Twitter | Telegram | Twitter | Telegram |
| Label 0 | 37985 | 194 | 49861 | 1759 | 6428 | 875 | 4716 | 15 |
| Label 1 | 484 | 7 | 834 | 25 | 82 | 10 | 102 | 0 |
| Label 2 | 898 | 3 | 1014 | 41 | 63 | 11 | 28 | 0 |
| Label 3 | 272 | 0 | 588 | 31 | 52 | 10 | 55 | 0 |
| Label 4 | 209 | 2 | 85 | 4 | 5 | 0 | 26 | 0 |
| Total | 39848 | 206 | 52382 | 1860 | 6630 | 906 | 4927 | 15 |
| | **Case 5** | | **Case 6** | | **Case 7** | | **Case 8** | |
| | Twitter | Telegram | Twitter | Telegram | Twitter | Telegram | Twitter | Telegram |
| Label 0 | 63924 | 44 | 75969 | 28 | 10781 | 1 | 51093 | 3 |
| Label 1 | 689 | 0 | 704 | 0 | 62 | 0 | 545 | 0 |
| Label 2 | 394 | 1 | 433 | 0 | 85 | 0 | 250 | 0 |
| Label 3 | 313 | 0 | 728 | 0 | 33 | 0 | 276 | 0 |
| Label 4 | 118 | 2 | 123 | 0 | 11 | 0 | 68 | 0 |
| Total | 65438 | 47 | 77957 | 28 | 10972 | 1 | 52232 | 3 |

**Table 3.6:** Counts of messages per label for both platforms.

clear difference in counts between Twitter and Telegram for each case. The Telegram shows very low counts for label 1 to label 4, where these are often zero or close to zero.

The table shows significant differences between the physical (Cases 1-4) and online (Cases 5-8) cases. In the physical cases, the total counts for Twitter messages range from 4,927 to 52,382, and Telegram ranges from 15 to 1,860 messages, with label 0 having the highest counts for both sources. In contrast, the online cases show much higher total counts for Twitter messages, namely 10,972 to 77,957. However, much lower counts for Telegram, 1 to 47 messages. Label 1 to label 4 have low counts across both sets, but they are significantly lower in the online cases.

Finally, the distributions of the labels of the classified messages are analysed by plotting the counts of the labels per date over time. This gives a clear image of how the distribution behaves around an incident. Significant dates within the timeframe relating to the targeted politicians are also included to observe relevant peaks in the distribution regarding significant events. However, it is important to note that there can be differences between events, media events, and social media events, which may occur simultaneously or sequentially. In this research, it is not possible to differenti-

ate between these various types of events. This is due to the parallel development that progresses so rapidly it becomes indistinguishable. Therefore, for the remainder of this research, significant events are defined as those that have gained attention in any form of media.

Another test will be used, the Simple Moving Average (SMA), which shows a trend in the distribution by smoothing out short-term fluctuations. The technique is often used in time series analysis where it estimates the level of a value over time by sliding a window over the dataset (Johnston et al., 1999). This value is used to forecast future trends in the data. For this research, it is useful to identify how the frequency of classification of messages changes over time towards and after an incident. The window size of the moving average is the number of days that are used for the averaging process; in this research, it is chosen to use a window size of 7 due to the short range of observations (26 days).

# 4. Results

This section presents the findings from the hyperparameter tuning and the performance of the radicalisation phase classifier. The analysis begins with an examination of the model's performance using the confusion matrix and across key metrics, including accuracy, precision, recall, and F1-score. These metrics are crucial for assessing the effectiveness of classifying distinct radicalisation phases in messages. Furthermore, a detailed section is contributed to the eight case studies which highlights the practical application of the classifier by analysing specific instances of threats against Dutch politicians.

## 4.1 Classifier Performance

The performance of the radicalisation phase classification model was evaluated using several methods including accuracy, precision, recall, and F1-score. As mentioned in Section 3.3.7, the hyperparameters of the model are tuned. The best-performing model with a maximise macro-averaged precision objective was the model with the hyperparameters tuned with Optuna:

- epochs = 10
- learning rate = 1e-05
- weight decay = 0.001
- batch size = 16
- optimiser = AdamW

To present the performance of the model, a confusion matrix is used, as shown in Figure 4.1. This matrix is colour-coded from yellow to purple, where yellow indicates a higher number of correctly classified messages, and purple represents a lower number of correctly classified messages.

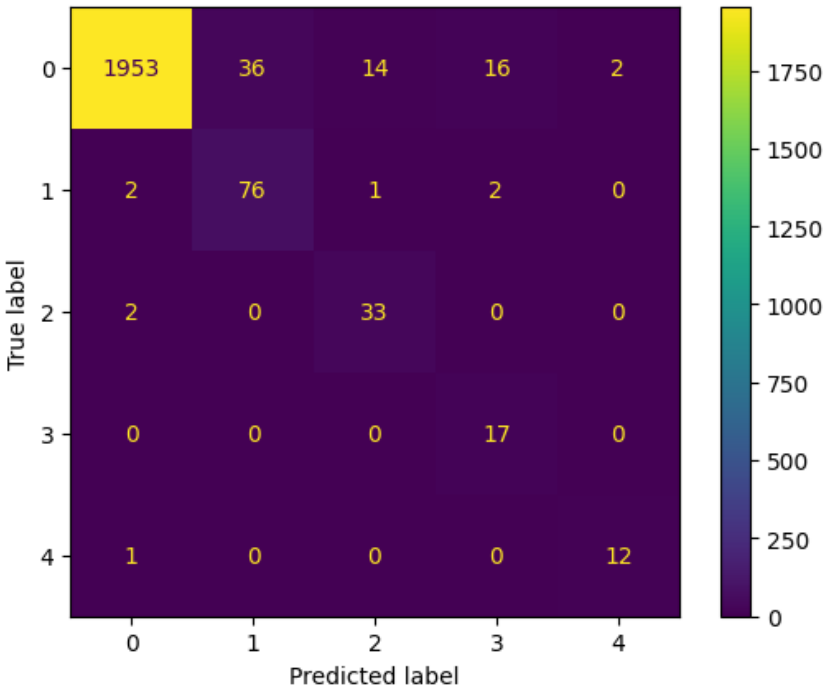


**Figure 4.1:** Confusion matrix

In the confusion matrix, label 0 shows a high number of correctly classified messages, with 1953 messages correctly classified as label 0. However, there are 36 messages misclassified as label 1, 14 messages misclassified as label 2, 16 messages misclassified as label 3, and 2 messages incorrectly classified as label 4. For label 1, the classifier correctly identified 76 messages. There were 2 messages misclassified as 0, 1 message incorrectly classified as 2, and 2 messages incorrectly classified as 3. Label 2 had 33 messages correctly classified, with 2 messages incorrectly classified as 0. Label 3 had 17 messages correctly classified, with no misclassifications. Finally, label 4 had 12 messages correctly classified, with 1 message incorrectly classified as 0. The high number of correctly classified messages for label 0 (1953) suggests that the classifier performs well in this category. The relatively lower numbers for other labels indicate areas where the classifier could be improved. To dive deeper into the strengths and weaknesses of the model, the accuracy, precision, recall, and F1 scores are calculated by using the formulas mentioned in Section 3.4 and the confusion matrix.

### 4.1.1 Accuracy

The accuracy of the model is 0.9891, which means that 98.91% of the messages were correctly identified by the model.

### 4.1.2 Precision (Column-wise view of confusion matrix)

As mentioned before, precision is the number of messages predicted with a label that were actually correct. The precision for label 0 was 99.74%, which indicates the classifier was correct for almost all its predictions for label 0. A lower precision score can be seen in labels 1 and 2 which were 67.86% and 68.75%, respectively. For both labels, the classifier predicted label 1 and 2 while these messages were actually label 0. An even lower precision score was when the classifier classified messages as label 3 (48.57%), where the message belonged to label 0 almost half of the time. The last label, 4, scored 85.71% precision.

### 4.1.3 Recall (Row-wise view of confusion matrix)

Recall is the number of messages with a true label that were correctly predicted by the classifier. The recall score for label 0 was 96.63%, indicating that the classifier correctly identified almost all actual label 0 messages. For labels 1 and 2, recall was higher at 93.83% and 94.29%, respectively, meaning the classifier successfully identified most of the messages that truly belonged to these labels. The recall for label 3 was perfect at 100%, suggesting the classifier identified all actual label 3 messages without missing any. For label 4, the recall was 92.31%, showing that most of the true label 4 messages were correctly classified.

### 4.1.4 F1-score

The harmonic mean of precision and recall, F1-score, provides balance between the two metrics. The F1-score for label 0 was 98.16%, meaning a high level of both precision and recall. Labels 1 and 2 had F1-scores of 78.67% and 79.41%, respectively, indicating a moderate balance between precision

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| 0 (Nothing) | 99.74% | 96.63% | 98.16% |
| 1 (Dehumanising) | 67%86 | 93.83% | 78.67% |
| 2 (Demonizing) | 68.75% | 94.29% | 79.41% |
| 3 (Violent Threat) | 48.57% | 100 | 65.43% |
| 4 (Death Threat) | 85.71% | 92.31% | 88.89 |

**Table 4.1:** Summary of metrics per label

and recall for these labels. The F1-score for label 3 was 65.43%, highlighting that while recall was perfect, the lower precision affected the overall balance. For label 4, the F1-score was 88.89%, showing a high balance between precision and recall for this label.

A summary of the metrics and the scores of each label are presented in Table 4.1. Since the accuracy is not calculated per label but over the whole model it is not present in the table.

The macro-averaged metrics for the classifier, as shown in Table 4.2, provide an overview of its performance across all classes and highlights both strengths and weaknesses. The macro-averaged precision of 0.741 is lower than the precision scores for labels 0 (0.9974) and 4 (0.8571), indicating that these classes are classified with higher accuracy than the average. Labels 1 (0.6786) and 2 (0.6875) have lower precision, contributing to the macro-averaged precision being lower than these individual values. The macro-averaged recall of 0.954 is higher than the recall for label 4 (0.9231) but lower than the perfect recall of 1.000 for label 3, suggesting that the model excels in identifying violent threat messages overall, with other labels achieving near-perfect detection. The macro-averaged F1-score of 0.821 is lower than the F1-scores for labels 0 (0.9816) and 4 (0.8889), meaning that these classes have more balanced precision and recall. However, the F1-scores for label 1 (0.7867) and label 2 (0.7941) are lower, showing that these labels have difficulties in balancing precision and recall effectively.

