Erasmus University Rotterdam

Can forests make protests safer?

Using emotional loadings, Latent Dirichlet Allocation and violence classification on tweets related to the BLM protests in 2020 and 2021 in the US to predict whether a protest will turn violent.

Erasmus School of Economics

Master Thesis - Data Science & Marketing Analytics

Author: Mateusz Daniel Zbaraski

Student number: 448731

Supervisor: Dr. Radek Karpienko

Second Assessor: Prof. Dr. Patrick Groenen

August 14, 2022

Abstract

Protests are common social events, often turning from peaceful to violent. It is a good medium for expressing societal concerns and discontent with existing political systems by parties who are oppressed or their voice is oftentimes unheard by governments. However, when protests turn violent, the credibility of cause for protests is diminished and societal support for cause decreases. Moreover, often lives are lost as well as emotional damage and value destruction occur. Given the data richness in social movements and behavioural patterns in social networks, this study aims at identifying features that distinguish violent from peaceful protests prior to one happening. The analysis is based on 334 protests in years 2020 and 2021 in the U.S. related to the Black Lives Matters movement and death of George Floyd. Social media source used was Twitter. There were approximately 20 000 tweets connected to each protest, from which emotional loadings, Latent Dirichlet Topics (LDA) and violent tweet classification were extracted. Final models were computed using Random Forest model and Boruta model was used to assess the importance of features. The results of a model on three groups of features provides staggering accuracy of more than 85%. From the analysis it was found that mostly LDA topics are the best predictors of violence, however, only in combination with emotional loadings and violent tweet classification good model fitness was attained. The best predictor of violence was topic 40 that was mainly focused around Breonna Taylor, criminality and cowardice and topics 1 related to the killing of George Floyd, protests and rioting, skin colour issues with focus in New York. This analysis proves it is possible to build an accurate violent protest predicting tool using social media data when focusing at the BLM protests in the past years. With very limited research on predicting violent protests as of now, analysis on wider range of protests and countries is needed to understand the general pattens and trends. With further development it might be possible to accurately predict violent unrests around the world in advance and take appropriate measures to diminish the negative outcomes of riots.

Keywords: protest, violence, BLM, random forest, riot identification

Table of Contents

1 Introduction

A protest is an event where people gather, frequently in large groups, to express their opinions on societal problems, show frustration, inform about certain injustices or share information (ADL, 2020). Such protests have been occurring regularly worldwide, increasing in multitude over time. On 02-01-2022, police had to disperse anti-vax and anti-lockdown protestants in Amsterdam (Wires, 2022), and in the previous year, on 20-11-2021, a significant protest in Rotterdam shook the city centre. In the latter case, police had to open fire, wounding seven people and having over 20 arrested. Besides multiple arrests and wounds, several police cars were burned, certain city centre areas were demolished, and one building was severely burnt. Rotterdam's mayor deemed this event "an orgy of violence" (BBC, 2021).

Protests have been occurring for many years, with over 230 significant anti-government manifestations in more than 110 countries since 2017 (Carnegie Endowment, 2022). In the recent decade, however, compared 2011 to 2018, the number of protests and riots more than doubled. The number of nonviolent demonstrations was at a level of roughly 500 a year in 2011, down to 310 in 2013 and increased to 900 in 2018. The exact rising trend is visible for riots (violent) increasing from 180 in 2011 to 350 in 2018 (Vision of Humanity, 2020).

Buchanan, Bui, and Patel (2020) show that on the 6th of July 2020, there was an intensification of BLM-related protests, with more than half a million people protesting in almost 550 locations simultaneously. They state that 15 to 26 million Americans took to the streets, meaning approximately 7% of the population participated over the following weeks after the George Floyd killing. Damage from related riots and civil disorder is believed to surpass the 1-billion-dollar mark (Kingson, 2020). The protests linked to the issue continue until now.

After that, the COVID pandemic happened, with the anti-vax protests turning violent too (van der Zwet et al., 2022). In the other part of the world, in Kazakhstan, over one week and two days, 208 citizens and 19 members of the security forces were killed, with over 9 900 arrested (Abdurasulov, 2022). The protest was sparked by the government's increased gas prices, fuelled by dissatisfaction with economic inequality and corruption. The event resulted in the resignation of prominent government figures and the resignation of the cabinet, massive looting in the capital Almaty, and restoring the price cap on gas prices.

Additionally, a bigger societal problem occurs when a peaceful protest turns into a violent riot. Simpson, Willer, and Feinberg (2018) note that violent protests affect the public perception of the event. In those

situations, the general public might view the protest group as 'less reasonable', decreasing both the identification with the group and social support for the cause.

As can be seen from history, outcomes vary a lot. Sometimes dissatisfactory and corrupt governments are over-throned. Other times, anti-discriminatory laws are established and sometimes - nothing changes. If a protest gets out of hand, it can result in massive value destruction in the area of public life, economic shrinking of certain areas followed by arrests and deaths and lower societal effects compared to peaceful events. Expressions of opinion, social problems and repression are necessary for a society to grow and evolve; however, are the costs necessary and unavoidable?

One source that can provide much understanding of the phenomenon is social media. In 2004, Myspace became the first-ever social media platform to achieve 1 million monthly active users (Ortiz-Ospina, 2019). Since then, many platforms have appeared, such as Facebook, Instagram, and Twitter. As of October 2021, roughly 4.48 billion people, accounting for 56.5% of the population, with up to 82% in North America, use social media. Every person has 8.4 social media accounts (Dean, et al., 2021). This enormous growth in users enabled further growth and transformation of platforms.

Moreover, people have become more dependent on using them to organise and plan events (Moretti, 2012). Since all social media interactions happen online, every piece of information is saved, providing data richness in many aspects of socio-economics. Social media became an effective tool for instant long-distance communication, sharing information on planned events and organising as a group.

In the last decades, computing power has been growing exponentially, as well as the storage and variety of data. This data availability led to a boom in the ability to analyse, visualise data and come up with actionable insights. As Wiederhold (2020) found and previously mentioned in the paper, social media plays a significant role in organising events and mobilising interested parties. It is possible to access that data, getting an overview of the tweet content, number of likes and retweets, the location of a tweet and who tweeted. By applying various text analytics methods, such as sentiment analysis or topic modelling, to the body of tweets, it is possible to extract even more information, some of which might be useful for predicting the violent inclinations of protest participants.

There has been an extensive body of research on societal effects of violence during protests, psychological analyses of the drives and machine learning (ML) solutions to classifications of tweets. (Mooijman et al., 2018) identified that one of the predictors of violence in protests is the degree to which people view a protest as a moral issue and perceived moral convergence. Ramakrishnan et al. (2014) built a tool successfully identifying probable protests and their location using social media four days in advance. Jiménez-Moya et al. (2015) found that low group identifiers are more likely to behave

violently during a protest as, counterintuitively, those more connected to a movement even though they are invested in the cause are often concerned about how the organization is perceived.

Many tools and methodologies identify with high precision where, when and what topic of a protest will most likely be, such as state of the art EMBERS system (Ramakrishnan et al., 2014). However, there is still very little research and machine learning methodologies that can precisely predict whether a protest will turn violent prior to an event. There are many capabilities countries could obtain by being able to identify probable violent protests, understanding the underlying issue and taking preventive measures rather than solely reacting. In my research, I intend to expand the knowledge on violent protest drivers and provide a cohesive base methodology for further development. Therefore, the research question sounds:

Which social media features are good predictors of violence outbursts during the 2020/2021 BLM protests in the U.S?

The research is divided into six parts. First, the literature review is provided. It is elaborated on how versatile and rich the social media data is, what protest-related methodologies are in use and what are the characteristics of a violent protest. It is the followed with description of the data sources and extraction, mainly from Twitter and ACLED database. The fourth part describes the technical process of wrangling the data to prepare it for a sound analysis followed by the tweet feature extraction procedures of emotional analysis, LDA topics and violent tweet classification. Methodology explains the machine learning methods applied to the datasets. The last two parts are the results that explain the findings and are followed by concluding remarks, limitation of the research and potential for further research.

2 Literature review

2.1 Introduction

The literature review focuses on explaining the theories and findings of machine learning applications to protests using Twitter data. First, the versatility of social media is shown. A discussion of protest identification tools then follows. The next part outlines the researched causes of violent outbursts during protests. Last parts focus on the natural language processing capabilities and meaningfulness of variables that can be extracted from text as well as practical measures applicable to prevent or de-escalate violence.

2.2 Twitter as a protest tool

The social media user count has constantly been growing in the past years. As of Jan 2022, there are approximately 4.62 billion social media users worldwide, with a 10% annual increase in the headcount (GWI, 2022). Twitter itself had 206 million monetizable users that were active worldwide (Statista, 2022). Even though the users' numbers come short compared to giants such as Facebook, Instagram or Tik-Tok, Twitter is heavily used by journalists, celebrities and politicians, having 83% of global leaders as frequent users (Davies, 2022). It is a great political debate platform as well as an organizational tool.

To further this statement, Tufekci and Wilson (2012) found that in the Tahrir Square protest in 2011, part of the Arab Spring, social media were one of the main drivers for individuals in making decisions on whether one will or will not attend the protest. People learned about the event usually through interpersonal communication on Facebook, phone calls or in-person conversations – something the government failed to incorporate in their response to the event effectively.

Breuer, Landman, and Farquhar (2015) argued that social media is not only an event-organization tool but also a platform for cyber activism. The Tunisian uprising of 2010-2011 was heavily influenced by the growing discontent of the Ben Ali regime. This was mainly enabled by the digitalization in the county that enabled citizens to learn about the scale of inequality within the country and abroad. The digital activists showcasing the narratives and government abuse further thickened the discontent. Social media played a crucial role in circumnavigating censorship to inform the people. Moreover, since much data, both textual and visual, was shared, citizens could calculate their "risk threshold" and gave a base to the creation of a national collective against the regime.

However, not all economies are ruled by regimes nor do they have heavy censorship. In terms of protest organization and expansion, social media is a tool of great use. As Mundt, Ross, and Burnett (2018) found, based on the Black Lives Matter (BLM) movement, social media is of great help in strengthening the position, building connections and mobilizing society, referred to as the "scaling up" process. They identified that there are multiple BLM groups that, through online presence, connect and create a feeling of participation in a more significant cause, building a greater sense of community. Shaw (2013) formulated that collective identity is critical to strengthening engagement and the position of movements. By often having very open membership (no log-in necessity or no paywalls), such groups can quickly gather followers and rapidly incentivize members to action. The size of amplification, reach, and speed of mobilization through social media are well depicted by the response to the Alton Sterling and Philando Castile shooting case on July 5th 2016. Within only hours after a black man was shot and

killed in his car by a police officer, protests erupted. One group administrator said in an interview that 'in a short amount of time, the event quickly grew to over 1500 people who were committed to attending' (Mundt et al., 2018).

This displays the vastness of emerging possibilities and rapid response to disturbing societal events using social media. As protests frequently occur semi-spontaneously and sometimes disturb the social life and leads to violence and destruction, it is valuable to identify them as soon as possible.

2.3 Protest identification and forecasting

The topic of protest forecasting has been research by many. With the ongoing rapid expansion of the digital world there is a lot of variability to it. Korkmaz et al. (2015) focused their research on Latin America. They used GSR (Gold Standard Reporting) combined with social media (Twitter), blogs, news (obtained from LANIC) and country stability indicators such as currency. By applying logistic regressions with lasso regularization, they have obtained a precision of between 68% and 95% compared to the GSR in predicting civil unrest.

Their research was then widened by Muthiah et al. (2015). They are the creators of one of the most prominent protest predicting systems, namely EMBERS – the Early Model Based Event Recognition using Surrogates. As of writing their paper, they approximated 75% of all protests are planned beforehand. They use the elements of phrase learning, probabilistic soft logic and time normalisation to find the upcoming protests. The tool they have proposed use a combination of data from Twitter, Facebook, RSS (news and feed) and mailing lists. By applying this approach, they were able to predict significant unrests 4.08 days in advance. This system is in operation since November 2012 and predicts civil unrest in 10 Latin American countries. It has successfully predicted the Brazilian unrest in June 2013 and the Venezuelan protests in February 2014 (Ramakrishnan et al., 2014).

Bahrami et al. (2018) have focused on the protests surrounding Mr. Trump prior to his election in 2016. They based their research solely on Twitter. By finding relevant hashtags related to the elections and applying machine learning models they have obtained a 75% - 100% accuracy in identifying future protests. Their conclusion is that Twitter solely can be a powerful prediction tool for when a protest will occur. They have applied their models on clustered user posts.

2.4 Violent behaviour prediction

In terms of violent behaviour prediction during protests, the ML research is rather limited. Anastasopoulos and Williams (2019) are thought to have created an approach for measuring violent and peaceful protests based on social media data. Their approach is heavily based on van Deth (2014) where they select only political tweets and then classify them into four groups based on 7 rules. The rules used for classifications answer questions such as whether we are dealing with behaviour, is the activity voluntary, is the activity done by citizens or is the activity aimed at solving community problem. Further, by answering a question on whether tweet relates to one or more people or if it is violent or not, all the tweets are classified into clusters singular peace, singular violence, collective peace and collective violence. Looking at the case of the Ferguson protests around 11-08-2014 they have found that prior to this day of the biggest unrest, total action types was mainly singular peace. At the day of the protest all the classified groups occurred almost 10 times more often, with the collective force being the most visible. Hence, the first hypothesis is:

H1: The peaceful/violent singular/collective a priori Tweet classification is indicative if the protest will turn violent or stay peaceful.

A different approach to protest violence is presented in Mooijman et al. (2018). The underlying theory is that the emergence of violence during protests can be perceived as a function of individual's moralization of a cause and the level to which other people from their social network moralize the cause. Looking at the 2015 Baltimore protests, it is found that the degree of moralizing in social media was higher on the days when violence occurred. Moreover, it was found that on an hourly level the moralization predicted well the future protest arrests. By running additional experiments, the bottom line is that people are most affected by this phenomenon when they believe that the people around them share their values.

Jiménez-Moya et al. (2015) also aspired at explaining what causes protests violence. Before, it was found that there is a positive relationship between group identification and collective behaviour (Sturmer and Simon, 2004). Jiménez-Moya et al. in their analysis, however, proved that as collective actions are important towards a social change, people of the disadvantaged groups are often aware of the risks that come with active participation. They found many pros and cons for whether high or low identification correlate with violence, hence they have done two studies in that direction. The result is that radical behaviour is most seen among low identifiers, often when a social disadvantage of a group is perceived as legitimate, since they have a nothing-to-lose attitude.

Bollen, Mao, and Pepe (2011) have focused on almost 10mln tweets published in the second half of 2008. From there, they have used psychometric instruments to extract emotional states from the tweets.

By using the Profile of Mood States (POMS) and comparing to a timeline of selected events they argue that 'social, political, cultural and economic events are correlated with significant (…) fluctuations of public mood levels along a range of different mood dimensions'. The conclusion is that it is efficient to apply non-ML techniques to extract such features form text limited to just 280 signs. They argue that applying empirical psychometric methods can be as effective as any ML based methods in understanding public sentiments.

Following this way of thought, Ives and Lewis (2020) identified multiple features connected to violence protest. Their main finding was that violent escalations are most common when protests are preceded by repression. They argue that violence is rather driven by repression in short before a protest, as in all of their 5 models time since repression was negatively correlated and significant with likelihood of violence outburst. The other important insight is that violence is more likely to occur if the protest is unorganized. They theorize that when a strong and structured hierarchy is missing it is of higher difficulty for protest organizers to control the people. As previously found by researchers, they confirm that electoral/political protests can be more tense whereas economy/jobs protests do not seem to become violent too often.

Looking at it from an emotional perspective, it is thought that contempt is the main emotion related to violent outbursts, not only at protests (Becker and Tausch, 2015). They provide rationale that contempt leads to lower chance of reconciliation. It is possible that in the presence of an injustice or a threat, contempt can result in hostile reactions. This is further confirmed by Tausch et al. (2011). They have identified that anger was strongly related to normative action but unrelated to nonnormative action and that contempt was either unrelated or negatively related to normative action but significantly positively predicted nonnormative action. Therefore, emotional states of protest participants or people discussing protest related content can be indicative of anger issues at hand. Hence, the second hypothesis sounds:

H2: Emotional features extracted from Tweets are meaningful predictors in identifying if a protest will stay peaceful or turn violent

To further the understanding of what are the sentiments before an event it is interesting to research topics discussed and their linkages to both peaceful and violent protests. Resnik et al. (2015) have done their research on exploring the links between Latent Dirichlet Allocation (LDA) classified topics and depression-related Tweets. Their conclusion is that LDA and LDA-like models are significantly outperforming simpler models in identifying a latent structure. Their second finding is that by aggregating tweets weekly, the precision at R=0.5 increased to 74% and at R=0.75 increased to 62%. Therefore, LDA was applied to twitter data to identify the topics related to protests. Moreover, for the

purpose of identifying latent structure LDA is much faster than Non-negative Matrix Factorization (NMF) and does not have the issue of negative factors and loading as in PCA.