This comparison underscores the model's high performance in classifying labels 0 and 4, but also highlights weaknesses in precision for labels 1 and 2, and a need to balance precision and recall for label 3 to achieve a more consistent performance across all classes.

| Metrics | Score |
|---|---|
| Macro-averaged Precision: | 0.741 |
| Macro-averaged Recall | 0.954 |
| Macro-averaged F1-Score | 0.821 |

**Table 4.2:** Macro-averaged metrics

## 4.2 Case Study: Threats Against Dutch Politicians

In this section, the individual cases introduced in Section 3.5 are presented. For each case, a summary will be given of the event after which the results for the Twitter and Telegram data will be given. Lastly, a short conclusion of the case will be given.

### 4.2.1 Case 1: Physical Threat against Geert Wilders

This case involves the party leader of the political right-wing party PVV in the House of Representatives, Geert Wilders. He has been threatened with severe bodily harm in the period June 11, 2021 up to and including July 15, 2021 in The Hague, The Netherlands. A postal package containing a syringe/needle with blood and a letter with the text:

> *"Ik heb geld nodig op me rekenig. Komt te kort wand ik heb hulp nodig kennen jullie mijn nu eens helpen en die rat van de linden zit achter me geld aan. Mvg Peter Marc".*

Keywords that were used to filter the dataset are: *'Geert Wilders'*, *'@geertwilderspvv'*, *'wilders'*.

#### Twitter

In the first discussed case, the target date was set on July 15 2021, which can be seen in Figure 4.2 as the red vertical dashed line. Label 0 shows a fluctuating distribution with a peak on July 6 2021, with over 4000 messages related to Geert Wilders. A second, less significant peak occurred in the period between June 29 2021, and July 1 2021, where the peak reached over 3000 messages. The SMA trend shows a fluctuating decline towards the

target date. In the second subplot, the count of label 1 is plotted over time where it shows a fluctuating trend with several peaks. A noticeable spike occurs on June 30 2021, when the counts reach almost 50 messages.

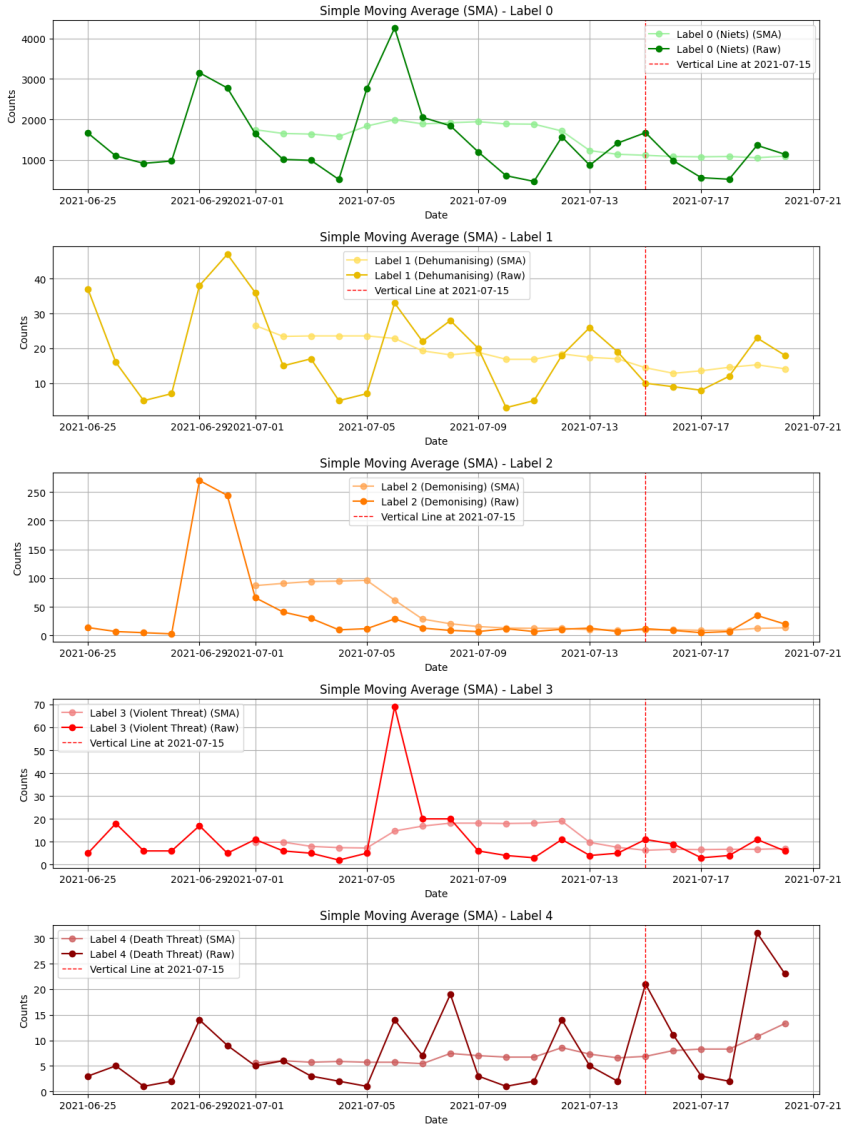


**Figure 4.2:** Twitter Plots Case 1

After this, a decline in counts with fluctuations can be observed. This can also be observed in the SMA trend, which shows a slight downward trend over the period. Label 2 shows a sharp peak around 29 June 2021, where the counts reach over 250 messages. Following this, the distribution quickly declines to a stable count of around 20 messages with no remarkable fluctuations. This decline and stable trend can also be observed in the SMA trendline. The third label plotted in the figure is label 3 which shows a simi-

lar distribution as label 2, however, the peak can be observed on 6 July 2021, where the counts reach just above 70 messages. After this peak, a quick decline in the counts can be observed to a stable number of 10 messages with no remarkable fluctuations. Aligning with the original distribution of the data, the SMA also shows a downward trend after the peak and a stable trend. Label 4 shows a fluctuating distribution with a significant peak on 19 July 2021 counting 35 messages. This fluctuating distribution can also be observed in the SMA trend which is a stable trend with a rise towards 19 July 2021. The target date does not correspond to any significant changes for all labels.

**Telegram**

Figure 4.3 presents the plots per label for Telegram messages. Only the distribution of label 0 is sufficient for analysis, where one major peak can be observed on July 6 2021 with over 40 messages related or mentioning Geert Wilders. The SMA trend of the label follows the same distribution as the Twitter SMA trend for label 0, a fluctuating decreasing trend over the time period. The other labels have an insufficient number of messages related to Geert Wilders. Therefore, no analysis is written about labels 1, 2, 3, and 4.

**Significant events in the time frame**

- 07/06/2021: Geert Wilders remains convicted by the court of group insult for his "fewer Moroccans" statement. [1]

The analysis of messages related to Geert Wilders during the period from June 11, 2021, to July 15, 2021, reveals several patterns. The activity on Twitter shows significant peaks on June 30 and July 6, which corresponds with the significant date when Wilders remained convicted for group insult. These peaks suggest a heightened public discourse and possibly increased higher-label messages around these dates. The Telegram data follows these findings, however they appear in a smaller volume of messages.

[1] https://www.recht.nl/rechtspraak/?ecli=ECLI:NL:HR:2021:1036)

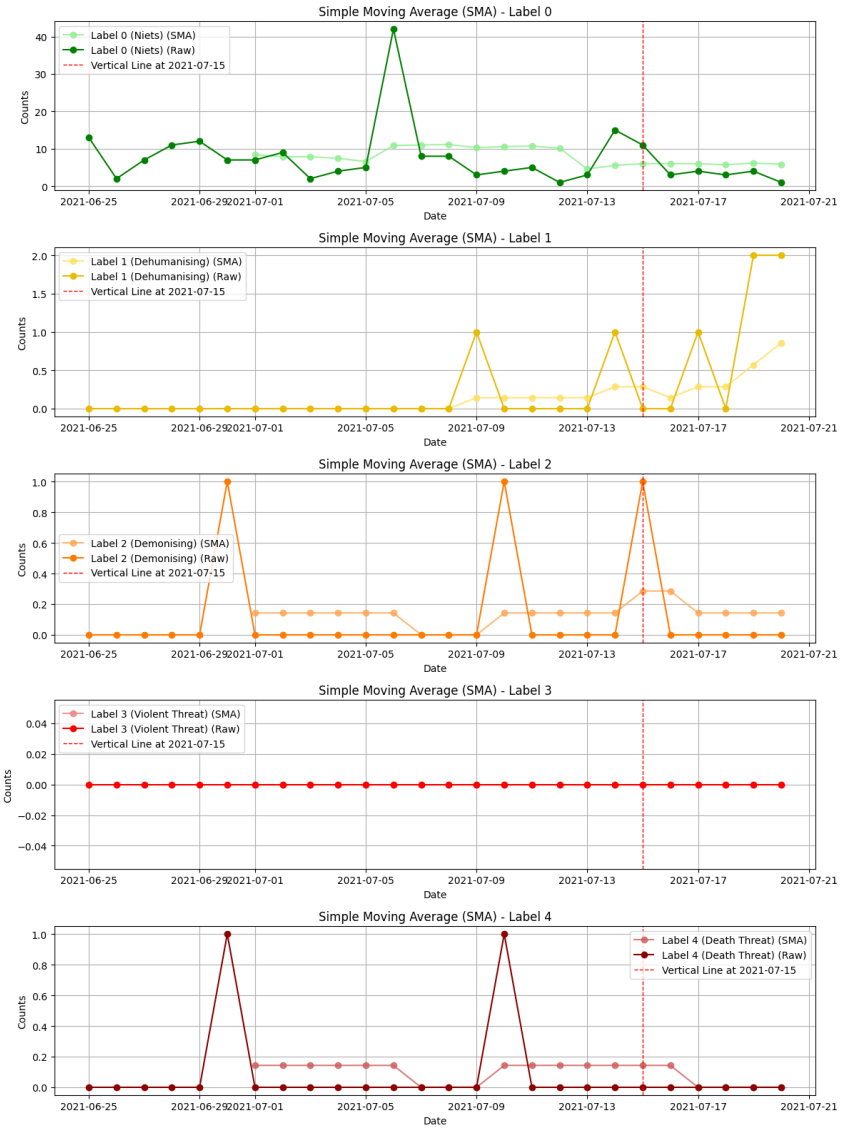


**Figure 4.3:** Telegram Plots Case 1

### 4.2.2 Case 2: Physical Threat against Sigrid Kaag

On Jan. 5, 2022, a man was arrested in front of the home of Sigrid Kaag, the Dutch politician, then Minister of Finance and leader of the D66 political party. The suspect walked to and stood at the front door of Sigrid Kaag with a burning torch. Furthermore, threats were shouted including:

- *"wij hebben wel een hele bijzondere fakkel bij ons, dat ga ik jullie zo in beeld brengen"*
- *"wauw [medeverdachte] wat is dat voor fakkel"*

- *"speciaal voor deze dame die hier woont" en/of (vervolgens)"*

While doing this, he filmed his actions and these images were distributed online.