H3: LDA topics extracted from Tweets are meaningful predictors in identifying if a protest will stay peaceful or turn violent

However, it is expected to attain the best predictive power when using a combination of all three feature groups to predict the violence potential of the protest. Hence, the four hypothesis sounds:

H4: The three groups of Tweets extracted features are performing better in classifying if a protest will stay peaceful or turn violent

For the purpose of answering hypotheses 1-4, two-class models were used. In the full dataset there were four classes in the dependent variable, namely *peaceful protest, violent demonstration, excessive force against protesters* and *protest with intervention.* The latter three groups indicate there was violence during an event, however, caused by protesters in some cases and sometimes by the police or the army. Hence:

H5: The three groups of Tweets extracted features are performing better in classifying if a protest will be peaceful, violent, with excessive force against protesters or protest with interventions

2.5 Preventing violence in protests

Ability to predict violent protests can greatly increase the understanding of this social phenomena and give time for authorities to prepare. However, having a precise prediction and not taking correct actions, the predictive effort will be futile. Nassauer (2019) states there are numerous activities essential for protests to stay peaceful. During a protest there has to be a flow of communication between protest participants and police, efforts to stop harmful rumours and setting hard territorial boundaries. By being able to estimate the likelihood of violence at a protest, understand better the participants and establishing appropriate communication is essential. Social media is one of the information sources that can both help predict unrests and provide means to reach out to the organisers.

3 Data collection

3.1 Computing

To have utmost precision, vast amount of data was collected and analysed. To store gigabytes of tweets and be able to run models, cloud computing was utilised. Amazon Web Services (AWS) were chosen. An RDS MySQL database was set up that was then connected to a cloud R and Python instances. By doing so, the computational power was expanded from 16GB of a desktop to up to 128GB of RAM and an organized storage expanded to more than 200GB.

3.2 ACLED Protest Data

The protest data was derived from The Armed Conflict Location & Event Data Project (ACLED). This project collects real-time information on variables such as location, dates, participants, fatalities of all reported political unrests and protests globally. The data is highly precise as over a month between 29-04-2022 to 27-05-2022 there were 9 745 distinct events recorded and more than 10 866 fatalities identified. The data set is updated weekly and was downloaded using the Data Export Tool.

For the purpose of the research the data of two years was selected, from 01-01-2020 until 31-12-2021. Additionally, since the BLM protests are researched, the most prominent examples were visible in the U.S. and Canada. However, Canada lacked protest data for the entirety of 2020, hence it was dropped. By applying such filters, the ACLED data was narrowed down to 35 736 observations with 31 variables. Events such as the 25-05-2020 George Floyd killing were incorporated as well as the following major unrest across the entirety of the US.

Variable	Description	Value
Event Date	Date of a protest in a format year-month-day	Between 01-01-2020 and 31-12-2021
Event Type	Describes the type of event	1. Protest 2. Riot
Sub Event Type	Describes the type of protest	1. Violent demonstration 2. Peaceful protest 3. Protest with interventions 4. Excessive force against protesters
Associated Actor	The organizer or associated party for an event	Non-standardized variable; Examples: ‘BLM’, ‘Health Workers’, ‘NAACP’, ‘Boogaloo Boys’
Country	Country of protest	United States
State	State of the protest	Standardized state name; Examples: ‘California’, ‘Texas’, ‘Florida’
Location	Exact location of event happening, either county or	Standardized location name; Examples: ‘Greene’, ‘Pasquotank’, ‘New York’, ‘Fulton’

***Table 1**: Overview of the ACLED protest dataset most important metrics*

In this dataset the first relevant grouping category is *event type*. The variable describes the general type of an event, such as a protest, riot or violence against civilians. In the US during the chosen years there were 16 646 peaceful protests, 472 riots and 82 violent actions against civilians. Figure 2 shows how the events unfolded in a timely manner. It can be seen that the highest spike in the number of protests occurred around the end of May 2020 with some days having more than 600 separate protests occurring. The data table was then filtered to exclude violence against civilians as those events aren't related to protests (30th December 2021 healthcare workers attack in Tustin, Orange; 29th November 2021 police officers shot a burglar on a scooter in Tucson).

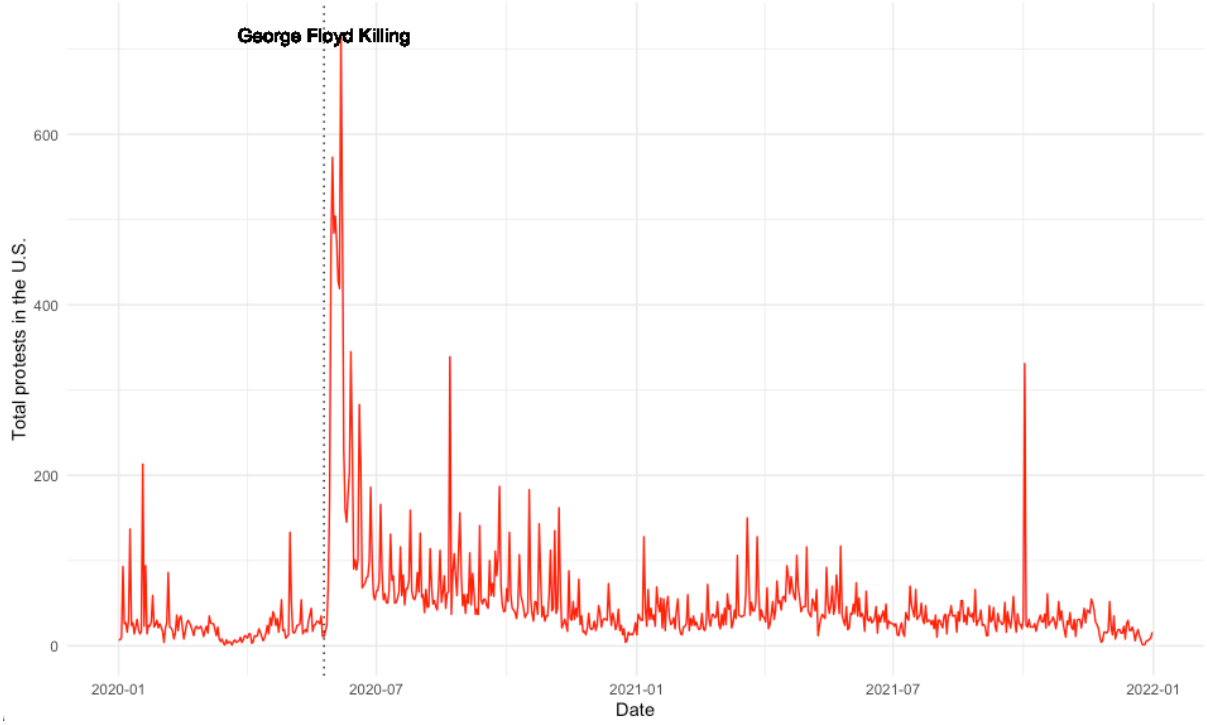


***Figure 1**: Total number of daily protests across the U.S. between 01/01/2020 and 31/12/2021*

A big increase in protests can be seen on the 25th May 2020 when George Floyd was killed. The amplitude diminished in the following months, however there were between 100 to 200 protests per day in the days right after the event. The data was further filtered to contain only BLM related (Black Lives Matter) events. A variable *actor1* describes the parties that organised the event and those that motivated people to action. Therefore, the *actor1* variable was filtered for the existence of 'BLM' string, narrowing the number of events down to 5901.

The other important variable is the *sub event type* that further specifies the type of an event. The groups in this variable are Excessive force against protesters (59), Mob violence (39), Peaceful protest (5233), Protest with interventions (270) and Violent demonstrations (339). Mob violence isn't associated with BLM protests as those are attacks and burglaries targeted at specific civilians. Hence, mob violence was

filtered out from the dataset. In Figure 1 can be seen a spike in protesting activity after the George Floyd killing. Even though the spike is vastly overshadowed by an increase in peaceful protests to a level of 330, around the nearest days following the event there were between 50 to 10 violent protests per day.

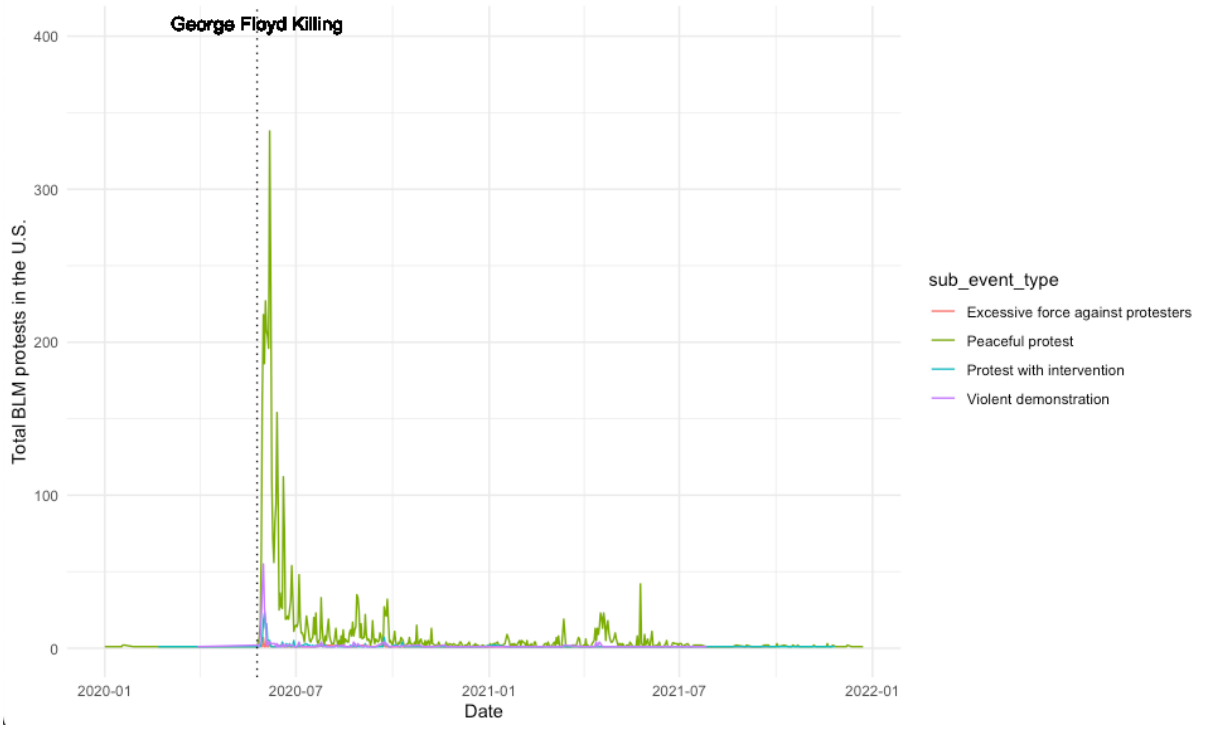


***Figure 2**: Total daily BLM protests in the U.S. between 01/01/2020 and 31/12/2021 per sub event category with an indication of date of George Floyd killing on 25/05/2020*

Moreover, since a few days before and including the day of a protest were scraped, downloading at a pace of 100 000 tweets per hour, it was decided that downloading entirety of tweets for almost 6 thousand events would take months and terabytes. Additionally, since there are only 59 occasions with excessive force against protesters, data balancing was performed. Therefore, the final ACLED data is comprised of 59 Excessive force, 100 peaceful protests, 100 protests with intervention and 100 violent demonstrations. To add to that, for some of the events there were only a few tweets linked to the protest, hence it was assumed there must be at least 10 tweets per protest for one to be included in the final dataset. After applying fitters there were 334 protests left.

State is a variable describing the second most general land division after country. As can be seen in Figure 4.1, majority of protest captured in the sampled dataset happened in California (52), New York (26) and Oregon (24). The dataset contains events from 46 distinct states, however for 7 of them there is only one event recorded. As seen in Figure 3.2, the within state division between sub event types is fairly balanced too with majority of violent protests happening in California and Oregon.

In Figure 3.3 a division between the U.S. State is made against the actor organising the event. As can be seen, vast majority of protests was organised or associated solely to the BLM movement. The other rectangles represent other organizations/groups that collaborated with the core BLM movement. Civilians, journalists, students, LGBT or even police groups were involved in organizing those protests in collaboration with BLM.

In the final ACLED dataset there were 334 protests recorded, yet there were four classes of protests. As the goal of the research of the paper is to assess whether it is possible to successfully predict whether a protest will be violent, the data was filtered to contain instances of only *violent demonstration* and *peaceful protest.* This further decreased the number of events to 184. The four-class dataset was used, however, to assess whether the same features can successfully distinguish between *violent demonstration, excessive force against protesters, protest with intervention* and *peaceful protest.*

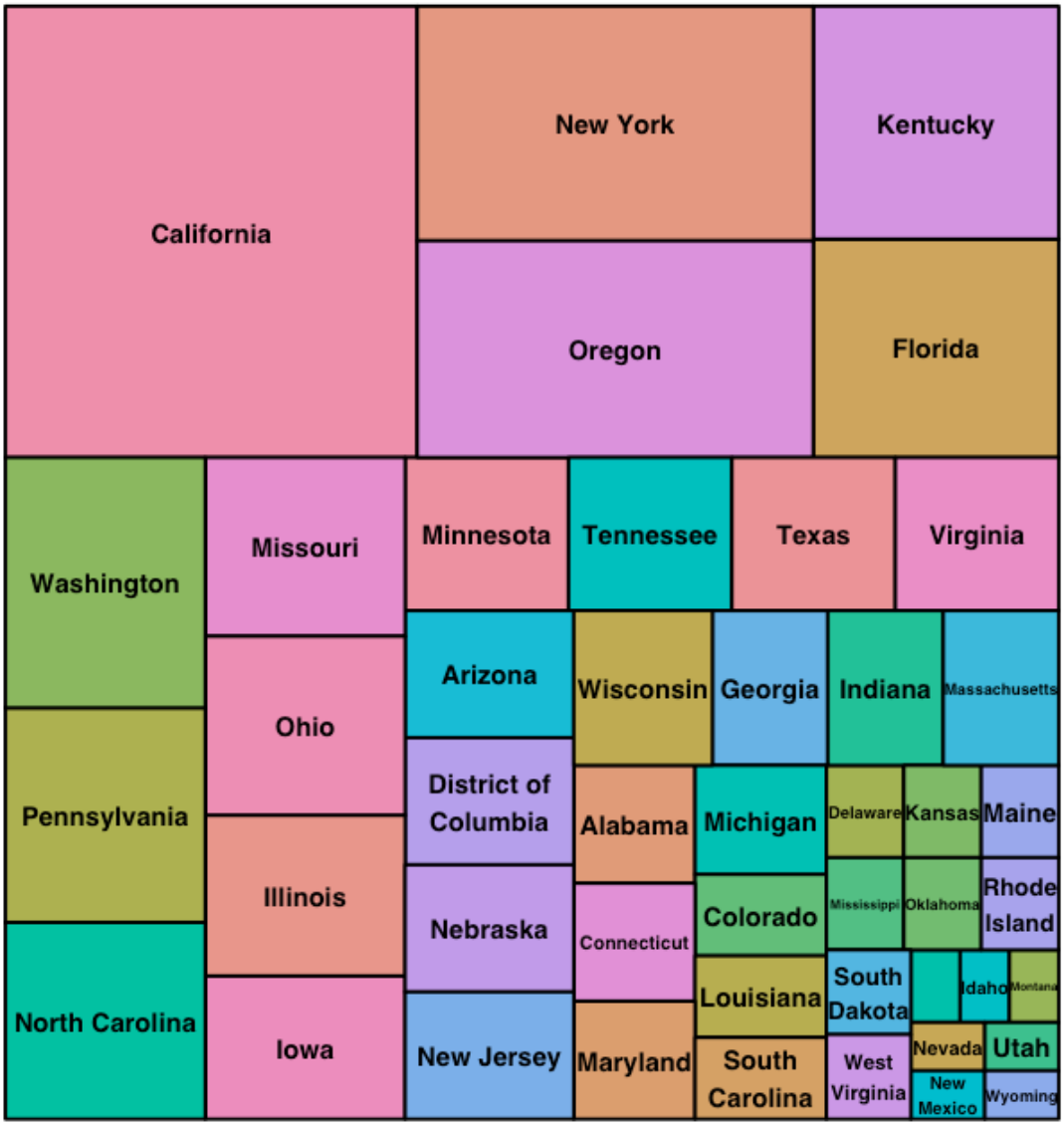


Figure 3.1*: Comparative number of U.S. protests in ACLED dataset. 46 distinct states were identified*