Keywords that were used to filter the dataset are: *'Sigrid Kaag'*, *'@SigridKaag'*, *'kaag'*

**Twitter**

The target date for the second physical case was set at January 5, 2022. The counts of labels 0, 1, 2, and 3 follow the same distribution over time as shown in Figure 4.4. A small peak can be seen on December 20, 2021, which reached almost 2500, 60, 75, and 110 messages for label 0, 1, 2 and 3, respectively. After the first peak, the distribution remained relatively stable up until January 6, 2022, where a significant peak can be observed. At this peak, the counts for labels 0, 1, 2 and 3 rose above 9000, 120, 150, and 140 messages, respectively. The SMA trends for these labels show the same results, with a stable period before January 6, 2022. The trend is starting to increase from January 5, 2022. Label 4 performs a bit differently than the other labels.

Leading up to the target date, label 4 does not have a peak, whereas the other labels had one on December 20, 2021. Furthermore, it does not have a significant peak on January 6, 2022, but does have a peak on January 9, 2022. The SMA trend does have a similar increase as the other labels.

**Telegram**

More Telegram messages were available in the time frame related to Sigrid Kaag in contrast to Case 1, however label 4 still has an insufficient number of messages to analyse (Figure 4.5). Label 0 shows three main peaks occurring on December 20 2021, January 6 2022, and January 9 2022, reaching almost 200 messages for the first two dates and over 200 messages for the last date. For label 0 and 2 there is a stable period of a low count of messages between the first and second peak. Label 1 follows a more fluctuating distribution with a peak on December 20 2021 reaching 4 messages, and a peak on January 9 2022 reaching 5 messages. Label 3 has a distinct distribu-

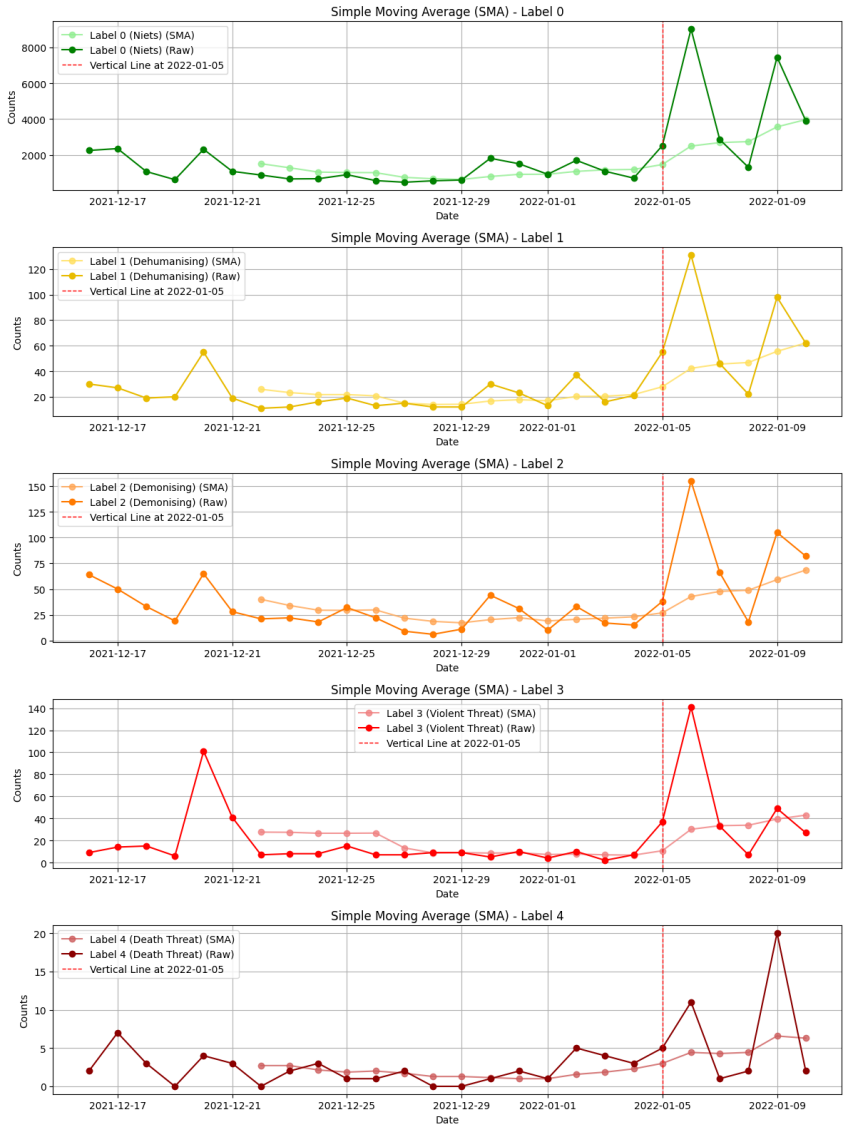


**Figure 4.4:** Twitter Plots Case 2

tion, where it shows one major peak on December 30 2021, which reached over 10 messages.

The SMA trends of labels 0, 1, and 2 all follow the same trend where a small decrease in counts can be seen from December 26 2021, after which the trend rises. Label 3 follows a different trend where a small rise on December 29 2021 can be observed which remains stable until a decrease on January 5 2022 after which the trend decreases.

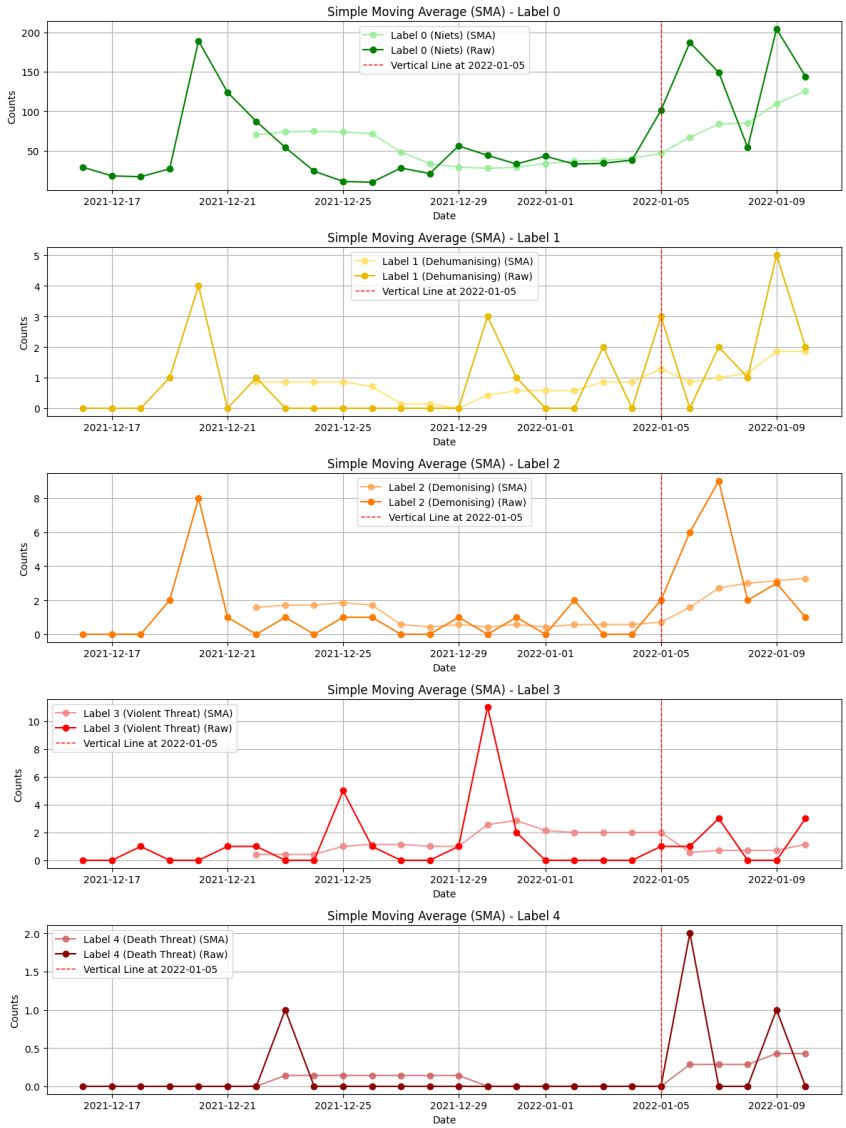


**Figure 4.5:** Telegram Plots Case 2

### Significant events in the time frame

- 06/01/2022: Various newspapers wrote about the torch incident related to Sigrid Kaag on this day.

The analysis of messages related to Sigrid Kaag from December 20, 2021, to January 9, 2022, reveals several patterns. Twitter activity shows peaks on December 20, 2021, and January 6, 2022, the latter correlating with responses of the torch incident involving Kaag on January 5, 2022. These peaks suggest heightened public discourse and increased higher-label messages around these dates. Telegram data shows similar trends with peaks on December

20, 2021, January 6, 2022, and January 9, 2022.

### 4.2.3 Case 3: Physical Threat against Christianne van der Wal

On July 6th, 2022, a man was arrested on suspicion of threatening Christianne van der Wal, Minister for Nature and Nitrogen Policy. On the back of his truck was a text mentioning the Minister of Nature and Nitrogen in the same line with Pim Fortuyn and Theo van Gogh, who were both murdered. Behind the names of Van Gogh and Fortuyn were the dates on which they were killed, and behind that of the minister a question mark. The minister pressed charges: the public prosecutor assessed the statement as a criminal threat. In addition, the suspect is charged with sedition.

Keywords that were used to filter the dataset are: *'Christianne van der Wal'*, *'Christianne van der Wal-Zeggelink'*, *'@MinisterNenS'*, *'van der Wal'*, *'Zeggelink'*.