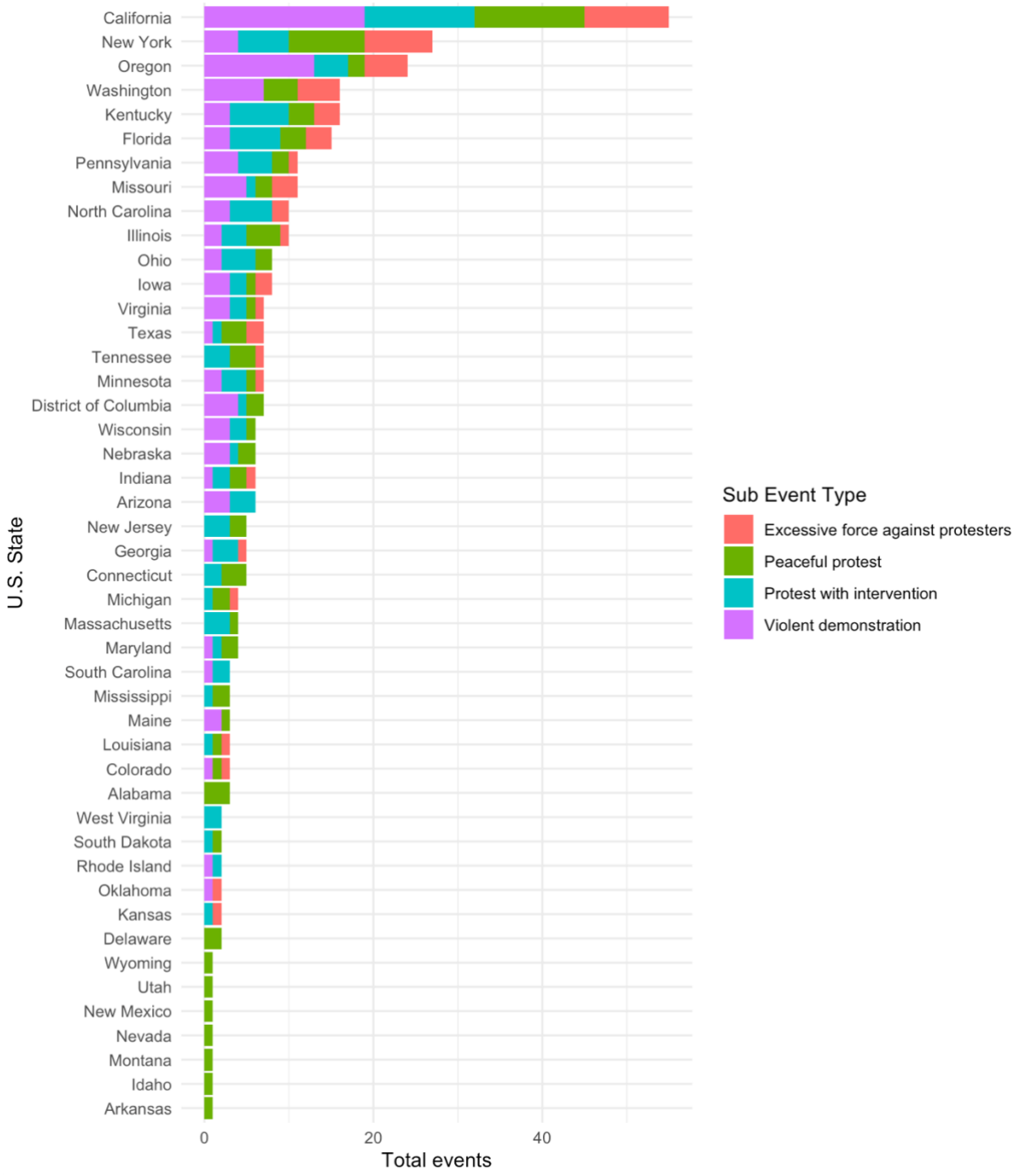


***Figure 3.2**: Count of protests in the U.S. divided per sub event category from the ACLED dataset.*

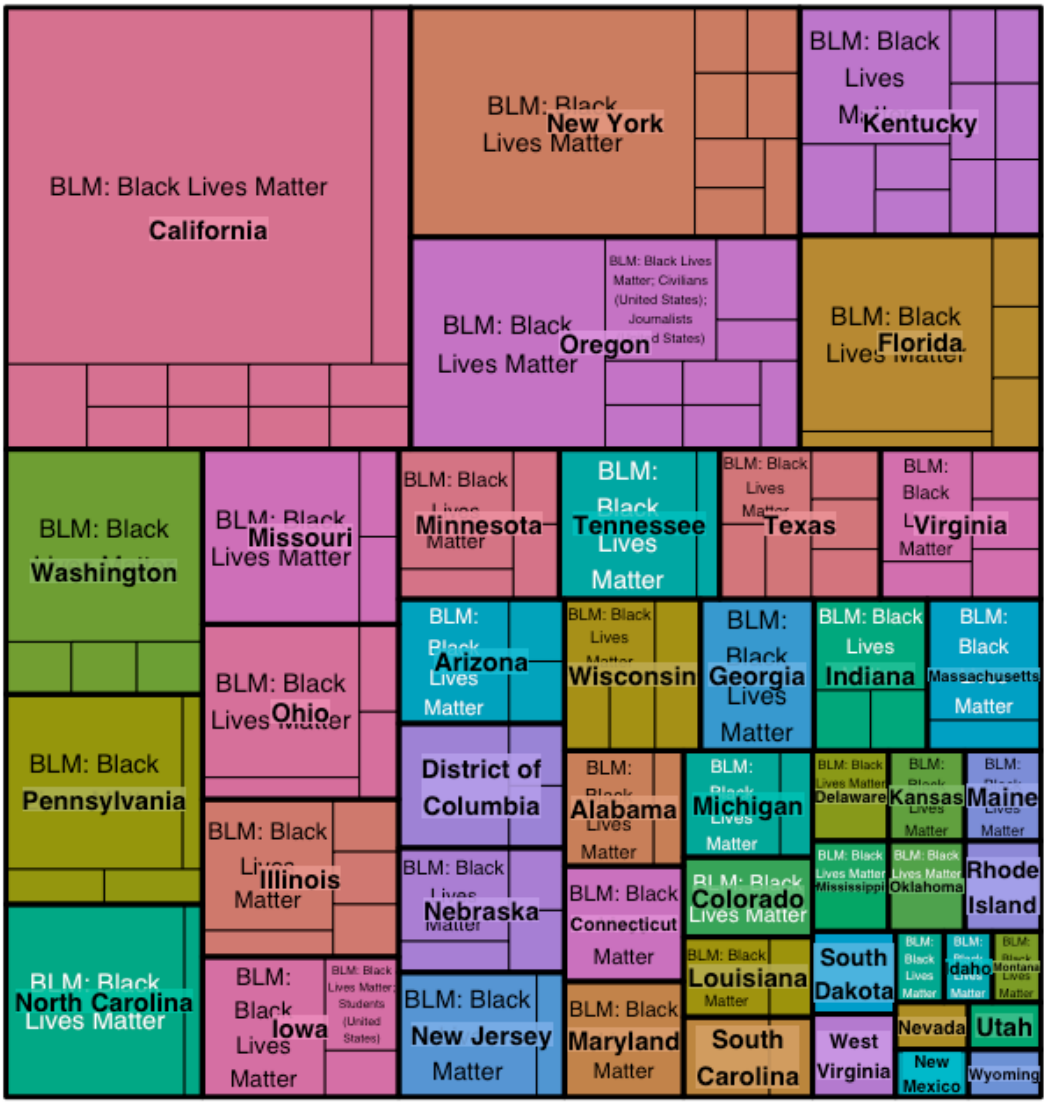


***Figure 3.3**: Division of main protest organizers across states in the U.S. from ACLED dataset*

3.3 Twitter data

The second data source used in the research is Tweets. Tweets were downloaded for the period between 01/02/2020 and 31/12/2021 for each of the protest identified in the ACLED dataset of 334 events in the U.S. in the four-class dataset and 184 in the two-class dataset. These datasets are respectively referred to as 4-class df and 2-class df in the remainder of the paper.

The best tool identified for the task of downloading numerous tweets was Snscrape – a python tool for downloading all kinds of social media data. Snscrape is superior to the official Twitter API as there is no limit of 3200 tweets per query or a one-week lookback. By combining a Python script with R looping and connecting it to an AWS RDS a download speed of 100 000 tweets per hour was obtained. In a

query to download the tweets it was possible to scrape tweets that contain a specified hashtag, word or belong to a specific user.

There are numerous variables possible to download from Twitter such as the *Tweet id, the date* and *time of posting the tweet, user name, URL, textual content, retweet count, like count, quote count, user location, verified flag, follows, friends* and *language*. For the purpose of the research only relevant variables were selected such as *Tweet id, date, textual content, language, geotagging* and *user location*.

Variable	Description	Value
Tweet ID	Unique string of numbers of each tweet	18 – 22 randomly generated digits
Date	Exact date and timestamp of when a tweet was posted	2021-12-31 23:59:59
Textual content	Body of a tweet posted by an individual	limited to 280 characters or Unicode glyphs (30 – 45 words)
Language	Automatically classified language based on textual content of a tweet	en, pl, fr, gr
Geotagging	GPS coordinates of where the tweet was posted	Lat: 44; 6; 43.82 Long: 79; 14.86
User Location	Location of the Twitter profile user, specified by each user	Non-standardized location

***Table 2**: Overview of the Twitter most important variables*

***Figure 4**: Personal information of an account – official Black Lives Matter profile. Looking at the textual information from top to bottom, there is the official name of user followed by unique id name, description of the profile. In the second to last line at the left there is the location, www link to an official web page to the right,*

followed by the time of joining Twitter. Last two values show how many profiles are followed by BLM and how many followers of BLM there are.

To link all the tweets to specific protests, the name of town or location where a protest was happening was passed as a keyword to scrape the tweets. This means that the body of tweets was scraped based on whether a location is mentioned in the body of tweet. For example, if a protest occurred in Los Angeles, all the tweets that mentioned 'Los Angeles' were scraped. This approach proved to be more precise as less than 1% of tweets are geotagged (Sloan and Morgan, 2015) and the *user location* variables is non standardized, many times meaningless.

To optimize the data download process, 3 days prior to the protest and the day of the protest were downloaded. Since for some cities there are hundreds of millions of tweets produced each day, a daily limit of 50 000 per day was set. In total 6 881 173 tweets were downloaded for 334 events, averaging 20 602.3 tweets per protest.

Even though almost 7 million tweets were scraped many were irrelevant to protests. Some of the locations where protests were happened were Travis, Kings or Jefferson. There are many tweets about King LeBron James, Jefferson airplane (band), or Travis Scott (a singer). Hence, the data was filtered for keywords specific to BLM protests. Badaoui (2020) has focused in his paper on identifying the corpus via web-scraping. He created a term frequency distribution table in the most commonly occurring tweet keywords as well as hashtags. Those are words such as black, police, protest or matter, and hashtags #georgefloyd, #blackouttuesday or #nojusticenopeace (complete list in Appendix A). All the collected tweets were then filtered for the existence of one of the specified words or hashtags resulting in a decrease of tweet count from 6 881 173 to 1 374 744.

Once filtered for the existence of protest-related keywords, tweets were further cleaned. The first step was to remove all the user mentions starting with '@'. Some of the tweets had 20+ mentions with little to no text to it.

From the tweet body the URLs were removed with qdapRegex. Additionally, all the double spaces were squished as well as the text was stemmed resulting in more normalized data. Furthermore, all stop words except for 'no', 'not' and 'never' were removed since they occur oftentimes and do not differentiate the tweets much. Additionally, all the punctuations, hashtags beginning (#) and time and dates were standardized to the same format. Moreover, some signs or figures that look 'regular' on Twitter look different when downloaded through API. One example can be '&', that looks normal on the site, but is seen in the API data s '&amp'. Some of those non-readable texts were removed, however, emojis (that

are displayed in R as a string of mixed letters and digits) can be expressions of emotions – hence they weren't removed. To have the data prepared for emotional analysis and LDA, all the punctuations and special signs were removed, together with de-capitalizing all the words.

3.4 Violent tweet classification dataset

Anastasopoulos and Williams (2019) have compiled a dataset of hand-classified tweets for identifying violent and peaceful tweets based on tweet body. The second grouping axis they chose is whether the tweet is related to one person or more, distinguishing between singular and collective tweets.

Their dataset comprises of 22 625 tweets classified into each of these four groups. In the dataset there are four groups and the tweet body. Those categories are *singular peace, collective peace, singular force* and *collective force*. The data is very unbalanced, as a vast majority of tweets does not belong to any group. In total, there are only 2 596 observations with a category assigned to them and 20 029 tweets with no category. There are 1 823 *collective peace*, 381 *singular force*, 474 *collective peace* and 795 *collective force* tweets classified as seen in Table 3. To attain better classification precision the data was slightly balanced to combine the 2 596 classified tweets and 5000 0-classified tweets. The data was then cleaned, in the same way as in the 3.3 part – removed links, removed mentions, stemmed, removed stop words.

Variable	Frequency	Mean
Tweet ID	7596	0.2400
Textual Content	7596	-
Singular Peace	1823	0.2400
Singular Force	321	0.0462
Collective Peace	474	0.0624
Collective force	795	0.1047

***Table 3**: Overview of the violent tweet classification dataset after class balancing.*

Violence tweet classification	Example of a tweet
Singular peace	I guess if you're not white, you're guilty until proven innocent. #Ferguson
Singular force	It's barley seven in the morning and I legit already dodged my death bullet.
Collective peace	4.5 minutes of silence at the intersection of 14th and Broadway. #FergusonVerdict http://t.co/fYQfE6XSod

Collective force	Scattering people after loud pops and maybe tear gas #Oakland #Ferguson

***Table 4**: Examples of four tweets in the twitter violence classification dataset, each from different classification category.*

3.5 Aggregating the Tweet dataset

In total there are 334 protests for which Tweets were scraped. The issue at hand is to predict what protests are thought to turn violent, rather than understand which tweets are related to violence during protests. Therefore, all the tweet data was grouped by the protest location and date, as there are more than one protests at locations.

First, emotion feature group was summarised. Since there are both emotional variables and flags, both variable groups were summed per event and date. Moreover, since there are different number of tweets for each event, the sole magnitude of emotional count loses much of its meaning when compared between protests. Therefore, the sum of emotional counts and flags was divided by the number of total observations to identify the emotional loadings of each protest.

Next, the violence classification was aggregated. All the four groups were summed and also divided by the number of observations. The third aggregation was done on the LDA topics. All of the 40 topics were summed and divided by the number of tweets identified per protest. All these groups are referred to in the remainder of the paper as total emotional count, emotional percentages, total violence count, violence percentages and total LDA loadings.

The dataset however, contains only 184 final observations. As there are two classes of protest type, peaceful and violent protests, the final prediction was based on fairly few observations. Hence, it was decided to create a second dataset where instead of only two classes there were four. This was attained using the full dataset with classes *excessive force against protesters* and *protest with intervention*. Hence, two dataset of size 184 and 334 observations respectively were used.

4 Data processing and feature extraction

After data wrangling was completed, the next step was to derive information from the body of tweets. There were three feature extraction methods applied to the Tweeter 2020-2020 data applied: LDA, emotional analysis and violence tweet prediction.

4.1 Emotional analysis

The second type of features extracted were emotions. As Bollen, Mao, and Pepe (2011) and Becker and Tausch (2015) found, emotional states of people affect their behaviour. Therefore, NRC sentiment were extracted. The method uses National Research Council Canada lexicon. The lexicon contains a dictionary of 14 000+ words with emotional sentiment. By applying this approach, it is possible to understand whether one or more of the eight emotions are present in a tweet. This resulted in identifying the counts of 8 emotions in each tweet: anger, anticipation, disgust, fear, joy, sadness, surprise and trust. Additionally, a count positive and negative emotions was found.

Just like Becker and Tausch found, contempt is an emotion oftentimes connected to violence. As stated in theoretical framework, contempt is a primary dyad emotion, meaning it's a combination of two emotions elicited together, here being disgust and anger. Hence, to create this 9th emotion, if the value of disgust and anger is equal to or above one, then contempt variable takes the value of mean of these two values.

Count of the emotionally loaded words can be useful in identifying the intensity of each emotion in a single tweet, however, not all tweets are equal. For example, there were 120-word tweets with 10 words loading on certain emotions and some tweets containing only one emotionally loaded word. Therefore, to account for the discrepancies in word count, new section of flag variables was created. Those were 8 flag variables such as flag_anger or flag_trust that took a value of 0 when there were no emotional loadings in the tweets and 1 when there was at least one could of an emotion.

In total there were 22 emotional variables found, 11 with the count of emotionally loaded words and 11 flagging existence of each emotion in a tweet. For the remainder of the research, this group is referred to as *emotions count* and *emotions percentage* groups.

4.2 LDA topics

The third feature group extracted were though the Latent Dirichlet Allocation method. LDA is generative probabilistic model that identifies topics that are present in the body of text as well as the contents of such topics (Blei et al., 2003). The first assumption is that the body of text is a Bag of Words. This means that each document (singular tweet) is taken as a vector that contains multiple tokens (words in each tweet). It is important to note that in a Bag of Words approach the order of tokens is irrelevant.

As mentioned, LDA is a generative model. In short, this means that the model can create new data case, where the join probability p(X,Y) is found, rather than conditional probability p(Y|X) (Jebara, 2004). Therefore, the Dirichlet distribution is used:

$$\beta_k \sim Dirichlet(\delta_1, \delta_2, \dots, \delta_n)$$
$$\theta_n \sim Dirichlet(\alpha_1, \alpha_2, \dots, \alpha_K)$$

1. **To choose the number of words in for a document**
2. **For each of the number of words:**

   Select a topic $z_{in} \sim Multinomial(\theta_i)$

   Select a word $w_{in} \sim Multinomial(\beta_{z_{in}})$

Where

K – number of topics

β_k – term distribution for topic K

θ_n – topic distribution for document n

δ – parameter guiding the distribution of β_k

α – parameter guiding the distribution of θ_n

Each of these steps has to be done for all the documents in the dataset (Proto, 2018). By following these steps LDA creates a mix of topics per each document. By providing the model with parameters α and β, it is possible to progress with the computations. This enables to compute the probabilities of each document being part of an LDA topic, as well as identifying what tokens make up each topic. The process is illustrated in figure x (Lee, Kang, and Jun, 2018).

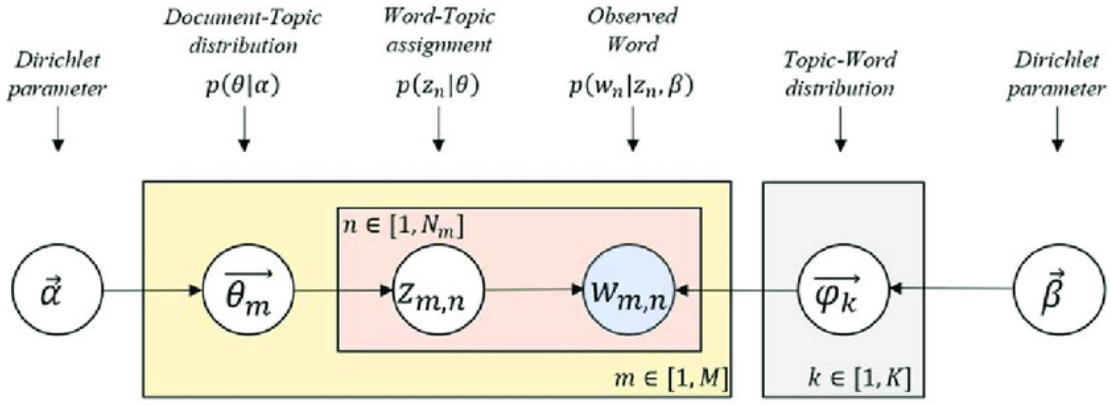


***Figure 5**: Visual explanation of the LDA model functioning*

In total there were 40 LDA-classified topics found. A complete list of LDA derived topics used in models can be found in Appendix 2. For the remainder of the research, this group is referred to as LDA topics.

4.3 Violent tweets classification

First, all the tweets were classified along the methodology of Anastasopoulos and Williams. To do so, a model had to be built that was applied to the tweet protest dataset. Three feature groups were used to build the model, namely one and multi-grams, LDA topics and emotional loadings. Those data transformation methods were used both for the violence and Tweet datasets.

The first method applied was identifying singular words that are most common in the tweet violence classification dataset. In total there were 10 479 one-grams identified such as 'peopl', 'fuck', 'protest' or 'brown'. Many of the words were however not meaningful, such as numbers 1-500, words like 'lot' or 'ur'. There were many words strictly connected to locations such as 'ferguson', 'hongkong' or 'la' that wouldn't be useful when predicting force/peace in other locations. All of these were removed and the top 500 one-grams was chosen. A complete list of one-grams used can be found in Appendix C.1. For the remainder of the research, this group is referred to as one-grams.

The other important features extracted from the body of tweets are multi-grams. Those are strings of text that contain two or more words and occur in the data. It was specified that the minimum number of words was two and maximum three, with the maximal length between them of four words. After removing multi-grams such as 'gt gt gt', 'realli realli realli' or 'york ny' there were 223 multi-grams left. A complete list of multi-grams used can be found in Appendix C.2. For the remainder of the research, this group is referred to as multi-grams.

LDA and emotional loadings were extracted from the violent tweets' classification dataset in the same manner as explained in sections 4.1 and 4.2.

5 Methods

5.1 Random Forest model

As stated in the theoretical framework, machine learning and data science field grew enormously over the past decades. The main distinction made between models is whether it is a black-box or white-box models.

Black-box models are usually more complex. This term is referred to models where inputs are passed to a model, then computations are made that are hard to impossible to understand or interpret and an output is generated. In those cases, the evaluation is mostly based on comparing the input and the output of such model omitting the explanations of what's happening inside. Neural networks or ensemble methods such as random forests and bagging are some examples of such models.

White-box models on the other hand are fairly powerful yet less precise than black-box due to lower computational complexity. The advantage they have over black-box models, however, is they are more easily to interpret in a human-understandable way. Examples of such models are linear regressions or decision trees.

The model of choice for the task is random forest. It is an ensemble method comprised of decision trees. Random forest was first created in 1995 (Ho, 1995), however popularized by Breiman (2001) just six years later. He added a tree bagging method, which increased the precision of the model, very close to other ensemble methods such as boosting or bagging. The main strengths of random forests is that they reduce the risk of overfitting, is very flexible as it is both a classification and regression model and it is easy, which is important for this research, to determine the feature importance (IBM, 2021).

Random forest algorithm is comprised of three steps. The first one is bagging, also known as bootstrapping. It is a method where *N'* observations are sampled with replacement from the original dataset of size *N* and put into multiple smaller datasets (Abney, 2002). Since the observations are sampled with replacement, it is possible to have the same observation repeated inside a bootstrapped dataset as well as in two separate bootstrap samples. This enables to build many decision trees on each of the datasets.

The random forest model of Breiman utilizes a second step already created by Ho, namely the random subspace method. This method randomly selects multiple features that are then used to grow the trees. This is the main difference between random forest and decision trees, as the latter one uses all the features to identify viable splits in the model. By utilizing random subspace method, a lower correlation between the decision trees is achieved.

There are two main methods in identifying the best split locations for the trees. The first metric is Information gain. It is approximated by calculating the decrease of *entropy* when transforming the dataset. In summary, this metric identifies the gain in explanatory power by adding a variable to the tree. This method performs well in identifying well-fitting trees and assessing variable importance. Entropy being *E* in the formula below

$$E(c) = -\sum_{i=0}^{n} p_i \log(p_i)$$

Where E(v) stands for entropy for the class c, n is the number of classes and p_i is the probability of an event happening. After calculating entropies, the information gain is calculated

$$IG(c1, c2) = Entorpy(c1) - Entropy(2)$$

$$IG(c1, c2) = -(\sum_{i=0}^{n} p_{c1} \log(p_{c1}) + \sum_{i=0}^{n} p_{c1,c2} \log(p_{c1,c2}))$$

Where c1 is the 1st category, $p_{v1}$is the probability of an event happening.

The other metric used for constructing the decision trees is Gini impurity (Laber and Murtinho, 2019). It is describing the possibility of randomly chosen datapoint to be misclassified by the tree. It is calculated as a sum of probabilities of classes classified correctly, multiplied by 1 – the probability of classes classified correctly:

$$Gini(c) = \sum_{i=1}^{n} p_c\,(1 - p_c)$$

$$= \sum_{i=1}^{n} (p_c - {p_c}^2)$$

$$= 1 - \sum_{i=1}^{n} {p_c}^2$$

After the data is bootstrapped and hundreds of trees are fitted, majority vote takes place. It is a standard ensemble prediction voting method. In this step the final random forest model is decided upon as the predictions of distinct trees are taken as votes and the accuracy of the forest is determined through majority voting (Brabec and Machlica, 2018). The Figure 6 depicts graphically the model creation process of random forest algorithm (Collaris, 2018).

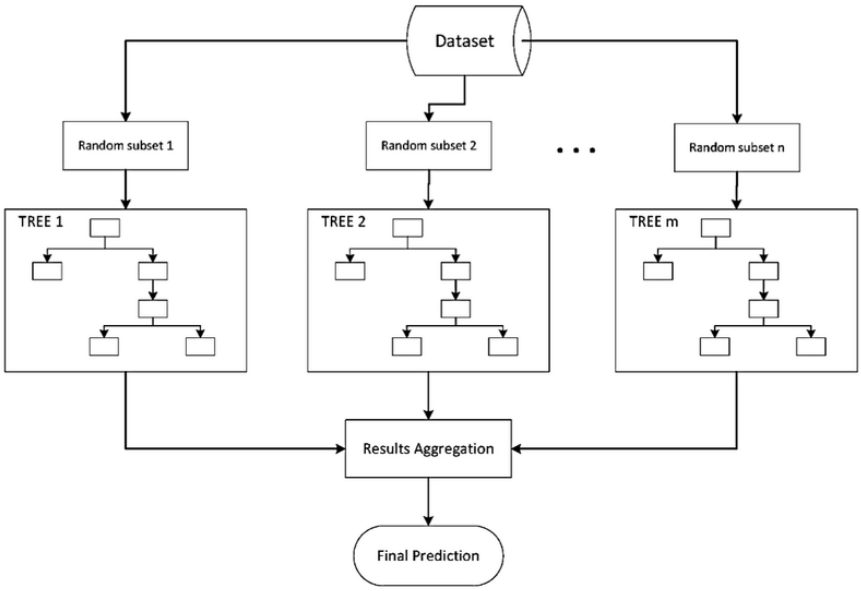


***Figure 6**: Visual description of random forest model functioning. From the dataset first bootstrapping is applied, followed with decision trees creation and results aggregation in the form of majority voting (Collaris 2018)*

5.2 Modelling the violence tweet classification

As explained in the previous steps, the three feature groups applied to the Tweets dataset were emotions, LDA topics and violent tweet classification. Both emotional analysis and LDA were applied to both Tweet dataset and Violence dataset. However, the violent classification had to be first modelled on the Violence dataset and then applied to the Tweets dataset.

There are 7596 observations in the violence dataset. Prior to training the model to predict violence group of each of the tweets, the data was split into train and test sets with a ratio of 7:3. On the train data random forests were deployed. Since all of the violence groups (singular/collective peace/force) are separate binary variables, four random forest models were trained on one-grams, multi-grams, emotions loadings and LDA topics. Afterwards, all four models were applied on the tweets dataset to create a third feature group for the final analysis.

5.2 Modelling the Tweet dataset

After aggregating all of the three feature groups, a final model was trained on the data. To do so, the final dataset containing 184 observations was randomly split into train and test sets with a ratio of 7:3. On this data five models were run.

The first model is based solely on emotional features from the tweets. As described in the previous parts, those contain the 22 variables describing the emotional loadings per protest, expressed as a total count and as a percent of the whole protest. The third feature group that was looked at was the violent tweets classification. Those were expressed as a count of separate categories per protest as well as a percent of the whole protest. The fourth model contains all three feature groups.

The last model was built on a dataset where all four classes are present. This model based on a wider dataset of 334 observations and its goal is to classify protests into both *violent protest* and *peaceful protest* groups as well as into *excessive force against protesters* and *protest with intervention.*

5.3 Model fitness

All of the models needed to be evacuated to understand how well they do perform. One of the most common performance measurements is confusion matrix. It is a table where comparison between the actual values and those predicted by a model. In models 1.1 – 1.4 and 2.5 the dependent variables are binary; hence the confusion matrices are tables with four combinations of values. For models 2.1 – 2.4 the *sub event type* variable takes two possible values. Only model 2.5 is deployed on a four-class dataset, hence the confusion matrix is of size 4x4.

However, there is much information lost when looking solely at a confusion matrix. To better understand the fit of the model, 5 machine learning metrics are used. Those are:

Accuracy – the simplest metric out of the six. It is the correct number of predicted values divided by the total number of predicted values

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100\%$$

Precision – there are instances where accuracy is high, however it cannot be assumed the model is of good fit. For example, when there is high class imbalance of the dependent variable. Precision accounts for that, as its focus is to show the accuracy for the class with lowest count.

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

Recall – it is a similar metric to precision. The difference between them is that precision looks solely at the correct true predictions out of all true predictions, whereas recall accounts for the missed true prediction possibility.

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

F1 score – in models both precision and recall are important to a different extent depending on the model and its functionality. To give a general description of a mode F1 is used as it combines both Precision and Recall into one variable.

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Specificity – the last metric used to identify how many actual negatives were predicted as the negative. It indicates how well the model identifies the negative class, which is the exact opposite of precision.

$$Specificity = \frac{TN}{TN + FP}$$

5.4 Features importance

Models can be evaluated based on metrics generated from predicting the test data. Regardless of performance of those models it is also imperative to understand what features are driving the results and are most influential for the model.

The feature assessment algorithm applied is Boruta. It is a feature selecting method that is a wrapper on random forest. Wrappers are algorithms that are computationally expensive, however, can yield very precise results. Since there are only 184 and 334 observations in the final datasets, Boruta was deployed successfully. The model orders all the features from least to most important for prediction and groups them into three classes. The first one groups the variables to drop, then potentially useful features and the third one – very important features that must be retained, as opposed to variable importance plots that lack the grouping part.

6 Results

6.1 Results of the violent tweet classification models

The four models described in section 3.5.4 were first applied to the test dataset. As seen in the Figure 7, the models are fairly successful in predicting singular peace and singular force. Big discrepancies start to show in models identifying collective peace and force. It seems that based on the 500 one-grams,

223 multi-grams, 22 emotional features and 40 LDA topics peace and force is best distinguished for tweets where only one person is involved.

Confusion matrices for all four models show that models are effective at predicting both 0 and 1 classes. This is further confirmed by high accuracy of all models, between 0.8323 of singular peace to 0.9390 of collective peace models, Table 5.

In the dataset there are significant imbalances between classes. For example, for singular force there are 2260 non-singular force reference observations and only 19 singular force reference observations. This, however, does not affect the fit of the mode, as precision is high for all four models with values above and equal to 0.8413.

Both recall and F1 scores are high, in ranges of 0.9594 to 0.9982 and 0.8965 to 0.9809 respectively. Therefore, it can be assumed the is little missed true prediction possibility.

The lowest variable for all four models is specificity that ranges from 0.1141 to 0.4561. This indicates the models performs much worse at classifying negatives as negatives compared to properly classifying positives.

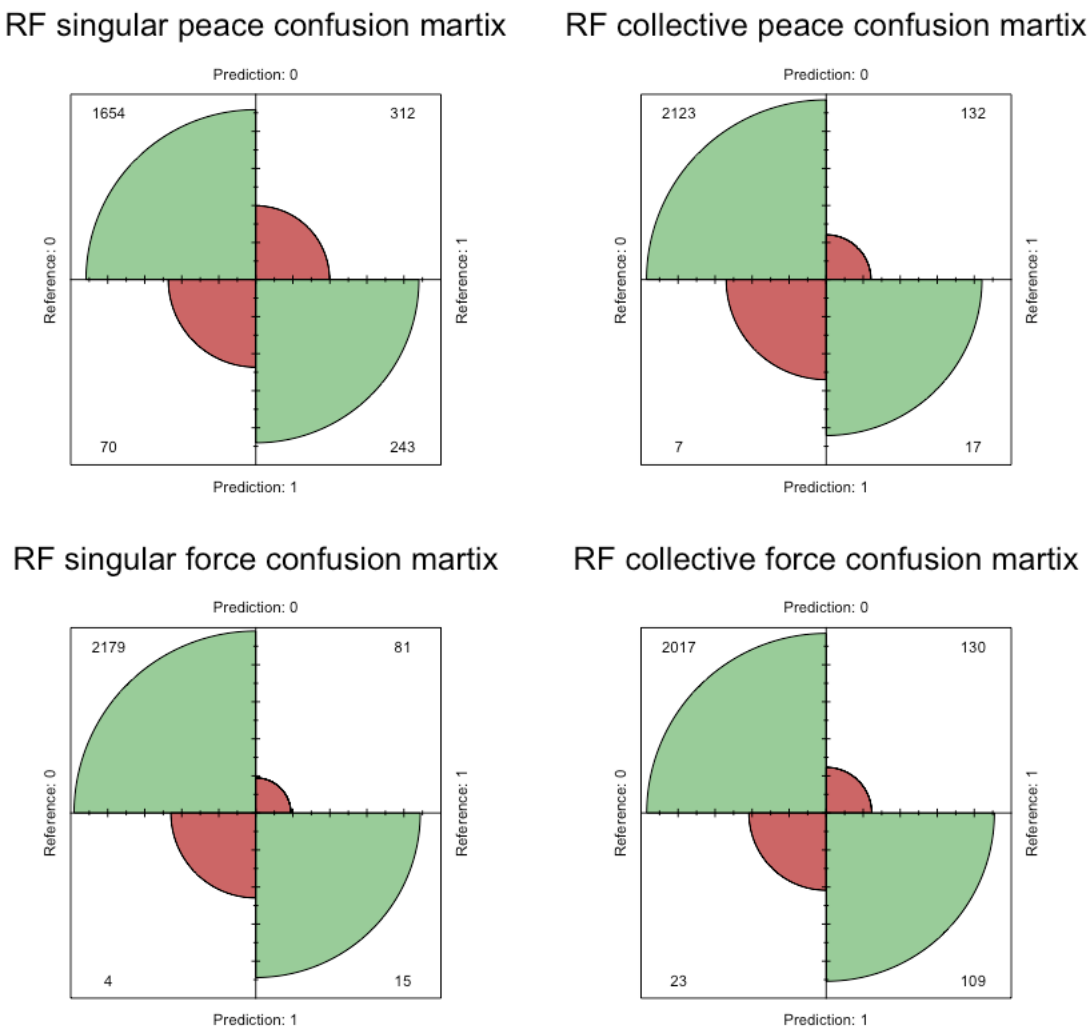


***Figure 7**: Confusion matrixes of the random forests tweet violence prediction on the test set for categories singular peace, collective peace, singular force and collective force*

Model	Accuracy	Precision	Recall	F1 Score	Specificity
Singular peace	0.8323	0.8413	0.9594	0.8965	0.4378
Collective peace	0.9390	0.9415	0.9967	0.9683	0.1141
Singular force	0.9627	0.9642	0.9982	0.9809	0.1563
Collective force	0.9329	0.9395	0.9887	0.9635	0.4561

***Table 5**: Twitter violence classification models fitness evaluation metrics*

6.2 Results of the violent protest models

6.2.1 Emotions

The first violence protest model aimed to identify whether it is possible to successfully distinguish whether the protest will stay peaceful or turn violent based on emotional loadings of tweets (H1). Therefore, it was run solely on emotion counts and percent of total count.