**Twitter**

The third case presented the threat against Christianne van der Wal with the target date of July 6, 2022. In Figure 4.6 it can be seen that every label behaves differently in this case, except for label 0 and 1. Label 0 and 1 show two main significant peaks on June 24, 2022 (750 and 12 messages for label 0 and 1, respectively), and June 29, 2022 (over 800 and 13 messages for label 0 and 1, respectively). The period before and after these dates has a relatively stable distribution with a decrease after the target date. The SMA trend also shows a rising trend leading up to June 29, 2022 for label 0 and June 30, 2022 for label 1, after which the trend decreases. Label 2 seems to have a different distribution, where one main peak on June 25, 2022, can be observed which reached 12 messages. The SMA trend does follow a similar pattern to the trend of label 0 and 1, where the increase up until June 29, 2022, can be seen after which the trend decreases. In the Figure can be seen that label 3 has one significant peak on June 29, 2021, reaching a total of 15 messages. The distribution shows a clear rise up until this date, and a clear downward trend after this date. The SMA trend rises until the date where it remains

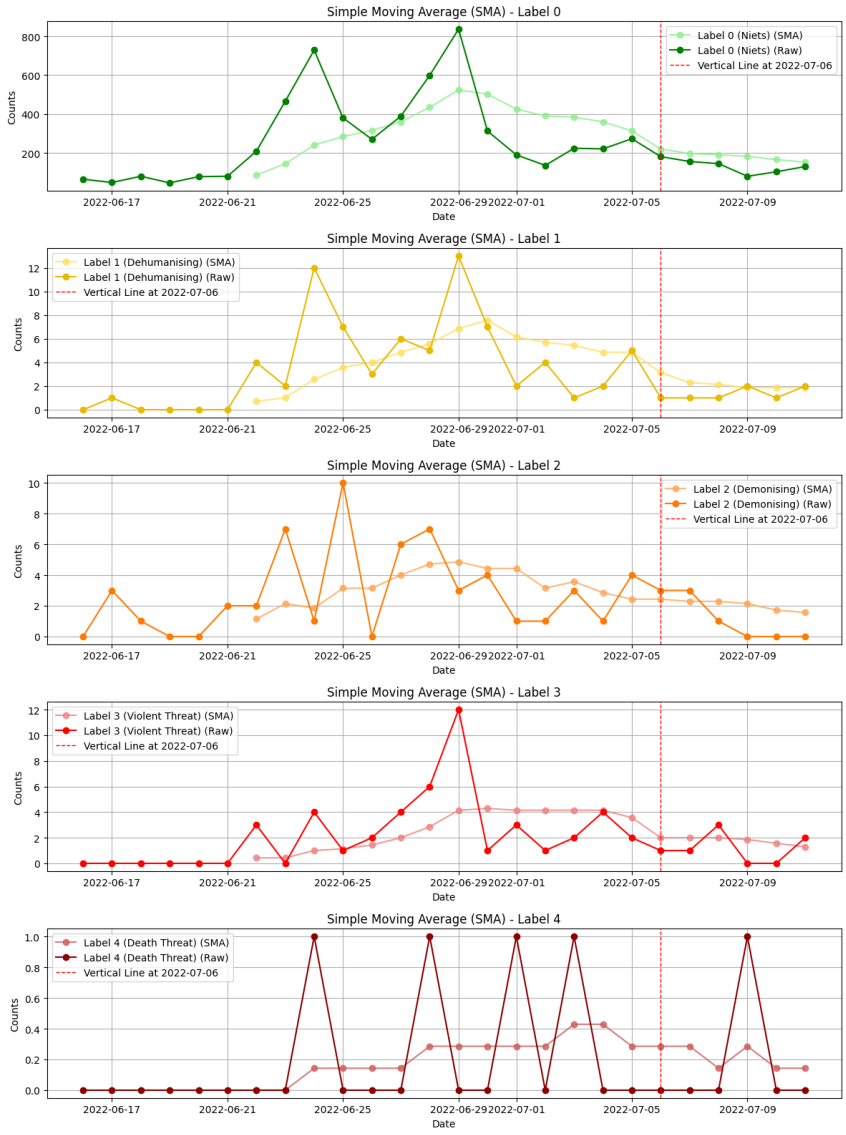


**Figure 4.6:** Twitter Plots Case 3

stable for six days, after which it declines. In the timeframe, the number of labels that were classified with label 4, was very small (5). Therefore, no results of this label will be discussed.

### Telegram

The Telegram data for this case is shown in Figure 4.7. Label 0 shows a fluctuating distribution where one major peak can be observed on June 28, 2022, where it reached over 120 messages related to Van der Wal. The SMA trend is showing a rising trend up to June 29, 2022 where it remains stable until July 4, 2022 after which it declines. Label 1 shows one peak on June 26 2022 where it reached 5 messages. The SMA trend for label 1 is relatively

stable over the time frame. A significant peak can be observed for label 2 after the target date on July 8, 2022 where it reached 6 messages. The trend of the label is stable accross the time frame. The fourth label, label 3, is also showing one peak, which is on June 26, 2022 reaching 7 messages. The SMA trend for this label is also a stable trend over the time frame. Label 4 has no messages which means that there are no results to present.

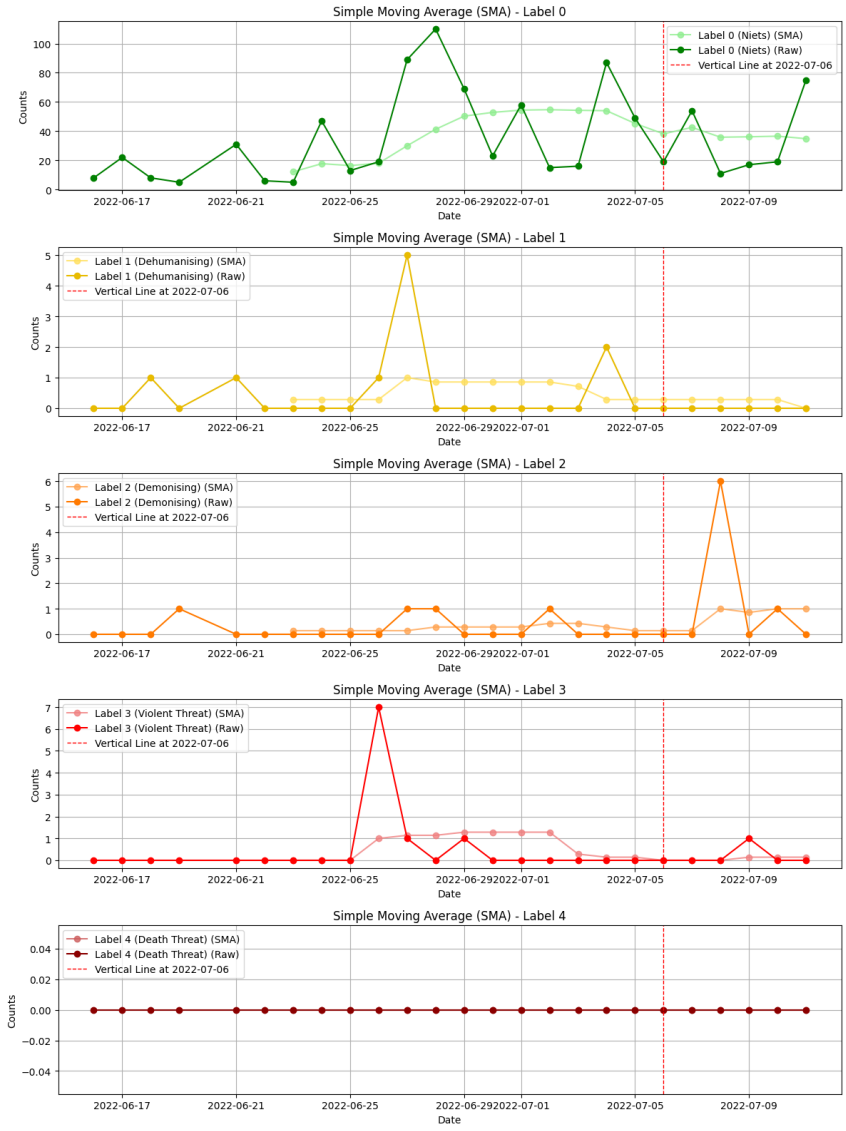


**Figure 4.7:** Telegram Plots Case 3

### Significant events in the time frame

- 10/06/2022: Christianne van der Wal announces new nitrogen plans.

- 14/06/2022 - 30/06/2022: Farmers protests against the new nitrogen plans. https://nipv.nl/wp-content/uploads/2022/09/20220912-NIPV-Boerenprotesten-in-de-zomer-van-2022.pdf

The analysis of messages related to Christianne van der Wal from June 10 to July 8, 2022, reveals peaks in activity coinciding with key events such as the announcement of new nitrogen plans and subsequent farmers' protests. Twitter and Telegram data both show significant spikes in late June, particularly around June 24 and June 29.

### 4.2.4 Case 4: Physical Threat against Dilan Yeşilgöz

This case involves the Secretary of State for Economic Affairs and Climate, Dilan Yeşilgöz from the VVD party. She has been threatened with a crime against her life as well as aggravated assault, via letter with the following text:

- *"ik kom binnenkort langs op jouw ministerie of bij jouw thuis om jou met een spuitje, een bijl of een handgranaat uit de weg te ruimen"*
- *'ik ga jou kapot schieten, verrekkes Turks kutwijf'*
- *'ik ga jou opzoeken in Den Haag om daar jouw buik op te snijden met een vlijmscherp stanleymes'*

The exact date of the threat is not known, however several sources state that the incident took place in June 2022. Therefore, 15 June is chosen as the date of incident.