In figure 7.1, it is visible there are many more positive emotions encountered than negative with 1 200 031 and 996 213 counts respectively. It is further shown that trust is the most common emotion detected in the protest related tweets. The next two emotions are both negative and very similar in total count: fear and sadness. Anticipation, anger and joy have similar have similar counts of approximately 500 000. After that, there is a significant drop in the remaining three emotions. Contempt, disgust and surprise were identified approximately 250 000 times, more than half of trust count. Having filtered those tweets to be only related to protests it is surprising there are more positive emotions than negative. The second outtake is that contempt has a very similar number to disgust. As contempt is a secondary dyad of disgust and anger, this indicates in almost all instances where anger was found, disgust was also present. Therefore, since there are possibly few observations where solely disgust or fear are present, the contempt variable is unlikely to be highly meaningful.

Figure 7.2 depicts the counts of emotions divided per sub event type. As seen, there are no significant differences between emotional loadings of tweets linked to each sub event type. Again surprisingly, there are more positive emotions linked to violent demonstrations, whereas there are more negative emotions liked to peaceful protests. The only outtake is that anger, fear and surprise are more common emotions for violent protests whereas sadness is more often linked to peaceful protests. There are no significant differences in emotional loadings between peaceful and violent protests, hence this is another proof the model is predicted to perform poorly. On other hand it is also possible those differences are big enough for random forest to pick up the smallest traces.

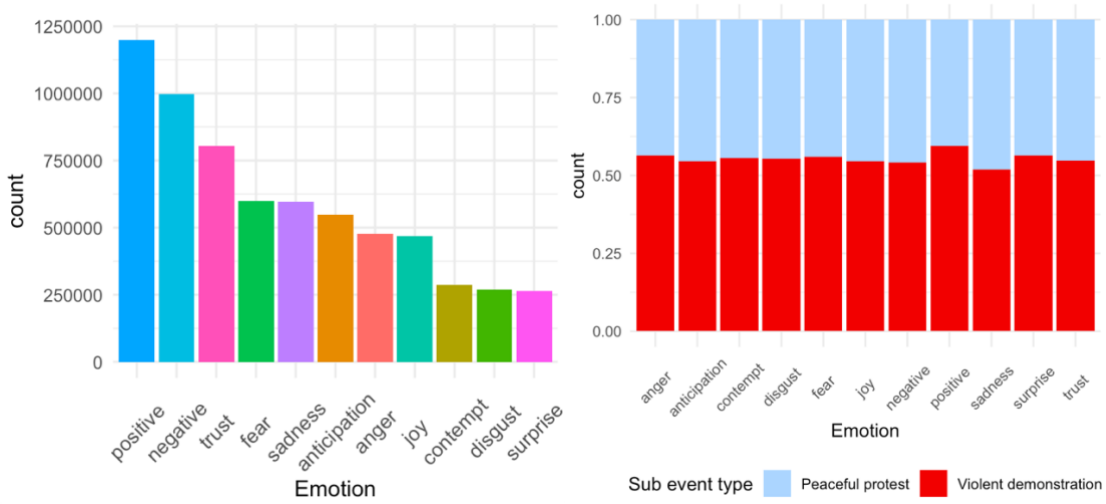
Looking at the confusion matrix in figure 7.3 it is visible the model is rather imprecise, as 17 violent demonstrations were wrongfully classified as peaceful protests and 8 instances of peaceful protests classified as violent demonstrations with counts of 12. There were 15 observations classified correctly both as peaceful protests and violent demonstrations.

This inaccuracy is further proven when looked at table 6. Accuracy is only slightly above 0.5 with a value of 0.5455. The other four metrics are in the range between 0.4687 and 0.6522 indicating, it is approximately as accurate as blindly guessing which protest will turn violent.

Further, in graph 10.2 there are five meaningful emotional features to the model. The most important features are *fear percentage*, *anger count*, *fear count*, *anger count*, *sad count* respectively. It was expected for *contempt* to be of high importance in this model (Becker and Tausch 2015). This is only partially confirmed by the model, as contempt is the 6th most important variable in the model.

Model	Accuracy	Precision	Recall	F1 Score	Specificity
Emotions model (1)	0.5455	0.4687	0.6522	0.5454	0.4687

***Table 6**: Fitness evaluation metrics of the emotion loadings feature group random forest model with two classes.*



***Figures 7.1; 7.2**: The first figure to the right is an ordered bar chart displaying the number of all emotions identified in the Tweets database. The figure to the right is a division of all emotional count per type of protest – either peaceful or violent*

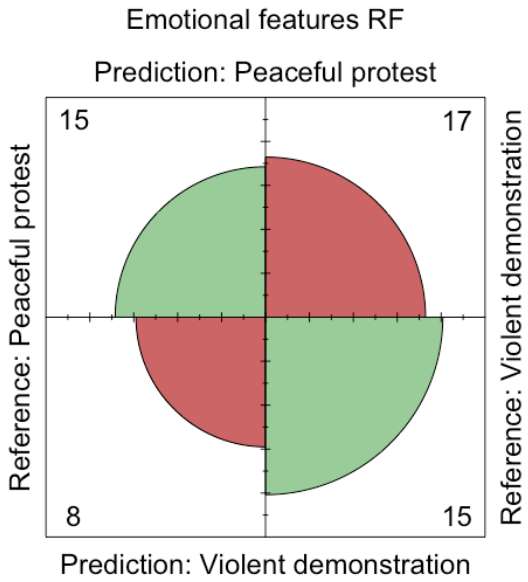


***Figure 7.3**: A confusion matrix of the emotions feature group random forest classification*

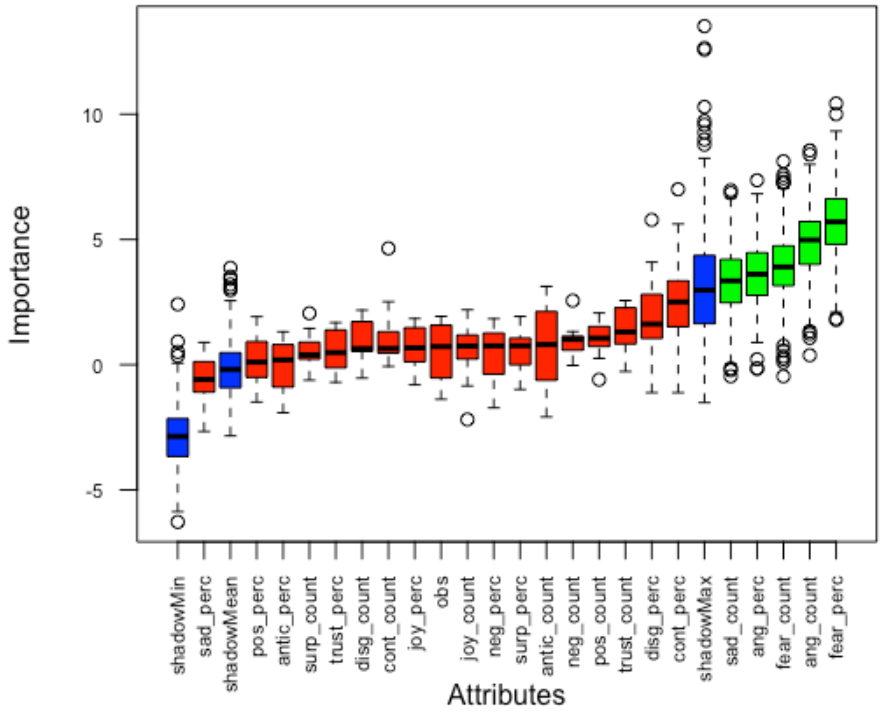


***Figure 7.4** Boruta feature importance plot of the emotion features group random forest model.*

6.2.2 LDA

To answer the question whether LDA classified topics perform well in predicting whether protest will be peaceful or violent (H2), LDA topics were used as independent variables. The confusion matrix of the LDA model (Figure 8.1) looks similar to the one on emotional model. It is however slightly better since the number of *peaceful protest* was correctly predicted for 16 instances and misclassified for 7, while *violent demonstration* was correctly predicted in 17 instances and incorrectly in 15.

The better fit is further proven by the fitness metrics with accuracy of 0.6000 and all other metrics not only higher than 0.5 but all also higher than the emotional model, shown in table 7.

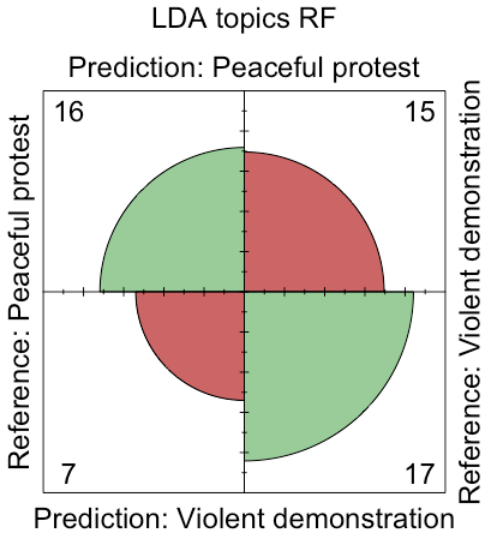


Figure 8.1*: Confusion matrix of the LDA topics feature group random forest classification*

Model	Accuracy	Precision	Recall	F1 Score	Specificity
LDA model (2)	0.6000	0.5161	0.6956	0.5926	0.5312

Table 7*: Fitness evaluation metrics of the LDA classified topics feature group random forest model with two classes.*

In the Figure 8.3, Boruta model selected 12 meaningful topics as well as one tentative. The most important topic is topic 40, which is significantly more important that the others in predicting violent protests. The top 9 most important topics were then selected to identify most important words in each topic (Graph X). In topic 40 those words are *plai, crimin, care, Michael, patriot, eric, taylor, beronna, clown, coward.* This topic is heavily linked to Breonna Taylor killing. It is also linked to possibly contemptuous description of people, such as *clown, coward* and *crimin.*

The next important topic is the first one. It is linked to a second killing of a black citizen, George Floyd. The topic is also related to protests (*protest, riot)*, colour (*black, white, orange)* and is focused mainly on New York.

The third most important topics is also linked to colour as well as probably George Floyd, Donald Trump and is mainly related to Los Angeles. It is less violence intense, as there are words *support* or *gt (great)* rather than *riot* and *protest* in topic 40.

All of the other topics are similar. They all describe topics related to colour, Donald Trump, police and Black Lives Matter movement. They all differ slightly in intensity of words used. For example, topic 28 talks about colours and police while there are words *petit, sign* indicating rather peaceful forms of action. Many others are linked to rioting or criminality. Those similarities showcase that all the tweets are heavily linked to the protests and the systematic issues in the U.S. The small differences between topics and their intensity are what gives significant power in distinguishing peaceful protests from violent ones (Figure 8.2)

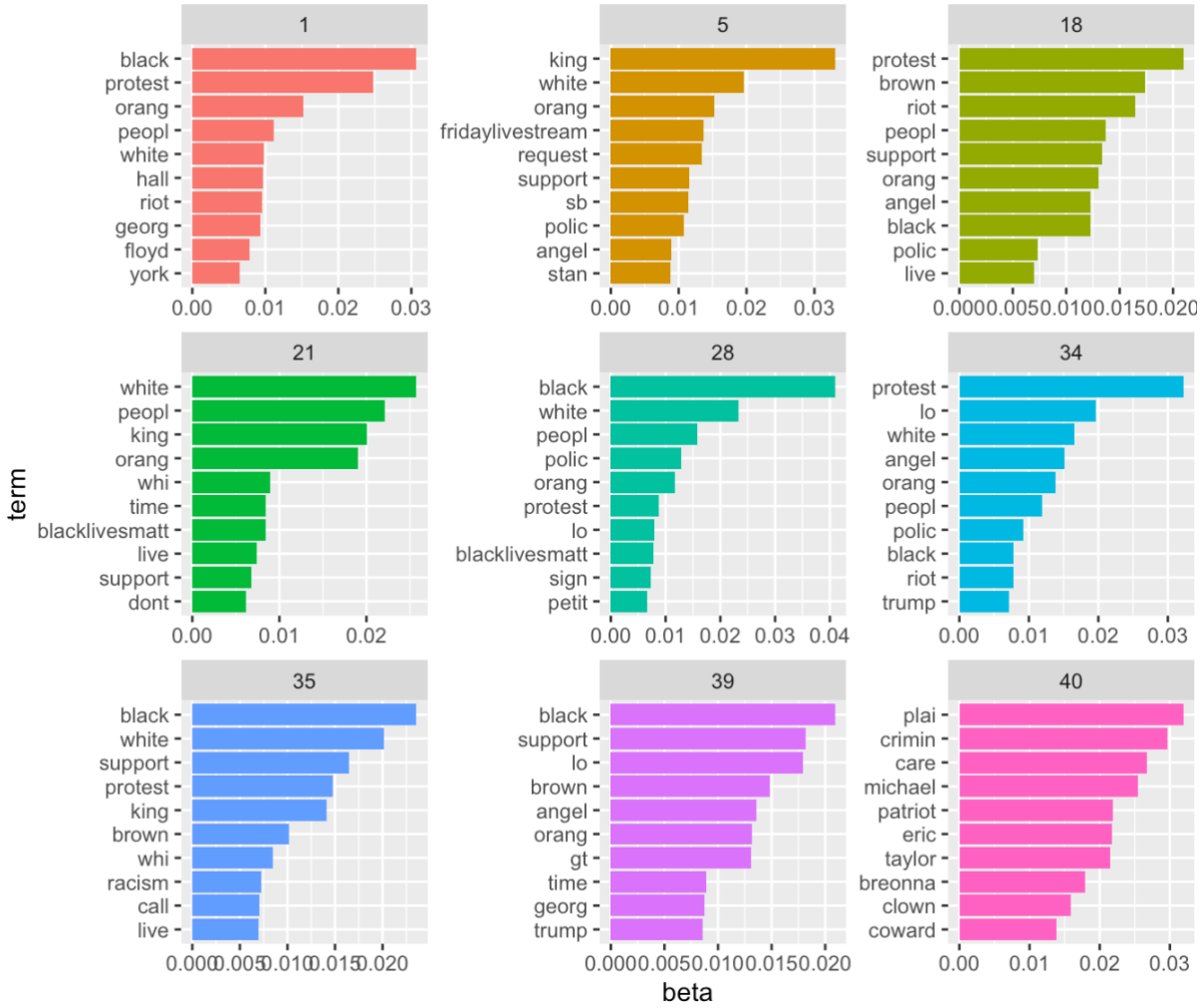


***Figure 8.2**: LDA topics identified in the Tweets dataset of 9 most important topics for the model, ordered by the beta significance.*