Keywords that were used to filter the dataset are: *'Dilan Yeşilgöz', 'Dilan Yesilgoz', '@DilanYesilgoz', 'Yesilgoz'*

**Twitter**

The plots of the last physical threat case discussed, are shown in Figure 4.8. The first plot of label 0 shows a stable trend up until June 17, 2022 after which a peak can be observed on June 18, 2022. The peak reached almost 1200 messages after which it declined fast. The SMA trend also shows a

stable trend with a small rise from June 17, 2022 to June 18 2022 after which it remains stable. Label 1 shows a fluctuating trend with a peak on June 7 2022 reaching 17 messages after which the distribution keeps fluctuating. The SMA trend shows a rising trend from June 3 2022 to June 10, 2022 after which a small decline can be seen. The third plot shows the distribution of label 2 which has one significant peak on June 8, 2022 which reached 12 messages, the rest of the distribution remains low and stable. This can also

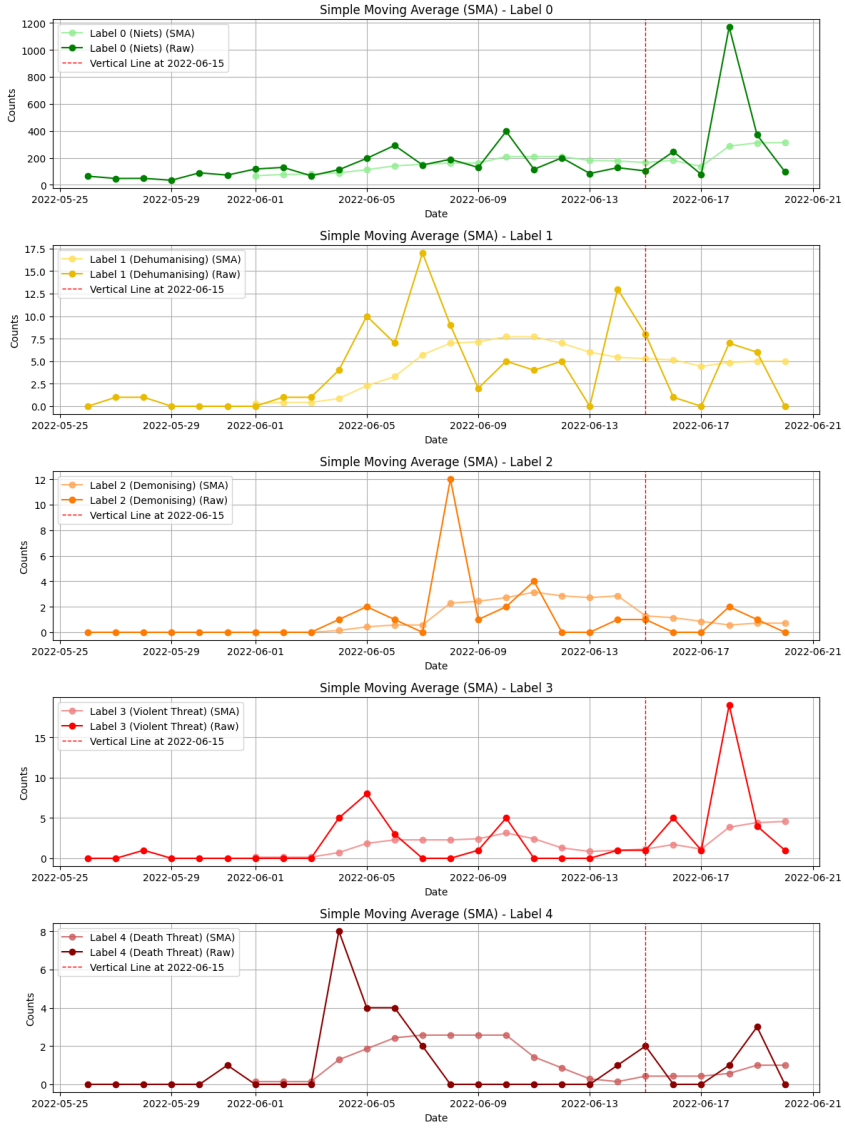


**Figure 4.8:** Twitter Plots Case 4

be seen in the SMA trend which does not show a significant trend. Label 3 shows a stable distribution with a couple of minor peaks and one significant peak on June 18, 2022 reaching 19 messages. The SMA trend shows a

stable trend with a small increase from June 17, 2022. Label 4 shows a significant peak on June 4, 2022 which reached 8 messages. Before this peak the distribution is stable, and after the peak minor peaks can be observed.

**Telegram**

An insufficient number of Telegram messages were obtained related to Yeşilgöz to perform any analysis on. Therefore, no results are written about the Telegram data for case 4.

**Significant events in the time frame**

No significant events were found related to Yeşilgöz in the time frame of this case.

The analysis of messages related to Dilan Yeşilgöz from June 1 to June 30, 2022, reveals several peaks in Twitter activity on June 4, June 8, and June 18, with the highest spike on June 18, reaching nearly 1200 messages. However, these peaks do not correspond to any specific events within the timeframe, indicating that no clear patterns are observable. Due to the insufficient number of Telegram messages, no analysis was performed for this platform.

### 4.2.5 Case 5: Online Threat against Mark Rutte and Hugo de Jonge

A man threatened Mark Rutte and Hugo de Jonge around November 2, 2021, with a crime against life/severe bodily harm, by posting on Twitter an image showing the following: a passenger car situated in the woods. Behind the car lie two oblong plastic-wrapped 'objects' in the shape of human bodies. A shovel is shown next to the 'bodies'. Written above the image were the threatening words:

> *"geen persconferentie? Sorry mensen! De persconferentie gaat niet door vanavond..."*

Keywords that were used to filter the dataset are: *'@MinPres'*, *'@hugodejonge'*.

**Twitter**

Presented in Figure 4.9, the 'Nothing' and 'Demonising' labels show a small increasing fluctuating distribution with a peak on November 2, 2021 reaching over 6500 and 40 messages, respectively. The SMA trend also follows an increasing trend which remains stable after November 3 2021. Label 1 shows a fluctuating distribution with two major peaks on October 29 2021 and November 2 2021, reaching over 55 and 50 messages, respectively. The SMA trend follows a steep rising trend towards November 4 2021. Label 3 follows a similar distribution as label 0 and 2, however leading up to the peak reaching 35 messages, the distribution fluctuates more with minor fluctuations. The SMA trend of label 3 shows a stable rising trend which remains stable after November 3 2021. The last label shows a fluctuating distribution with a significant doubled peak on November 1 2021 and November 3 2021 reaching 14 and 13 messages, respectively. The SMA trend follows the same trend as label 2.

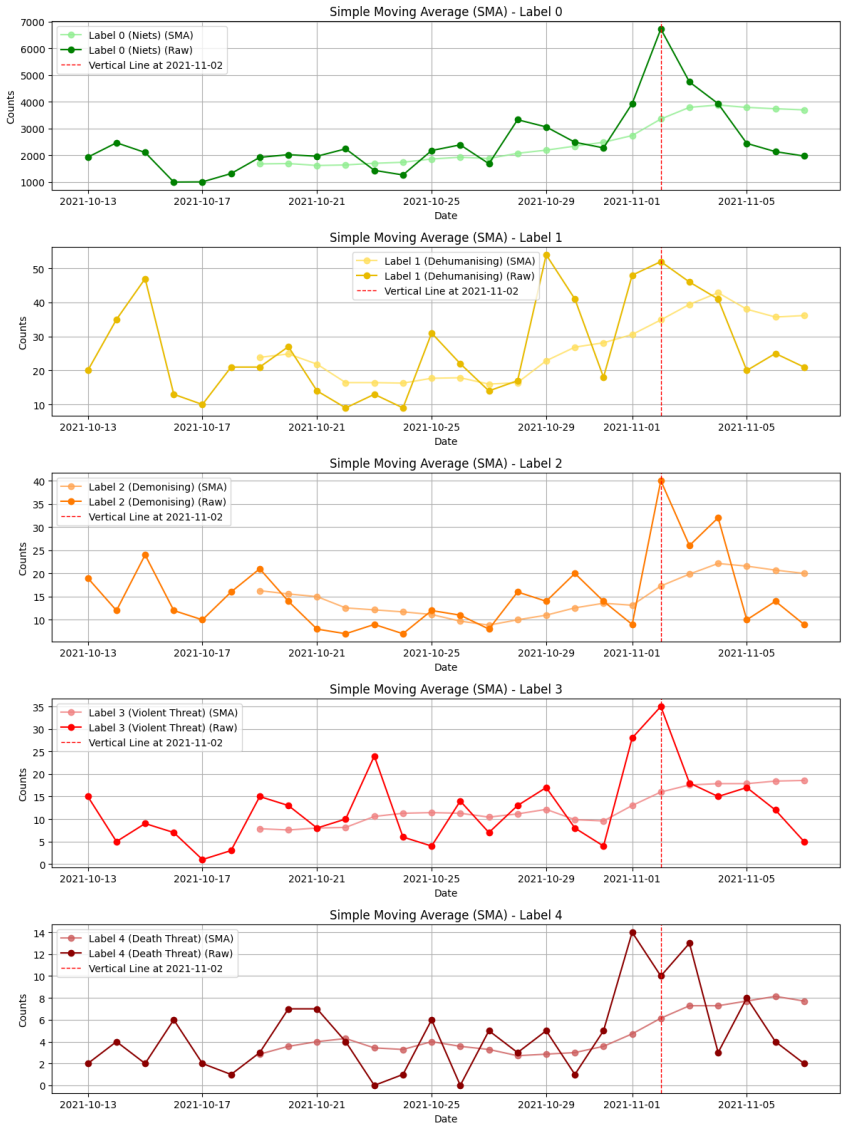


**Figure 4.9:** Twitter Plots Case 5

### Telegram

An insufficient number of Telegram messages were obtained related to Rutte and De Jonge to perform any analysis on. Therefore, no results are written about the Telegram data for case 5.

### Significant events in the time frame

- 02/11/2021: COVID Press conference Mark Rutte and Hugo de Jonge.

The analysis of messages related to Mark Rutte and Hugo de Jonge around November 2, 2021, reveals significant patterns. Twitter activity shows peaks

on October 29, November 1, and November 2, with the highest peak on November 2, which coincides with a COVID press conference by Rutte and De Jonge. The 'Nothing' and 'Demonising' labels show increased activity, particularly on November 2, with over 6500 and 40 messages, respectively. The SMA trends indicate a rising and then stable trend after November 3. Due to the insufficient number of Telegram messages, no analysis was performed for this platform. The peak on November 2, suggests that Rutte's and de Jonge's public appearance and related COVID-19 announcements were likely triggers for the increase in hostile messages.

### 4.2.6 Case 6: Online Threat against Mark Rutte

A man has threatened Mark Rutte in the period October 6, 2021 up to and including November 21, 2021with inflammatory and threatening messages on Twitter. Messages include:

- *"[slachtoffer] is een misdadiger die de doodstraf volledig verdiend. Beter dat hij onder gedoken blijft. Bekijk de video van Politiektok! Tiktok"*
- *"Ik word door [slachtoffer] ook psychopaat, Er gaat geen dag meer voorbij dat ik niet van fantaseer van zijn executie. Bijna elke dag droom ik er van hem te mogen ophangen! Naast zijn grootste vrienden [naam 1] en [naam 2]"*
- *"Als Volkert, [slachtoffer] morgen vermoord zal ik hem nooit meer vervloeken!"*

Keyword that is used to filter the dataset: *'@MinPres'*.

**Twitter**

Labels 0, 1, and 2, all follow the same distribution with a significant peak on November 12 2021 reaching over 6000, 60, and 35 messages relating Mark Rutte, shown in Figure 4.10. Furthermore, the SMA trends of the labels follow the same trend, namely a rising trend towards November 18 2021 after which it remains stable. Label 3 follows a different distribution where it reached a peak on November 18 with almost 140 messages. After this peak a

steep decrease can be observed after which the distribution rises again. The SMA trend also follows this trend of distribution, where a stable increase can be seen across the time frame. Label 4 follows a fluctuating distribution with a major peak on November 3 2021. Due to the fluctuating distribution of the messages, the SMA trend is relatively stable.