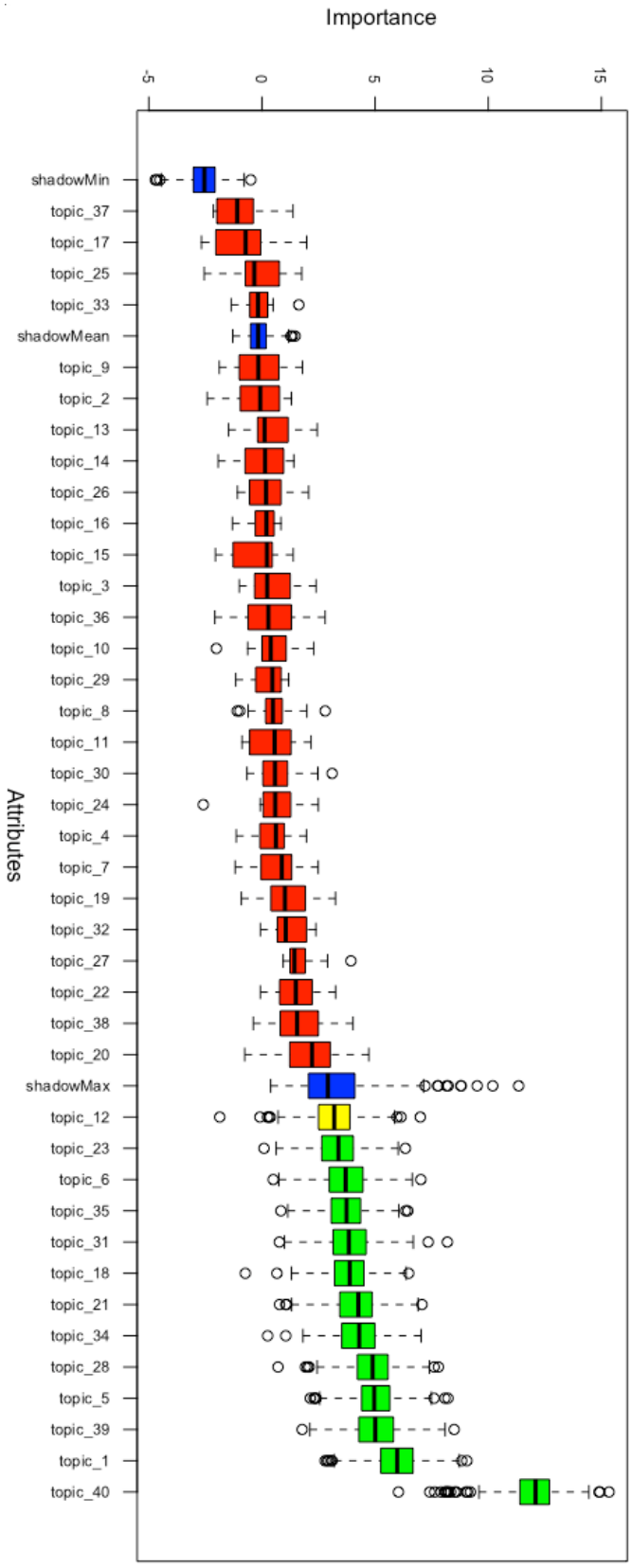


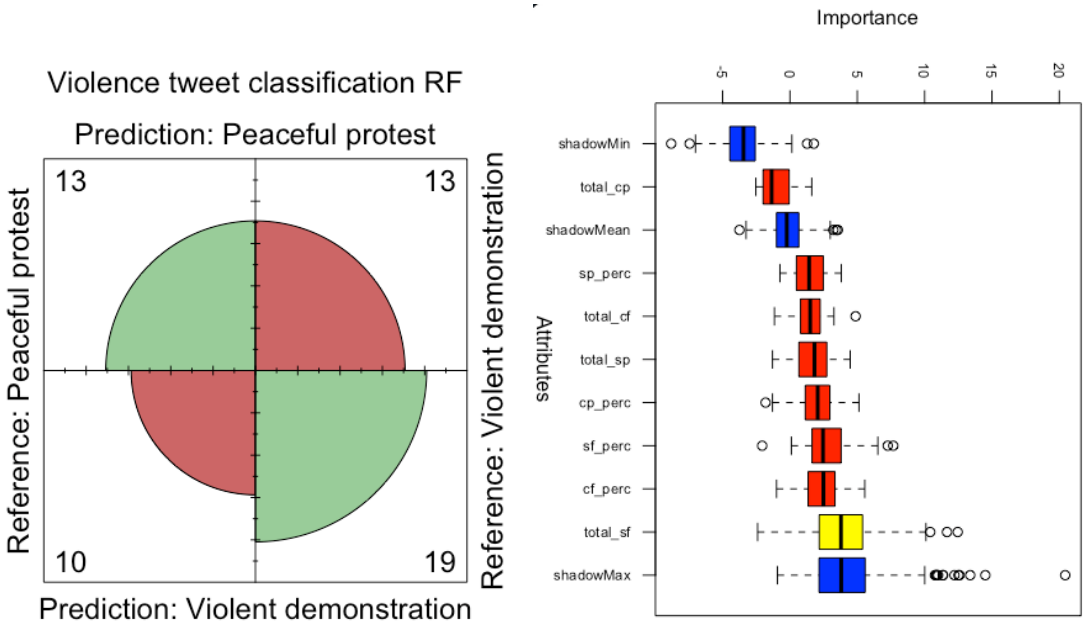
***Figure 8.3**: Boruta output on LDA topics feature group random forest model; shows the variable importance for the model. Each variable was classified into insignificant (red), tentative (yellow) and main features (green).*

6.2.3 *Violence classification*

The last feature group that a distinct model was ran on was the violence tweet classification. The aim was to identify whether violence tweet classification can yield significant results in predicting violence and peace at protests (H3). Visible form Figure 9.1, model classified correctly 19 out of 32 protests as *violent demonstration* and 13 out 23 *peaceful protest* classified correctly.

The fitness metrics in Table 8 indicate this model is even more precise than the LDA one. Accuracy increased from 0.600 to 0.6364 while recall is identical. F1 score increased to 0.6154 as well as specificity which is close to 0.6.

Even though all the fitness metrics are all an upgrade from the LDA model, the Boruta algorithm in Figure 9.2, identified none of the features to be significant, with only *total singular force* being in the tentative group. If omitted the significance of importance of the metrics, the next most important variable is *collective force percentage*, followed by *singular force percentage* and *collective peace percentage*. The last two metrics are *singular peace percentage* and *total collective peace*. This can indicate that it is the percentage and number of violence loaded tweets that is a better predictor of violence at protests rather than peace classification.



Figures 9.1; 9.2*: The figure to the left is a confusion matrix of the violent tweets' classification feature group random forest classification. The figure to the right is the Boruta feature importance plot of the violent tweets' classification feature group random forest model.*

Model	Accuracy	Precision	Recall	F1 Score	Specificity
Violent tweet classification model (3)	0.6364	0.5517	0.6956	0.6154	0.5937

***Table 8**: Fitness evaluation metrics of violent tweet classification feature group random forest model with two classes.*

6.2.4 Three feature groups model

The last two-class model is a random forest with all three feature groups as independent variables. This model answers the question whether a two-class dependent variable is better predicted by the model than the four-class model (H4).

Looking into table 9 it is visible this model attained a good fit in all of the five metrics. The model is accurate in 87.27% instances, has a recall of 0.6958 and F1 score of 0.8205. Moreover, both precision and specificity are equal to 1. The confusion matrix in figure x confirms the model performs well in predicting *peaceful* and *violent protests* with 16 correctly classified *peaceful protest* and 7 *peaceful protest* predicted to be *violent demonstration*. *Violent demonstration* was classified correctly in 32 instances and not one *violent demonstration* was predicted as *peaceful protest*.

Model	Accuracy	Precision	Recall	F1 Score	Specificity
Three feature groups model (4)	0.8727	1.0000	0.6957	0.8205	1.0000

***Table 9**: Fitness evaluation metrics all three feature groups random forest model with two classes.*

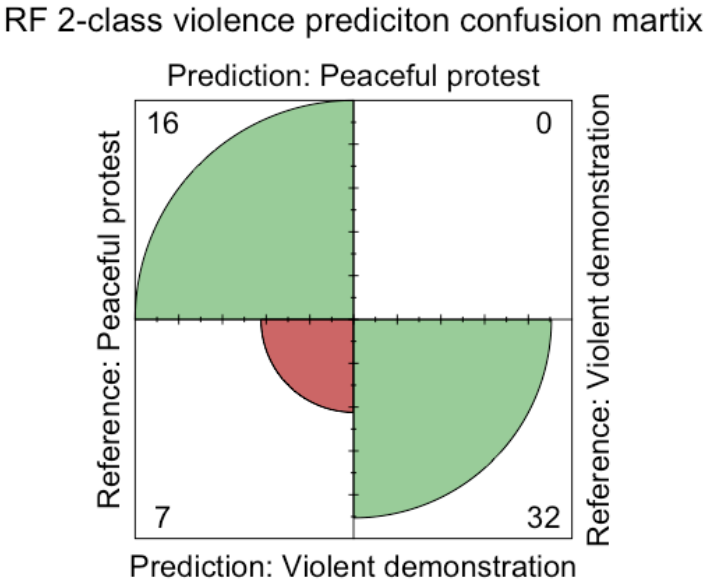


***Figure 10.1;**: Confusion matrix of the three feature groups random forest classification in the two-class model.*

Further, Boruta importance plot in graph 10.2 indicates a very high model importance for topic 40. It is the best distinguishing feature between *violent demonstration* and *peaceful protest* by far. In total there are 10 essential variables, 5 tentative and 56 irrelevant ones. This graph indicates *topic 40* is very important, and is within the acceptance range along topics 1, 5, 39, 28, followed by *anger count*, topics 34, 21, 35 and 18. The tentative group comprises solely of LDA identified topics. There are no emotional features, except *anger count*, and no violent tweets classification variables are present in the essential and tentative model variables.

Even though the violent tweet classification model was the most accurate, the most important metric is ranked 19th out of all features and is *collective force percentage*. It is followed 12 places later by *singular force percentage* and *singular force count*.

Therefore, it is assumed LDA topics are the most important features in a model predicting violent protests, however only in combination with emotional features and violent tweet classification the model performs very well. All the three feature groups are important since the fitness of all metrics increased compared to the best single-feature group model as in Table 10.

Model	Accuracy	Precision	Recall	F1 Score	Specificity
Violent tweet classification model (3)	0.6364	0.5517	0.6956	0.6154	0.5937
Three feature groups model (4)	0.8727	1.0000	0.6957	0.8205	1.0000
% change	+ 37%	+81%	=0%	+33%	+68%

***Table 10**: Model fit comparison between the best one-feature group random forest and three-feature groups random forest*

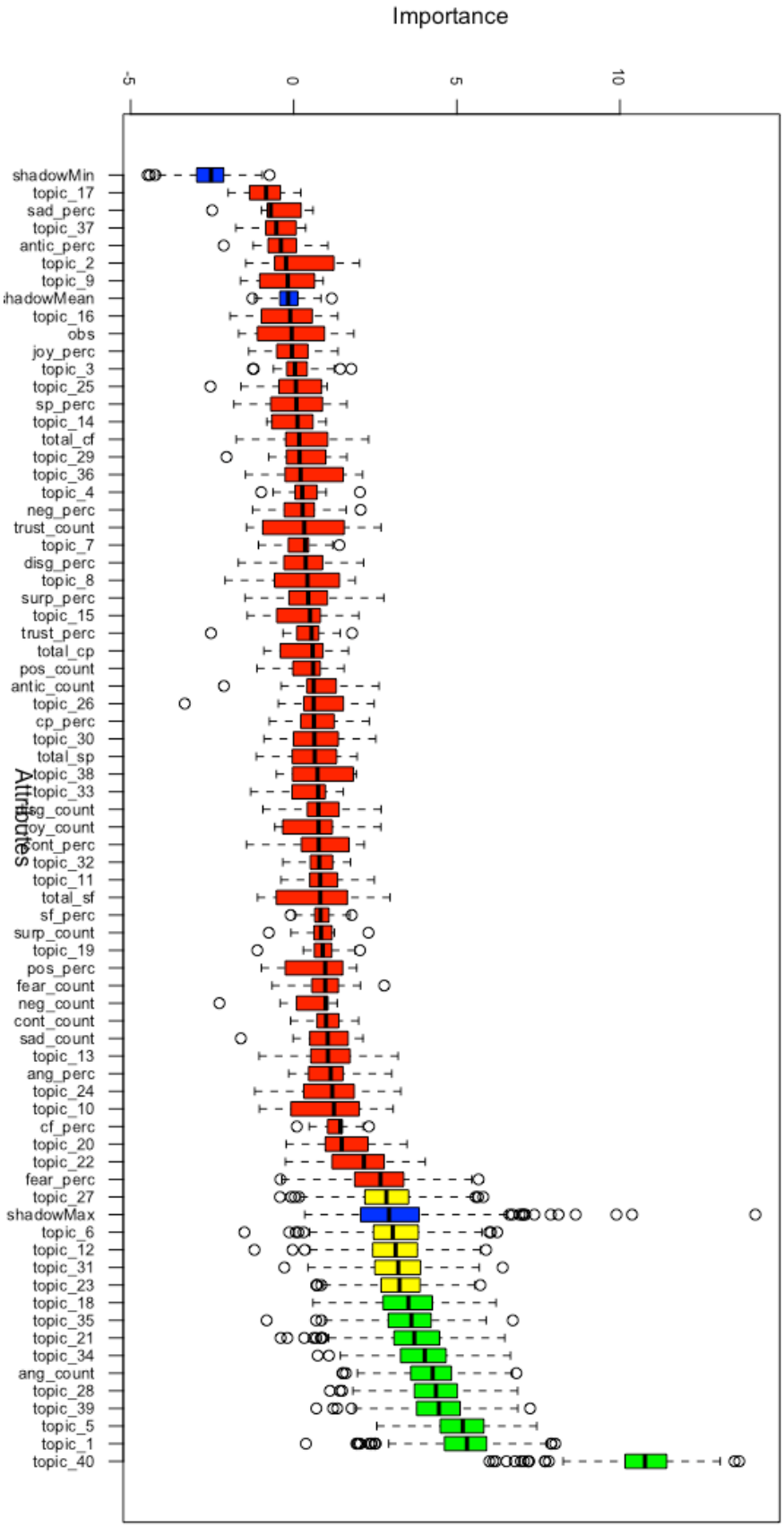


Figure 10.2*: Boruta output on three feature groups on two-class random forest model; shows the variable importance for the model. Each variable was classified into insignificant (red), tentative (yellow) and main features (green).*

6.2.5 *Three feature groups four-class model*

Given the successfulness of the two-class model, it was necessary to assess the utility of the model in distinguishing between four classes. The task is much more complex, as the differences between classes are not as clearly defined as between *violent demonstration* and *peaceful protest.* The task at hand was to answer whether a combination of features in the model can attain a reasonable fit at predicting violence at protests in a four-class setting (H5).

Comparing to the two-class models the accuracy obtained was just 26%. Precision is also rather small as it ranges between 0.1428 to 0.3871 (Table 11)

Excessive force against protesters is the weakest class in terms of predicting correctly, as well as it is the lowest represented class with just 20 observations in the test set. Its recall is equal to 0.0500 while the metric ranges between 0.2083 to 0.3750 for other classes.

In terms of F1 score, *excessive force against protesters* has the lowest value of the four classes. It is then followed by *protest with intervention* with a score of 0.1695 and *peaceful protest* with score 0.3137.

Specificity is high for all groups, but that is rather due to high number of true negatives.

***Table 4**: Fitness evaluation metrics of the emotions, LDA topics and violence tweet classification random forest model with four classes.*

Class	Accuracy	Precision	Recall	F1 Score	Specificity
Excessive force against protesters	0.2600	0.1428	0.0500	0.0741	0.9250
Peaceful protest	0.2600	0.2963	0.3333	0.3137	0.7500
Protest with intervention	0.2600	0.1428	0.2083	0.1695	0.6053
Violent demonstration	0.2600	0.3871	0.3750	0.3810	0.7206

***Table 11**: Fitness evaluation metrics of the emotions, LDA topics and violence tweet classification random forest model with four classes.*

The Figure 11.1 depicts the confusion matrix of the full model. It is visible not all of the classes are well classified. The model performs worse than models on separate feature groups. *Violent demonstration* is most accurately predicted with 12 such protests classified correctly, however, also 12 *violent demonstrations* were classified as *protest with intervention*. *Protest with intervention, peaceful protest* and *excessive force against protesters* are predicting very unwell, as they have successfully classified 5, 8 and 1 observation respectively.

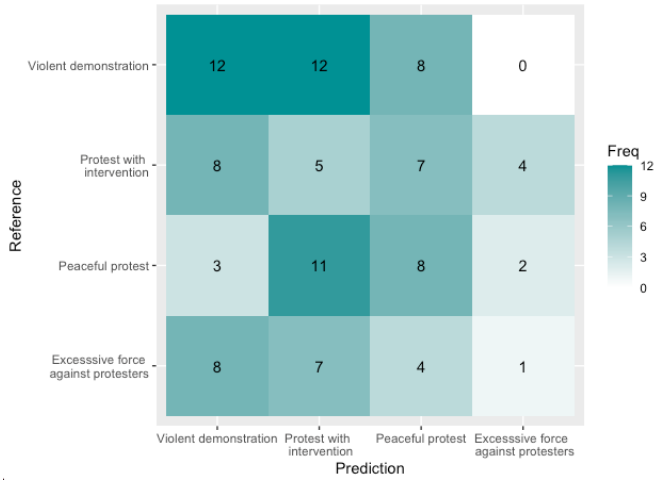


***Figure 11.1**: A multiclass confusion matrix of the three feature groups in a four-class random forest classification*

Graph 11.2 is the visualization of the Boruta random forest wrapper. All of the blue boxplots are indicators of min, max and mean z scores of features. All the variables marked with a red boxplot are deemed irrelevant, yellow represent tentative variables and green – confirmed features.

There are 5 topics that are below the 0-importance mark, along the *singular force count* variable. Afterwards, all the features have positive importance, however many are dropped. There are 53 variables to be dropped consisting of 16 emotional features, 31 topics and 5 violent tweet classes. The tentative variable group consists of *singular force*, *trust count*, *topic 34*. The most important feature is *total collective peace*, *topic 32*, *topics 1*, *topics 23*, *disgust percentage*, *fear percentage*, *anticipation percentage*, *fear percentage*, *anticipation count*, *surprise count* and *topics 35*. Those 9 variables are the main features on which random forest predicts the data.

This order is very different to the results of the two-class model. It can be assumed, that when trying to predict four classes, emotional and violent tweet classification features are much more important than topics. However, this cannot be soundly confirmed as the fitness of the model is very poor.