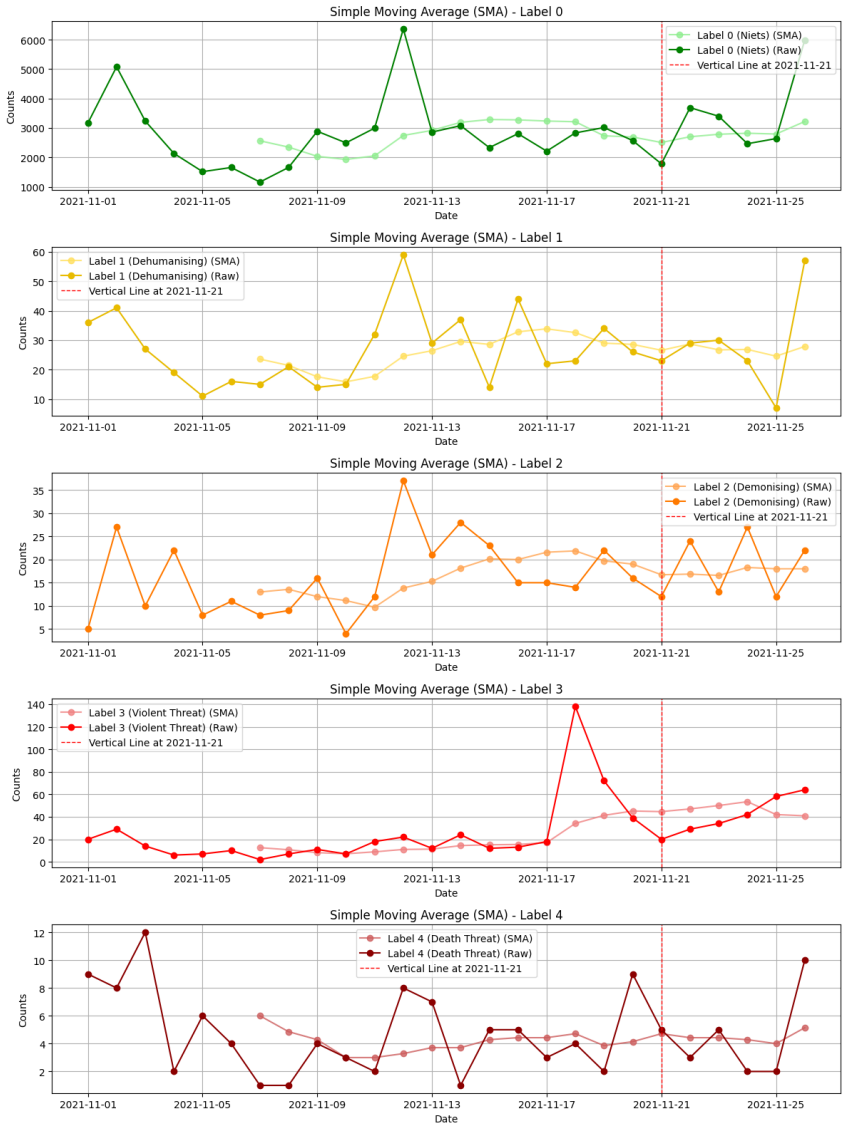


**Figure 4.10:** Twitter Plots Case 6

### Telegram

An insufficient number of Telegram messages were obtained related to Rutte to perform any analysis on. Therefore, no results are written about the Telegram data for case 6.

**Significant events in the time frame**

- 12/11/2021: COVID Press conference Mark Rutte and Hugo de Jonge.
- 26/11/2021: COVID Press conference Mark Rutte and Hugo de Jonge.

Between October 6, 2021, and November 21, 2021, Mark Rutte experienced a significant rise in threatening messages on Twitter. Analysis shows a peak in such messages on November 12, 2021, which correlates with a key COVID-19 press conference by Mark Rutte and Hugo de Jonge. The number of threatening messages increased steadily leading up to November 18, after which it remained stable. This trend suggests that Rutte's public engagements and related COVID-19 announcements were likely triggers for the increase in hostile messages.

### 4.2.7 Case 7: Online Threat against Hugo de Jonge

A man threatened Hugo de Jonge (Minister of Public Health, Welfare and Sports) and/or his daughter with a crime against life around January 10, 2021. This occured on Instagram, where the man wrote:

> *"Im gonna kill your father"*

Keyword that is used to filter the dataset: *'@hugodejonge'*.

**Twitter**

In Figure 4.11 it can be observed that, the first two labels follow the same distribution which show two main peaks on January 5 2021 reaching 1500 and 13 messages, respectively, and January 12 2021, reaching over 1750 and 14 messages, respectively. The SMA trend of both labels follow a stable trend with no fluctuations. Label 2 follows a low and stable distribution with a peak on January 1 2021 reaching 50 messages related to De Jonge. This stable trend can also be seen in the SMA trend. Label 3 follows a similar distribution as the first two labels only with three peaks on January 3 2021, January 5 2021, and January 12 2021, reaching 4, 5, and 7 messages, respectively. The SMA trend is also following the same trend as labels 0 and 1, a stable trend with no fluctuations. Label 4 has too few messages to present

any results about.

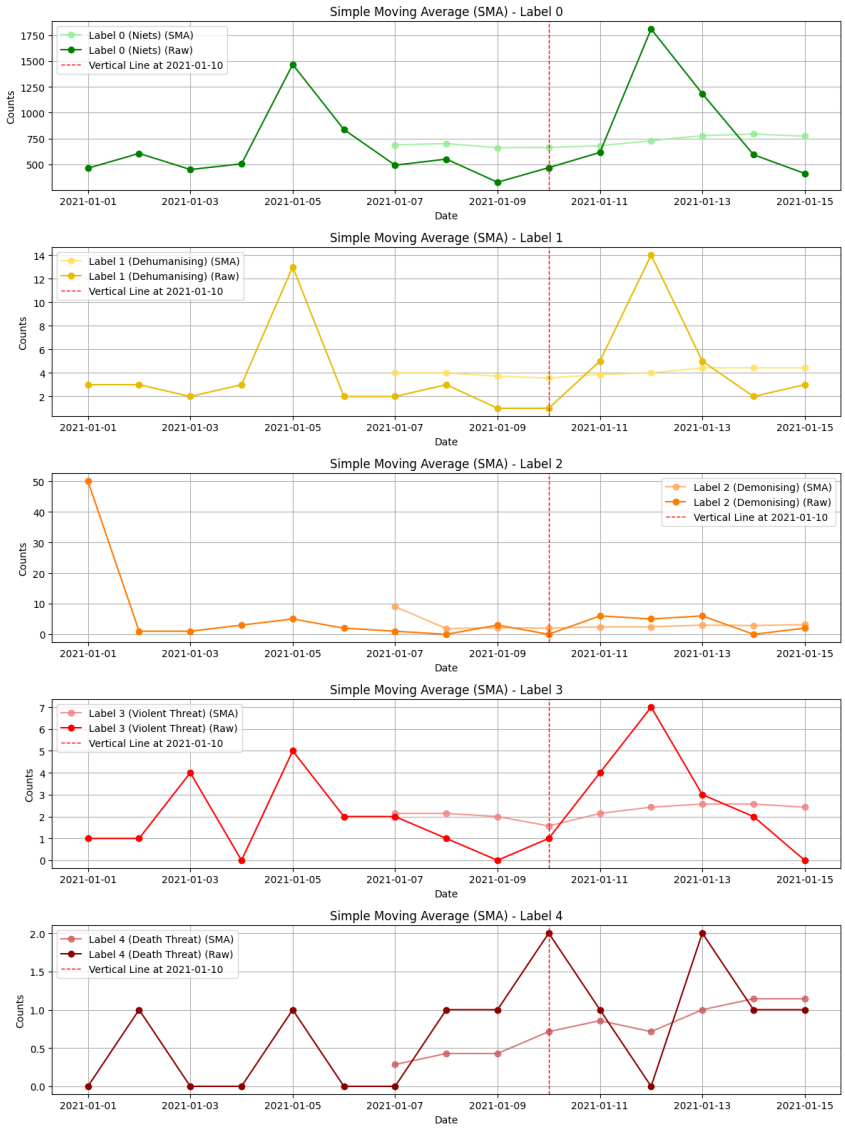


**Figure 4.11:** Twitter Plots Case 7

### Telegram

An insufficient number of Telegram messages were obtained related to De Jonge to perform any analysis. Therefore, no results are written about the Telegram data for case 7.

### Significant events in the time frame

- 05/01/2021: House of Representatives debate on developments surrounding Coronavirus.

- 12/01/2021: COVID Press conference Mark Rutte and Hugo de Jonge.

This case presented the threat on Instagram directed at Hugo de Jonge, where Twitter data shows peaks in messages related to De Jonge. On January 5 and January 12, 2021, corresponding to significant events: a House of Representatives debate and a COVID-19 press conference. The fluctuating message trends on Twitter reflect a heightened public reaction linked to these key events, although Telegram data was insufficient for analysis.

### 4.2.8 Case 8: Online Threat against Hugo de Jonge, Mark Rutte, and Jaap van Dissel

In May, 2021, a man threatened Hugo de Jonge, Mark Rutte and Jaap van Dissel through a video he posted on Telegram. The video showed Prime Minister Rutte, Minister De Jonge and OMT chairman Van Dissel with a rope around their necks. At the end of the video the following text appeared:

> *"Eindelijk hun verdiende straf! Van harte gefeliciteerd, Joost en familie! Dit toekomstige cadeau is voor jouw!"*

Keywords that were used to filter the dataset are: *'@MinPres'*, *'@hugodejonge'*, *'@RIVM_vDissel'*.

### Twitter

The last online case shows a similar distribution for the first four labels, 0, 1, 2, and 3, where a major peak can be observed in Figure 4.12 on May 14 2021 where the labels reached 7000, 90, 50, and 50 messages, respectively. The SMA trends of all the labels are following the same trend, A stable trend up to May 13 2021 after which it rises and remains stable. Label 4 follows a more fluctuating distribution with three peaks on April 29 2021, May 8 2021, and May 14 2021, reaching 7, 7, and 8 messages, respectively. The SMA trend of label 4 follows a stable distribution with no fluctuations.