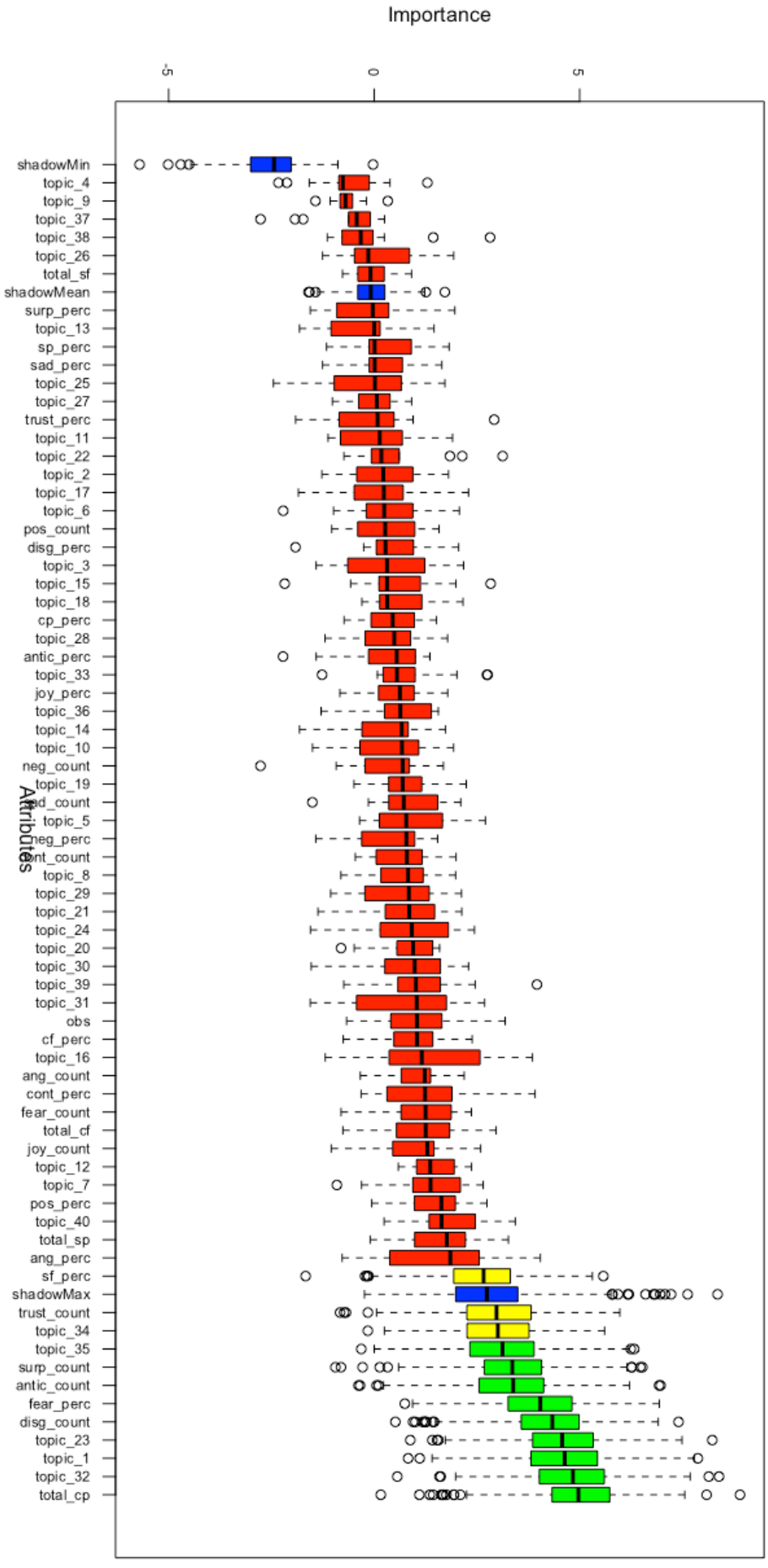

Importance
-5
0
5
Attributes
shadowMin
topic_4
topic_9
topic_37
topic_38
topic_26
total_sf
shadowMean
surp_perc
topic_13
sp_perc
sad_perc
topic_25
topic_27
trust_perc
topic_11
topic_22
topic_2
topic_17
topic_6
pos_count
disg_perc
topic_3
topic_15
topic_18
cp_perc
topic_28
antic_perc
topic_33
joy_perc
topic_36
topic_14
topic_10
neg_count
topic_19
topic_5
topic_8
topic_29
topic_21
topic_24
topic_20
topic_30
topic_39
topic_31
obs
cf_perc
topic_16
ang_count
cont_perc
fear_count
total_cf
joy_count
topic_12
topic_7
pos_perc
topic_40
total_sp
ang_perc
sf_perc
shadowMax
trust_count
topic_34
topic_35
surp_count
antic_count
fear_perc
disg_count
topic_23
topic_1
topic_32
total_cp

***Figure 11.2**: Boruta output on three feature group of the four-class random forest model; shows the variable importance for the model. Each variable was classified into insignificant (red), tentative (yellow) and main features (green).*

7 Conclusion and Discussion

7.1 Synthesizing the results

The goal of the research was to answer the following research question:

Which social media features are good predictors of violence outbursts during the 2020/2021 BLM protests in the U.S?

The analysis proved, that when building models on separate feature groups (emotional loading, LDA topics and violent tweet classification) it is possible to attain accuracy only slightly above 50%. However, when all features were used in a single model that accuracy increased to 87%. This indicates the model is successful at distinguishing peaceful protests to violent protests solely form Twitter data.

The Boruta model, assessing the importance of variables in the model, indicated that LDA topics are the most important predictors for the task. Even though the emotional and violent tweet classification features were shown to be irrelevant, only after adding them to the model accuracy increased by 37% as well as other fitness measures.

Therefore, all three feature groups are important for the prediction, however have significant explanatory power only when used together. Since the increase in model fit is so substantial it can be assumed that each of the group captures distinct differences between protests.

The four-class model proved to be significantly less precise. Moreover, the variable importance in this model was different to the two-class model. The fifth model favoured emotional loadings and violent tweet classification. Contrary to Becker and Tausch (2015) findings, contempt did not prove to be a good predictor of violence. However, due to poor fit, this model cannot be evaluated with confidence.

7.2 Practical implications

This research proves it is possible to precisely distinguish violent from peaceful protests using solely Twitter data and aforementioned feature extraction techniques. The main result is that violence during a protest is most precisely predicted by LDA topics. This means that more often than not it is the topic

of a protest that can spark a riot. Best predicting topics were related to events such as Breonna Taylor and George Floyd killings. There were many topics related to those situations but it can be assumed the words that are related to those topics and their valence is more important at distinguishing potential violence than the words that occur often in each of the topics.

By being able to accurately identify protest related topics, the next logical step could be to identify parties organising protests. After that, policymakers could run LDA, emotional analysis and violent tweet classification to understand which organisers are linked to igniting aggression at events. This approach could be beneficial to establishing clear communication with the organisers, which is the first step to decreasing a likelihood of an outburst according to Nassauer (2019).

Additionally, policymakers could already start using the tool to make predictions whether there will be violence at a protest. By having the prediction in advance, the police could shift their focus towards the protests that are likely to be violent. The next step would be to build a framework of potential de-escalation tools, apply them to protests predicted to be riots and compare the results of different actions. With further research and numerous features extracting methods it might be possible to build a tool that could predict not only the outcome of a protest, but suggest the most appropriate measures against escalation.

Lastly, quite a simple model (random forest) was applied for the task. An interesting step to take by the policy makers or the police intelligence would be to transform this model to collect large amounts of social media data using the three feature extraction methods and train the model further.

7.3 Limitations of the research

The researched topic of predicting violent and peaceful protests solely on social media data is one of the first ones of its kinds. The analysis is based on a collection of topic related methodologies, however there is much room for improvement.

Technical limitations:

The first limitation of the research is the number of protests taken into account. 7 million tweets can seem like a lot, yet that is only a fraction of all the tweets related to protests. The first step in improving the model would be to increase the number of protests from 334 to at least 10 000 events with different outcomes.

Moreover, the tweets were linked to specific protests solely by scraping tweets with location name in the body. This omits tweets where location is a name of a district, a shopping mall or a different non-city like name. Moreover, the tweets were first filtered for the existence of a location word and then filtered for the existence of a protest related word. Additionally, the snscrape limits the Regex combinations used for tweet identification making it impossible to pass the name of a city and a string containing protest-related keywords.

Additionally, for computational efficiency, the limit of 50 000 tweets per day of each event was set. This limit is big enough for a small location, however, there are multi-million cities where hundreds of protest-related tweets are generated each hour.

Moreover, the only social media source used was Twitter. The general limitation of this social media is that, even though oftentimes informative, tweets are limited to only 280 signs. This heavily decreases the meaningfulness of all feature extraction methods, especially that the language used there proves to be poorly understandable for computers.

Theoretical limitations:

All the models were built only on three feature groups, being emotions, LDA extracted topics and violent tweet classification. There are many other NLP methods that provide a variety of features that can be much more meaningful in distinguishing between violence and peace, as well as too strong police response.

Another interesting feature to add can be more secondary and tertiary emotional dyads. This approach, however, can be futile when applied on short Twitter texts but could yield interesting results with expansion of data sources.

Last limiting aspect, possibly an entire area of research, is the time. This research has a cross-sectional setting. By expanding it into a panel data analysis it could be interesting to learn how early before an event it is possible to identify whether a protest will be violent.

7.4 Suggestions for future research

The first step for future research is to expand the database of protests. Even though 334 protests were identified, there were some observations with just 10 or 11 tweets linked to each. Increasing the daily limit as well as better identification of protest linked tweets can provide big datasets of rich information.

Another interesting feature to look for is the tweeting person. In many instances those were related to news agencies or other information outlets. It could be interesting to deepen the methodology by classifying tweeting entities based on their size, type or political polarization.

Major improvements could be made by applying the methodology on EMBERS system (Ramakrishnan et al. 2014). EMBERS system is very precise in identifying when, where and what will be the topic of future protest event four days prior to an event. Its accuracy is so high most likely because all the information is downloaded from CSR, Facebook, Twitter, news outlets and more. By including varied information sources combined with the precise identification of protest-related posts, tweets and news can prove to significantly increase the predictive power of the model. This could be obtained by using panel-data methodologies.

8 Bibliography

Abdurasulov, Abdujalil. 2022. “Kazakhstan Unrest: ‘If You Protest Again, We’ll Kill You’ - BBC News.” Retrieved April 25, 2022 (https://www.bbc.com/news/world-asia-60058972).

Abney, Steven. 2002. “Bootstrapping.” Pp. 360–67 in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

ADL. 2020. “The Purpose and Power of Protest.” Retrieved August 12, 2022 (https://www.adl.org/resources/tools-and-strategies/purpose-and-power-protest).

Anastasopoulos, Lefteris Jason, and Jake Ryland Williams. 2019. “A Scalable Machine Learning Approach for Measuring Violent and Peaceful Forms of Political Protest Participation with Social Media Data” edited by L. K. Gallos. *PLOS ONE* 14(3):e0212834. doi: 10.1371/journal.pone.0212834.

Anon. 2021. “Rotterdam Police Clash with Rioters as Covid Protest Turns Violent.” *BBC News*, November 20.

Badaoui, Saad. 2020. “Black Lives Matter: A New Perspective from Twitter Data Mining.” *Policy Paper*.

Bahrami, Mohsen, Yasin Findik, Burcin Bozkaya, and Selim Balcisoy. 2018. “Twitter Reveals: Using Twitter Analytics to Predict Public Protests.” 24.

Becker, Julia C., and Nicole Tausch. 2015. “A Dynamic Model of Engagement in Normative and Non-Normative Collective Action: Psychological Antecedents, Consequences, and Barriers.” *European Review of Social Psychology* 26(1):43–92. doi: 10.1080/10463283.2015.1094265.

Blei, David M., Andrew Y. Ng, Michael I. Jordan, and John Lafferty. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3(4/5):993–1022.

Bollen, Johan, Huina Mao, and Alberto Pepe. 2011. “Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena.” 4.

Brabec, Jan, and Lukas Machlica. 2018. “Decision-Forest Voting Scheme for Classification of Rare Classes in Network Intrusion Detection.” Pp. 3325–30 in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45(1):5–32. doi: 10.1023/A:1010933404324.

Breuer, Anita, Todd Landman, and Dorothea Farquhar. 2015. “Social Media and Protest Mobilization: Evidence from the Tunisian Revolution.” *Democratization* 22(4):764–92. doi: 10.1080/13510347.2014.885505.

Buchanan, Larry, Quoctrung Bui, and Jurgal K. Patel. 2020. “Black Lives Matter May Be the Largest Movement in U.S. History.” July 3.

Carnegie Endowment. 2022. “Global Protest Tracker.” *Carnegie Endowment for International Peace*. Retrieved August 12, 2022 (https://carnegieendowment.org/publications/interactive/protest-tracker).

Collaris, Dennis. 2018. “Instance-Level Decision Visualization of Random Forest Models.” *Eindhoven University of Technology Research Portal*. Retrieved August 13, 2022 (https://research.tue.nl/en/studentTheses/instance-level-decision-visualization-of-random-forest-models).

Davies, Katie. 2022. “Facebook vs. Twitter: How Do They Stack Up in 2022.” Retrieved March 28, 2022 (https://www.websiteplanet.com/blog/facebook-vs-twitter-stack/).

van Deth, Jan W. 2014. “A Conceptual Map of Political Participation.” *Acta Politica* 49(3):349–67. doi: 10.1057/ap.2014.6.

GWI. 2022. “Global Social Media Stats — DataReportal – Global Digital Insights.” Retrieved March 28, 2022 (https://datareportal.com/social-media-users).

Ho, Tin Kam. 1995. “Random Decision Forests.” Pp. 278–82 vol.1 in *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 1.

IBM. 2021. “What Is Random Forest?” Retrieved August 9, 2022 (https://www.ibm.com/cloud/learn/random-forest).

Ives, Brandon, and Jacob S. Lewis. 2020. “From Rallies to Riots: Why Some Protests Become Violent.” *Journal of Conflict Resolution* 64(5):958–86. doi: 10.1177/0022002719887491.

Jebara, Tony. 2004. *Machine Learning: Discriminative and Generative*.

Jiménez-Moya, Gloria, Russell Spears, Rosa Rodríguez-Bailón, and Soledad de Lemus. 2015. “By Any Means Necessary? When and Why Low Group Identification Paradoxically Predicts Radical Collective Action: Predicting Radical Collective Action.” *Journal of Social Issues* 71(3):517–35. doi: 10.1111/josi.12126.

Kingson, Jennifer A. 2020. “Exclusive: $1 Billion-plus Riot Damage Is Most Expensive in Insurance History.” *Axios*. Retrieved August 4, 2022 (https://www.axios.com/2020/09/16/riots-cost-property-damage).

Korkmaz, Gizem, Jose Cadena, Chris J. Kuhlman, Achla Marathe, Anil Vullikanti, and Naren Ramakrishnan. 2015. “Combining Heterogeneous Data Sources for Civil Unrest Forecasting.” Pp. 258–65 in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. Paris France: ACM.

Laber, Eduardo, and Lucas Murtinho. 2019. “Minimization of Gini Impurity: NP-Completeness and Approximation Algorithm via Connections with the k-Means Problem.” *Electronic Notes in Theoretical Computer Science* 346:567–76. doi: 10.1016/j.entcs.2019.08.050.

Lee, Junseok, Ji-Ho Kang, and Sunghae Jun. 2018. “Ensemble Modeling for Sustainable Technology Transfer.” *ResearchGate*.

Mooijman, Marlon, Joe Hoover, Ying Lin, and Morteza Dehghani. 2018. “Moralization in Social Networks and the Emergence of Violence during Protests.” *Nature Human Behaviour* 2:13. doi: https://doi.org/10.1038/s41562-018-0353-0.

Mundt, Marcia, Karen Ross, and Charla M. Burnett. 2018. “Scaling Social Movements Through Social Media: The Case of Black Lives Matter.” *Social Media + Society* 4(4):205630511880791. doi: 10.1177/2056305118807911.

Muthiah, Sathappan, Bert Huang, Jaime Arredondo, David Mares, Lise Getoor, Graham Katz, and Naren Ramakrishnan. 2015. “Planned Protest Modeling in News and Social Media.” in *Twenty-Seventh IAAI Conference*. Citeseer.

Nassauer, Anne. 2019. “How to Keep Protests Peaceful.” doi: 10.1093/oso/9780190922061.003.0009.

Ortiz-Ospina, Esteban. 2019. “The Rise of Social Media.” *Our World in Data*. Retrieved August 12, 2022 (https://ourworldindata.org/rise-of-social-media).

Proto, Stefano. 2018. “Enhancing Topic Modeling through Latent Dirichlet Allocation with Self-Tuning Strategies.” 89.

Ramakrishnan, Naren, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, Gizem Korkmaz, Chris Kuhlman, Achla Marathe, Liang Zhao, Ting Hua, Feng Chen, Chang Tien Lu, Bert Huang, Aravind Srinivasan, Khoa Trinh, Lise Getoor, Graham Katz, Andy Doyle, Chris Ackermann, Ilya Zavorin, Jim Ford, Kristen Summers, Youssef Fayed, Jaime Arredondo, Dipak Gupta, and David Mares. 2014. “‘Beating the News’ with EMBERS: Forecasting Civil Unrest Using Open Source Indicators.” Pp. 1799–1808 in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York New York USA: ACM.

Resnik, Philip, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. “Beyond LDA: Exploring Supervised Topic Modeling for Depression-Related Language in Twitter.” Pp. 99–107 in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Denver, Colorado: Association for Computational Linguistics.

Shaw, Randy. 2013. “The Activist’s Handbook: Winning Social Change in the 21st Century - Randy Shaw - Google Books.” Retrieved March 30, 2022 (https://books.google.pl/books?hl=en&lr=&id=27owDwAAQBAJ&oi=fnd&pg=PR9&dq=Shaw,+R.+(2013).+The+activist%E2%80%99s+handbook:+Winning+social+change+in+the+21st+century.+Berkley,+CA:+University+of+California+Press.&ots=GFmXZRxiBf&sig=-_KcCSS289COgFNhZ3YiJgVVLKY&redir_esc=y).

Simpson, Brent, Robb Willer, and Matthew Feinberg. 2018. “Does Violent Protest Backfire? Testing a Theory of Public Reactions to Activist Violence.” *Socius: Sociological Research for a Dynamic World* 4:237802311880318. doi: 10.1177/2378023118803189.

Sloan, Luke, and Jeffrey Morgan. 2015. “Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter.” *PLoS ONE* 10(11):e0142209. doi: 10.1371/journal.pone.0142209.