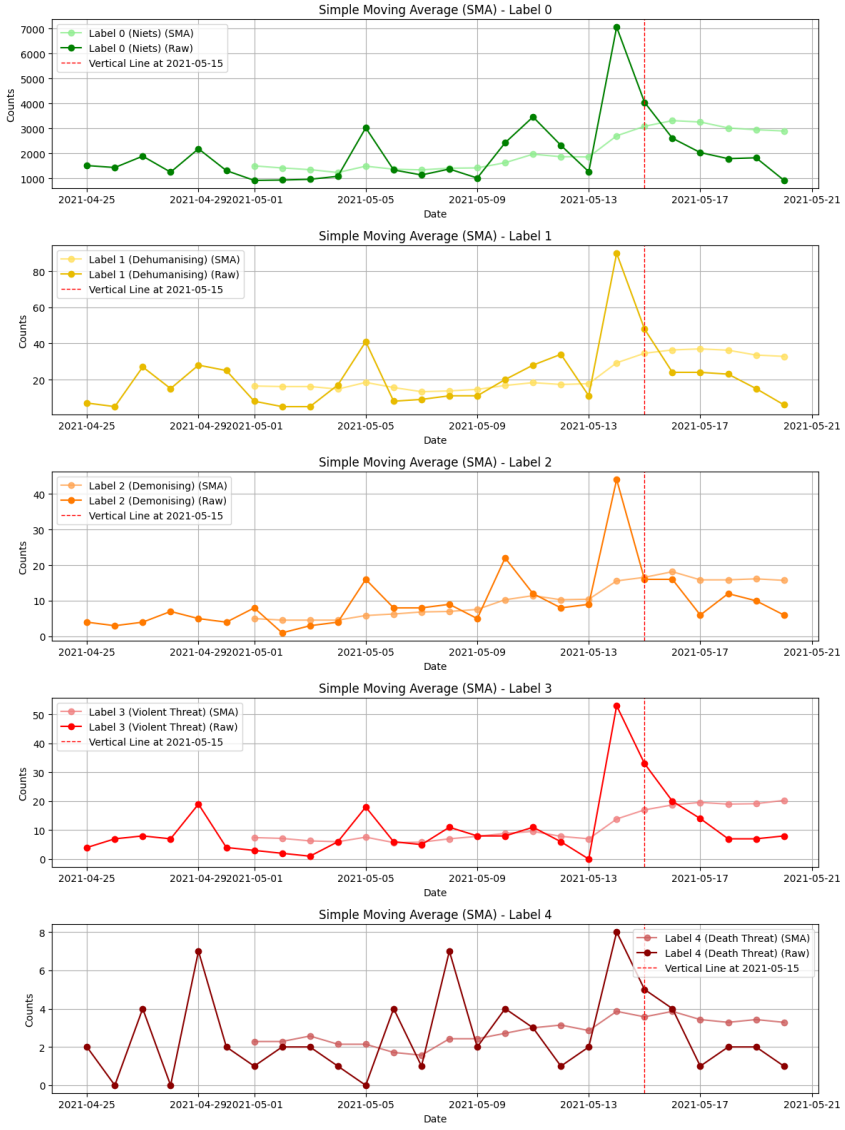


**Figure 4.12:** Twitter Plots Case 8

**Telegram**

An insufficient number of Telegram messages were obtained related to De Jonge, Rutte, and Van Dissel to perform any analysis on. Therefore, no results are written about the Telegram data for case 8.

**Significant events in the time frame**

- 05/01/2021: 14/05/2021: Public report by OMT regarding the advice for COVID rules.

In May 2021, a video threatening Hugo de Jonge, Mark Rutte, and Jaap van Dissel was posted on Telegram. Twitter data shows a significant peak in messages of all labels on May 14, 2021, the same day as an OMT report on new COVID-19 rules. The Telegram data was insufficient for analysis, but the Twitter trends reflect heightened public reaction around the time of the OMT report.

# 5. Discussion

This section discusses the results presented in Section 4 in two parts. Firstly, an interpretation of the classifier results and a comparison with the results from the literature is given. Next, the limitations and future work of the classifier are presented. Secondly, an interpretation of the results of the case studies, limitations and future work are presented in this section.

## 5.1 Interpretation of Classifier results

The primary objective of the classifier for detecting radicalisation phases was to minimise false positives and accurately classify messages. Although the model's accuracy score of 98.91% is promising, precision is the crucial metric for evaluating its effectiveness in meeting the objective. However, when comparing the accuracy to the previous mentioned classification techniques, it can be seen that the model exceeds the best performance of 63% by Parmar et al. (2018) and 96% by Kamath et al. (2018). The model does not exceed the performance of the MCNN-LSTM model by Hasib et al. (2023) which scored an accuracy of 99.71%. Furthermore, the classifier outperforms the RoBERTa model on multi-class classification studied by Rajendran et al. (2022) regarding the 95% accuracy score. The XLM-RoBERTa-large model used in the study by Wang and Banko (2021) is not outperformed based on the F1 score of 93.5%.

The high precision scores for Label 0 (99.74%) and 4 (85.71%) demonstrate the model's capability to identify non-radicalised messages and death threats. The macro-averaged precision of the model, 74.10%, can be used for comparing the performance of the model on the different labels. From this it can be deducted that the model's precision needs to be improved for label 1, 2 and 3. The reason why the model is performing less than the other labels could be due to several factors, including overfitting of words and inconsis-

tent manual labelling. These reasons are discussed in the limitations of this research.

An interesting finding of this research is the correlation of words or sentences with the different phases. After researching individual sentences, the model seems to be overfitting on certain words, and therefore relates these words to distinct labels. For example, a sentence where a lot of words relating to death ('dead', 'suicidal thoughts', 'cancer') are used but without any death threat, is labelled as label 4. This shows that the model does rely heavily on words and less on context and sentiment of the sentence, which results in poor generalisation to unseen data. This limitation of overfitting is difficult to discuss, since the model has a large number of parameters many of which are not fully understood or adjustable. Several points of limitations were found during the process of developing, training and testing the classifier.

### 5.1.1 Inconsistent Manual Labelling

After testing the classifier on the test set, several individual misclassifications were researched. It appears that several labelled messages were wrongly or inconsistently labelled. The model does, however, predict the messages which are clearly belonging to a phase correctly while the label is another phase. A clear example of mislabelling can be seen in the following message:

> *"Rinus .... aanhanger van de landverrader Rutte .. de nieuwe nazi's melden zich al ... https://t.co/2Axi4q1EZm"*

Due to the use of the word 'landverrader' before mentioning Mark Rutte, the message clearly belongs to label 1 (Dehumanising). While the model classified this message as phase 1, it was manually labelled as label 0 (Nothing). An example of inconsistent labelling can be seen in the following two sentences. The first sentence was labelled as a label 0 message:

> *'@ngundogan77 Ik, maar dan tegen jou WEF pop'*

The second sentence was labelled as a label 1 message:

*'Geweldig! GFYS WEF puppet @MinPres https://t.co/PSFUDDrJVM'*

While the context of the messages differs, it is clear that both messages belong to phase 1 since both political figures are being described as WEF doll/puppets.

### 5.1.2 Synthetic Data Issues

While the process of paraphrasing worked effectively in generating more messages for the minority labels, a bias might be introduced. Paraphrased sentences might focus on specific words which become overrepresented for certain labels, which alters the learning process of the classifier. Furthermore, as stated earlier in Section 3.3, language is a limited set of words. One sentence cannot be paraphrased multiple times; therefore, it is most likely that some paraphrased messages are almost the same as other paraphrased messages from the same message. Lastly, the number of paraphrased sentences was not theoretically grounded, which affected the data distribution and might have affected the performance of the model.

### 5.1.3 Optimisation of the Model

Due to time and computational constraints two main optimisations of the model were excluded from the research. The first optimisation is the hyperparameter tuning. The limited exploration of the hyperparameter space could have resulted in a parameter set which performs poorly compared to other sets which could have been explored without the constraints. The other optimisation that should have been included is cross-validation, which assists in hyperparameter tuning, prevents overfitting, and provides an estimate of how well the model generalises to unseen data.

To build upon this research, future work should focus on the manual labelling process to reduce inconsistencies, exploring techniques to mitigate overfitting, and conducting comprehensive hyperparameter tuning. Additionally, incorporating cross-validation will help evaluate the model's performance and generalizability. Improving the model's ability to understand context and sentiment, rather than relying heavily on specific words, can

also enhance its precision and effectiveness.

## 5.2 Interpretation of Case Study results

From the results of each case study, it can be deducted that the analysis did not show significant correlations between the distribution of radicalisation phases and the specific dates of the threats. The timing of the threats, therefore, does not follow a predictable pattern that is relative to the target dates. Due to the fact that there is no correlation, radicalisation phases do not specifically cluster significantly around the dates of incidents. However, the results show significant correlation between the volume of messages and certain external events. Peaks in the distribution and rising SMA trends were observed during periods such as demonstrations and COVID-19 press conferences. This shows that public events can boost political discourse online, including the spread of messages higher on the ladder.

An interesting finding is this the data showed that when the number of messages that belongs to the phase "Niets" (nothing) rise, other phases rise with it. This suggests that the overall political discourse rises during an event which leads to an increase in messages containing all of the radicalisation phases.

From the results can be seen that there is no difference between online and physical threats in terms of correlation between the data and target dates. Both types reflect the correlation between public events and the data.

A major limitation of the case study is the limited amount of Telegram messages that were available in each time frame. Often, zero messages were available while the figure in question is a high targeted politician in general political discourse. A reason for this could be found in the reason that Telegram does have another messaging structure than Twitter on Telegram, there is often a dialogue between (multiple) users, where a message could contain information that is a reaction to a previous message without directing it at a target figure. Furthermore, the possibilities and discretion of Telegram allows users to quickly make new groups which could contain more

concerning messages.

Future research should focus on the exploration of the distribution of messages on different social media platforms like Instagram, Tiktok, YouTube and Facebook. This could enhance the discovery of specific patterns in messages of radicalisation phases. One feature that this research did not look at was the retweets, reposts, and replies of the messages. This could be an interesting feature since it tells something about how messages are distributed in a certain social network and how these social networks are linked to each other. Emoticons provides a significant amount of sentiment in a sentence; therefore, future work could leave in emoticons for a more precise classification of different radicalisation phases.

# 6. Conclusion

This thesis aimed to develop a model that identifies and classifies radicalisation phases of messages in online political discourse. Furthermore, the analysis of distribution of messages around incidents are presented. The model achieved an accuracy of 98.91%, surpassing previous models by Parmar et al. (2018) and Kamath et al. (2018), though falling slightly short of the MCNN-LSTM model by Hasib et al. (2023). Precision is the critical measure for evaluating the model's capability in minimising false positives. High precision scores were obtained on label 0 (nothing) and label 4 (death threat), while label 1 (dehumanising), label 2 (demonising), and label 3 ((violent threat) obtained lower scores suggesting that there is room for improvement. The analysis of the case studies showed that public (political) events influence the volume and distribution of online political discourse and radicalised messages. Although no specific correlation between messages and incident dates was found, there was a clear correlation between significant events and peaks in data. This finding implies that the classifier is not redundant, and it remains a valuable tool for detecting and understanding the impact of significant events on online political discourse and the phases of radicalisation.

This research contributes to the field of Applied Data Science and Media Studies on several points. Firstly, the uniquely developed classifier shows significant potential in identifying non-radicalised messages and death threats. Furthermore, the use of synthetic data to handle imbalance in the data showed challenges and biases when paraphrasing sentences. The analysis of the case studies provides an understanding of how public events influence online political discourse and radicalisation phases. Lastly, a foundation is set for future research, by identifying limitations such as inconsistent labelling and optimisation constraints, to address the challenges and improve the performance of the model. This research may help foster the devel-

opment of an epistemic culture (Knorr-Cetina, 2009) consisting of media scholars and applied data scientists, who can communicate effectively and collaborate on further interdisciplinary research.

This research provides a foundation for advancing and exploring tools to detect and analyse online radicalisation and its different phases. To refine these tools and improve their capability to identify nuanced patterns in online discourse, constant efforts have to be made.

# Bibliography


Agarwal, S., & Sureka, A. (2015). Using KNN and SVM based one-class classifier for detecting online radicalization on twitter. In *Distributed computing and internet technology* (pp. 431–442). Springer International Publishing.

AIVD. (2023). Anti-institutioneel extremisme in nederland: Een ernstige dreiging voor de democratische rechtsorde? https://www.aivd.nl/documenten/publicaties/2023/05/25/anti-institutioneelextremisme-in-nederland-een-ernstige-dreiging-voor-de-democratische-rechtsorde

Bakker, Cornelisse, Mohamed, Schäfer, & Veerbeek. (2021). Van scherm naar straat. hoe sociale media-conversaties protest op straat mobiliseren. https://dataschool.nl/wp-content/uploads/sites/425/2021/03/20210318_Vanscherm-naar-straat.pdf

Bakker, de Graaf, Prins, Schäfer, & van der Weijden. (2023). Playing with fire: How the interplay between the dutch house of representatives and social media fuels rage. https://dataschool.nl/wp-content/uploads/sites/425/2024/04/Playing_With_Fire_research_report_29-04-24.pdf

Benotti, L., & Blackburn, P. (2022). Ethics consideration sections in natural language processing papers. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Berjawi, O., Fenza, G., & Loia, V. (2023). A comprehensive survey of detection and prevention approaches for online radicalization: Identifying gaps and future directions. *IEEE Access*, *11*, 120463–120491.

Borum, R. (2011). Radicalization into violent extremism i: A review of social science theories. *J. Strat. Secur.*, *4*(4), 7–36.

Bott, C., Leetz, J., & Kirkpatrick, S. A. (2006). *Radicalization: An overview and annotated bibliography of Open-Source literature*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale.

Das, M., Mathew, B., Saha, P., Goyal, P., & Mukherjee, A. (2020). Hate speech in online social media. *ACM SIGWEB Newsl.*, *2020*(Autumn), 1–8.

de Graaf, B. A., & van den Bos, K. (2021). Religious radicalization: Social appraisals and finding radical redemption in extreme beliefs. *Curr. Opin. Psychol.*, *40*, 56–60.

de Vries, W., Wieling, M., & Nissim, M. (2023). DUMB: A benchmark for smart evaluation of dutch models.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Proceedings of the 2019 conference of the north. *Proceedings of the 2019 Conference of the North*.

Dhyani, B. (2021). Transfer learning in natural language processing: A survey. *Philipp. Stat. (Quezon City)*, *70*(1), 303–311.

Erdogan, H. (2023). Dark side of the web in the context of online radicalization. *J. Islam. Thought Civiliz.*, *13*(1).

Gerritsen, J. B. A. (2021, August). *Tactvol contactloos onderzoek : Handleiding voor onderzoekers die de AVG respecteren*. Utrecht University.

Glazkova, A. (2008). *A comparison of synthetic oversampling methods for multi-class text classification. ArXiv, abs*.

Hasib, K. M., Azam, S., Karim, A., Marouf, A. A., Shamrat, F. M. J. M., Montaha, S., Yeo, K. C., Jonkman, M., Alhajj, R., & Rokne, J. G. (2023). MCNN-LSTM: Combining CNN and LSTM to classify multi-class text in imbalanced news data. *IEEE Access*, *11*, 93048–93063.

Johnston, F. R., Boyland, J. E., Meadows, M., & Shale, E. (1999). Some properties of a simple moving average when applied to forecasting a time series. *J. Oper. Res. Soc.*, *50*(12), 1267–1271.

Kamath, C. N., Bukhari, S. S., & Dengel, A. (2018). Comparative study between traditional machine learning and deep learning approaches for text classification. *Proceedings of the ACM Symposium on Document Engineering 2018*.

Knorr-Cetina, K. (2009, July). *Epistemic cultures*. Harvard University Press.

Kranenburg. (2022). Monitor integriteit veiligheid. io research 2022/118. https://open.overheid.nl/documenten/ronl90215dba1d2cbd7fac64bdd0bccd712ffea9e520/pdf

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach.

Lygre, R. B., Eid, J., Larsson, G., & Ranstorp, M. (2011). Terrorism as a process: A critical review of moghaddam's "staircase to terrorism". *Scand. J. Psychol.*, *52*(6), 609–616.

McCurdy, E. (2021). Radicalizing online. *muse*, *5*(1).

McKinney, W. (2010). Data structures for statistical computing in python (S. van der Walt & J. Millman, Eds.). *Proceedings of the 9th Python in Science Conference*, 56–61. https://conference.scipy.org/proceedings/scipy2010/mckinney.html

McMilin, E. (2022). Selection bias induced spurious correlations in large language models.

Moghaddam, F. M. (2005). The staircase to terrorism: A psychological exploration. *Am. Psychol.*, *60*(2), 161–169.

Mujahid, M., Kına, E., Rustam, F., Villar, M. G., Alvarado, E. S., De La Torre Diez, I., & Ashraf, I. (2024). Data oversampling and imbalanced datasets: An investigation of performance for machine learning and feature engineering. *J. Big Data*, *11*(1).

Mukherjee, S., Ojha, A., & Dusek, O. (2023). UFAL-ULD at BLP-2023 task 1: Violence detection in bangla text. *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*.

Muxel, A. (2020). Political radicalism among the younger generations. *Youth Glob.*, *2*(2), 123–136.

Nabilah Risky, N. R. (2023). The role of social media in the spread of religious extremism thoughts and their impact on social pathology. *J. Intellect. Sufism. Res.*, *6*(1), 29–34.

NCTV. (2023).

Nguyen, T. A., Zhurkina, O., Bui, T. C., & Sokolovskiy, K. (2022). Social media and political communication: Studying the interactive component. *J. Ethn. Cult. Stud.*, *9*(4), 187–200.

Nieuwsradio, B. (2023). Sigrid kaag stopt als leider d66 om bedreigingen, wil dat 'niet nogmaals vragen' van gezin [Accessed: 2024-07-25]. https://www.bnr.nl/nieuws/politiek/10518734/kaag-stopt-als-d66-leider#:~:text=23%2010%3A25-,Sigrid%20Kaag%20stopt%20als%20leider%20D66%20om%20bedreigingen%2C%20wil%20dat,niet%20nogmaals%20vragen'%20van%20gezin&text=D66%2Dpartijleider%20Sigrid%20Kaag%20stelt,minister%20tijdens%20een%20persconferentie%20zojuist.

Paraphrase [Accessed: 2024-7-25]. (2024).

Parmar, P. S., Biju, P. K., Shankar, M., & Kadiresan, N. (2018). Multiclass text classification and analytics for improving customer support response through different classifiers. *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*.

Peeters, S., & Willaert, T. (2022). Telegram and digital methods. *M/C*, *25*(1).

Perliger, A., & Pedahzur, A. (2016). Counter cultures, group dynamics and religious terrorism. *Polit. Stud.*, *64*(2), 297–314.

Rajendran, A., Sahithi, V. S., Gupta, C., Yadav, M., Ahirrao, S., Kotecha, K., Gaikwad, M., Abraham, A., Ahmed, N., & Alhammad, S. M. (2022). Detecting extremism on twitter during u.s. capitol riot using deep learning techniques. *IEEE Access*, *10*, 133052–133077.

Simon, M., Welbers, K., C. Kroon, A., & Trilling, D. (2022). Linked in the dark: A network approach to understanding information flows within the dutch telegramsphere. *Inf. Commun. Soc.*, 1–25.

Stafford, P., Campbell, C., & Lindén, M. (2019). A critical review of empirical evidence on radicalisation. In *Social identity theory perspective*.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks.

Vader, C. (2024, February). Gebruik van social media in 2024 groeit gestaag door [Accessed: 2024-7-25].

van Dijck, J., Poell, T., & de Waal, M. (2018). The platform society.

van Nederland, U. (2018). Nederlandse gedragscode wetenschappelijke integriteit [Accessed: 2024-7-25]. https://www.universiteitenvannederland.nl/files/documenten/Nederlandse%20gedragsco%20de%20wetenschappelijke%20integriteit%202018.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need.

Wadhwa, P., & Bhatia, M. P. S. (2015). Measuring radicalization in online social networks using markov chains. *J. Appl. Secur. Res.*, *10*(1), 23–47.

Walther, S., & Mccoy, A. (2021). *US extremism on telegram: Fueling disinformation, conspiracy theories, and accelerationism. perspectives on terrorism*.

Wang, C., & Banko, M. (2021, June). Practical transformer-based multilingual text classification. In Y.-b. Kim, Y. Li, & O. Rambow (Eds.), *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies: Industry papers* (pp. 121–129). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-industry.16

Wieringa, M., van Geenen, D., Schafer, M. T., & Gorzeman, L. (2018). Political topic-communities and their framing practices in the dutch twittersphere. *Internet Pol. Rev.*, *7*(2).

Wikipedia contributors. (n.d.). Lijst van tweede kamerleden 2021-2023.

Witteveen, S., & Andrews, M. (2019). Paraphrasing with large language models.