Statista. 2022. “• Twitter: Most Users by Country | Statista.” Retrieved March 28, 2022 (https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/).

Sturmer, Stefan, and Bernd Simon. 2004. “Collective Action: Towards a Dual-Pathway Model.” *European Review of Social Psychology* 15(1):59–99. doi: 10.1080/10463280340000117.

Tausch, Nicole, Julia C. Becker, Russell Spears, Oliver Christ, Rim Saab, Purnima Singh, and Roomana N. Siddiqui. 2011. “Explaining Radical Group Behavior: Developing Emotion and Efficacy Routes to Normative and Nonnormative Collective Action.” *Journal of Personality and Social Psychology* 101(1):129–48. doi: 10.1037/a0022728.

Tufekci, Zeynep, and Christopher Wilson. 2012. “Social Media and the Decision to Participate in Political Protest: Observations From Tahrir Square.” *Journal of Communication* 62(2):363–79. doi: 10.1111/j.1460-2466.2012.01629.x.

Vision of Humanity. 2020. “Civil Unrest Around the World Has Doubled in Last Decade.” Retrieved April 25, 2022 (https://www.visionofhumanity.org/civil-unrest-on-the-rise/).

Wiederhold, Brenda K. 2020. “Social Media and Social Organizing: From Pandemic to Protests.” 23(9):2. doi: 10.1089/cyber.2020.0461.

Wires. 2022. “Dutch Police Disperse Thousands Protesting in Amsterdam against Covid Lockdown.” *France 24*. Retrieved August 12, 2022 (https://www.france24.com/en/europe/20220102-dutch-police-disperse-thousands-protesting-in-amsterdam-against-covid-lockdown).

van der Zwet, Koen, Ana I. Barros, Tom M. van Engers, and Peter M. A. Sloot. 2022. “Emergence of Protests during the COVID-19 Pandemic: Quantitative Models to Explore the Contributions

of Societal Conditions." *Humanities and Social Sciences Communications* 9(1):1–11. doi: 10.1057/s41599-022-01082-y.

8 Appendix

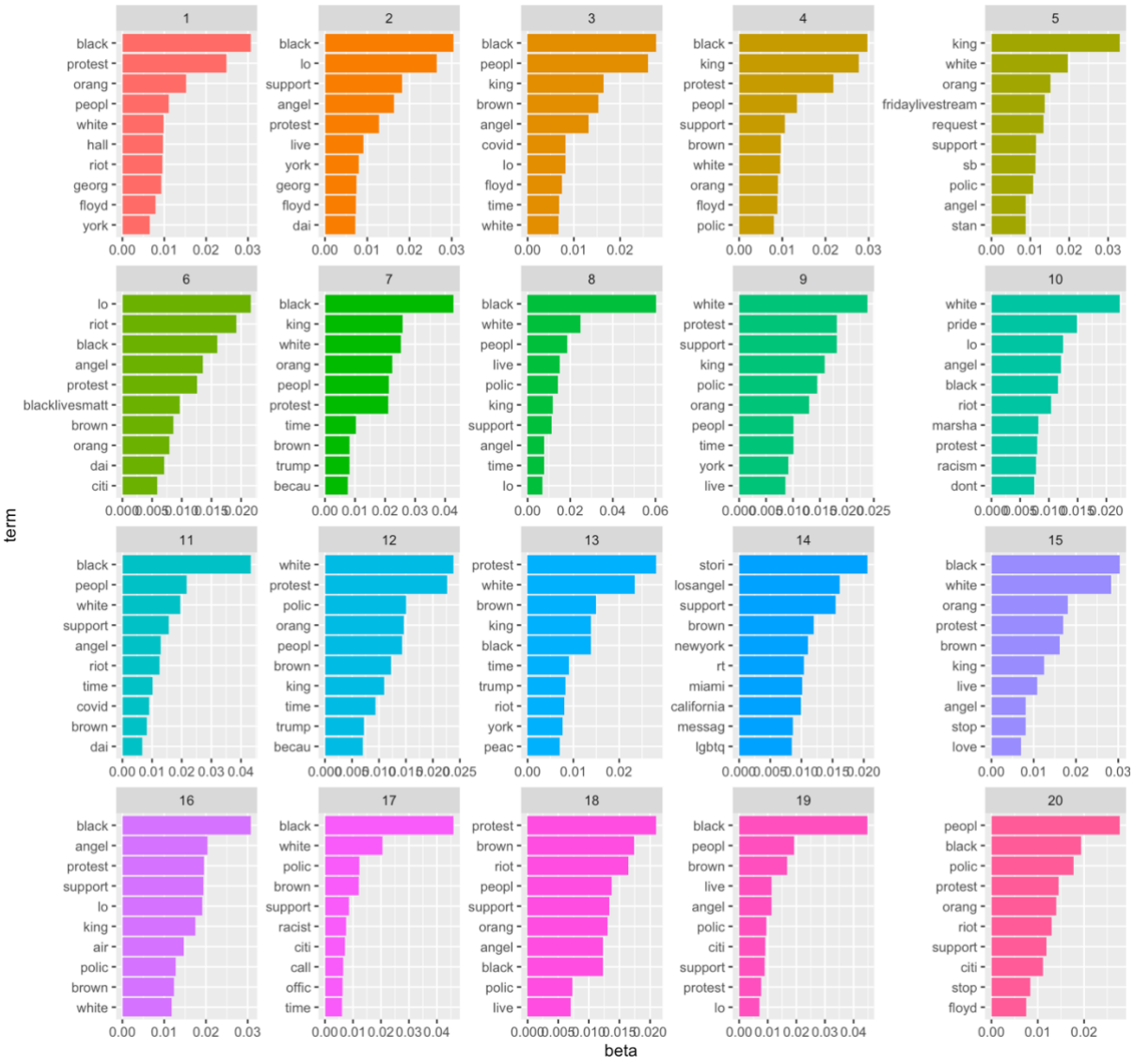
Appendix A:

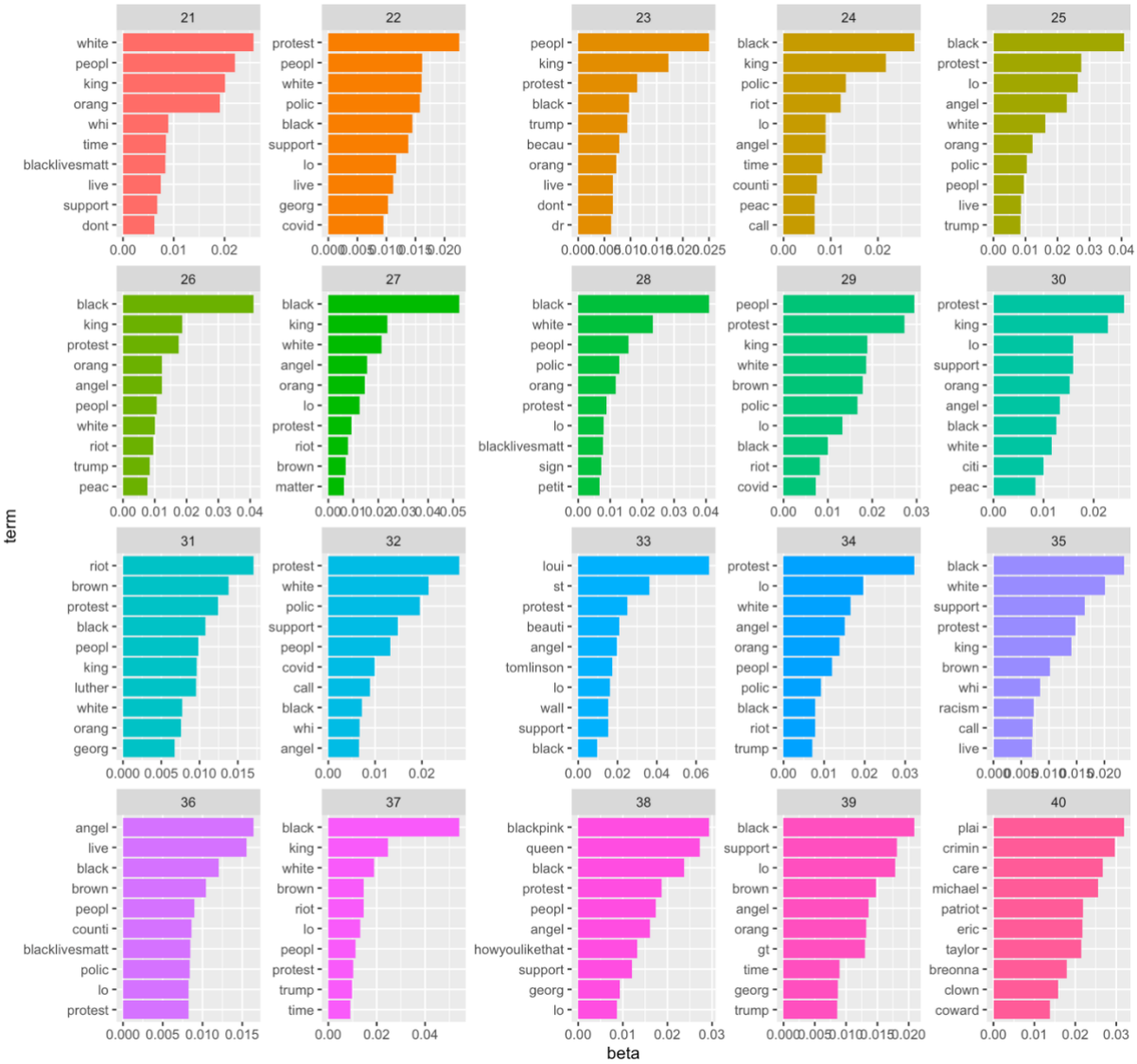
Filter terms used to identify protest-related tweets

Protest terms – *black, people, lives, police, georgefloyd, matter, protest, amp, white, racism, justice, support, protests, riot, riots, violence*
Protest hashtags – *#georgefloyd, #justiceforgeaorgefloyd, #nojuticenopeace, #breonnataylor, #cantbreathe, #blacklivesmatter, #blm, #protest, #defundthepolice, #allivesmatter, #justiceforbreonnataylor, #policebrutality, #protest2020, #protest2021, #maga, #metoo, #saytheirnames, #justice*

Appendix B:

Complete list of all 40 LDA classified topics from Tweet corpora




21
white
peopl
king
orang
whi
time
blacklivesmatt
live
support
dont
22
protest
peopl
white
polic
black
support
lo
live
georg
covid
23
peopl
king
protest
black
trump
becau
orang
live
dont
dr
24
black
king
polic
riot
lo
angel
time
counti
peac
call
25
black
protest
lo
angel
white
orang
polic
peopl
live
trump
26
black
king
protest
orang
angel
peopl
white
riot
trump
peac
27
black
king
white
angel
orang
lo
protest
riot
brown
matter
28
black
white
peopl
polic
orang
protest
lo
blacklivesmatt
sign
petit
29
peopl
protest
king
white
brown
polic
lo
black
riot
covid
30
protest
king
lo
support
orang
angel
black
white
citi
peac
31
riot
brown
protest
black
peopl
king
luther
white
orang
georg
32
protest
white
polic
support
peopl
covid
call
black
whi
angel
33
loui
st
protest
beauti
angel
tomlinson
lo
wall
support
black
34
protest
lo
white
angel
orang
peopl
polic
black
riot
trump
35
black
white
support
protest
king
brown
whi
racism
call
live
36
angel
live
black
brown
peopl
counti
blacklivesmatt
polic
lo
protest
37
black
king
white
brown
riot
lo
peopl
protest
trump
time
38
blackpink
queen
black
protest
peopl
angel
howyoulikethat
support
georg
lo
39
black
support
lo
brown
angel
orang
gt
time
georg
trump
40
plai
crimin
care
michael
patriot
eric
taylor
breonna
clown
coward
term
beta

Appendix C.1

Complete list of all one-grams used for the violent tweet classification model

1	peopl	100	they'r	201	cute	301	issu	401	save
2	fuck	101	doe	202	mind	302	fear	402	fan
3	love	102	understand	203	christma	303	prai	403	total
4	dai	103	boi	204	light	304	train	404	everydai
5	protest	104	block	205	guess	305	dad	405	solidar
6	time	105	ya	206	student	306	sometim	406	share
7	polic	106	matter	207	crowd	307	bui	407	fail
8	black	107	reason	208	store	308	yall	408	tast
9	shit	108	michael	209	haha	309	found	409	unit
10	whi	109	hear	210	racist	310	window	410	coff
11	feel	110	die	211	honest	311	bring	411	exact
12	white	111	lmao	212	hand	312	front	412	parti
13	live	112	racism	213	kong	313	road	413	washington
14	life	113	actual	214	commun	314	prove	414	experi
15	lol	114	wrong	215	meet	315	innoc	415	starbuck
16	realli	115	violenc	216	stupid	316	destroi	416	join
17	job	116	mad	217	hit	317	readi	417	chicago
18	brown	117	tear	218	sinc	318	funni	418	song
19	justic	118	umbrellarevolut	219	move	319	hot	419	project
20	kill	119	anoth	220	eat	320	oppress	420	till
21	riot	120	run	221	march	321	presid	421	book
22	cop	121	stand	222	respect	322	somebodi	422	direct
23	peac	122	wow	223	video	323	nn	423	event
24	tri	123	yeah	224	line	324	injust	424	due
25	stop	124	listen	225	read	325	power	425	moment
26	home	125	fight	226	dream	326	alert	426	creat
27	happi	126	verdict	227	action	327	west	427	send
28	tonight	127	ppl	228	stori	328	forc	428	drop
29	night	128	countri	229	learn	329	hell	429	lil
30	talk	129	shot	230	yo	330	market	430	nah
31	girl	130	hard	231	educ	331	beat	431	hella
32	world	131	post	232	angel	332	half	432	wit
33	plea	132	juri	233	sit	333	michaelbrown	433	angri
34	citi	133	act	234	left	334	gt	434	dude
35	hate	134	hour	235	wtf	335	shame	435	rock
36	someon	135	darren	236	histori	336	deal	436	travel
37	chang	136	pretti	237	probabl	337	dark	437	manag

38	start	137	photo	238	liter	338	pick	438	bout
39	wait	138	babi	239	ago	339	struggl	439	realiz
40	nigga	139	helicopt	240	scare	340	team	440	weekend
41	gui	140	final	241	center	341	perfect	441	mondai
42	watch	141	everyth	242	flag	342	em	442	fergsuon
43	york	142	protect	243	occupi	343	demonstr	443	everybodi
44	happen	143	cri	244	govern	344	report	444	plan
45	miss	144	safe	245	downtown	345	dinner	445	fun
46	everyon	145	trend	246	bro	346	sweet	446	accept
47	offic	146	head	247	opinion	347	grow	447	son
48	freewai	147	race	248	busi	348	hold	448	past
49	street	148	protestor	249	park	349	death	449	type
50	god	149	broadwai	250	togeth	350	kok	450	excit
51	bitch	150	dont	251	situat	351	mong	451	enjoi
52	care	151	class	252	cold	352	war	452	vote
53	sad	152	fridai	253	expect	353	throw	453	drink
54	friend	153	close	254	set	354	central	454	quot
55	system	154	hou	255	tweetmyjob	355	babe	455	bae
56	shoot	155	food	256	disgust	356	outsid	456	fat
57	hope	156	nice	257	heard	357	blame	457	fuckin
58	morn	157	anyon	258	gun	358	goe	458	possibl
59	real	158	we'r	259	announc	359	art	459	voic
60	ass	159	murder	260	bless	360	guilti	460	union
61	walk	160	obama	261	equal	361	continu	461	attack
62	damn	161	mom	262	text	362	serv	462	super
63	wanna	162	pic	263	agr	363	rai	463	kei
64	cau	163	new	264	truth	364	bai	464	station
65	bad	164	american	265	children	365	sundai	465	lunch
66	word	165	support	266	wake	366	south	466	pari
67	sleep	166	media	267	mayb	367	dure	467	anti
68	school	167	loot	268	rest	368	worth	468	selfi
69	call	168	mikebrown	269	thanksgiv	369	welcom	469	loud
70	tomorrow	169	indict	270	color	370	de	470	ugh
71	blacklivesmatt	170	omg	271	st	371	design	471	awesom
72	grand	171	hei	272	lose	372	dumb	472	month
73	veri	172	charg	273	pai	373	jail	473	piss
74	burn	173	leav	274	account	374	east	474	town
75	mani	174	hong	275	told	375	pl	475	import
76	plai	175	game	276	idea	376	women	476	earli
77	famili	176	sick	277	win	377	traffic	477	ud83dude29
78	person	177	heart	278	social	378	view	478	bore

79	crazi	**178**	bed	**279**	bullshit	**379**	finish	**479**	abl
80	occupycentr	**179**	lie	**280**	question	**380**	disappoint	**480**	public
81	london	**180**	anyth	**281**	begin	**381**	suppo	**481**	topic
82	stai	**181**	sound	**282**	shop	**382**	favorit	**482**	tv
83	week	**182**	shut	**283**	nobodi	**383**	health	**483**	idk
84	someth	**183**	occupyhk	**284**	minut	**384**	bruh	**484**	everywh
85	noth	**184**	break	**285**	hair	**385**	entir	**485**	glad
86	smh	**185**	human	**286**	lt	**386**	rule	**486**	lmfao
87	hurt	**186**	speak	**287**	surpri	**387**	deserv	**487**	sustain
88	befor	**187**	monei	**288**	squar	**388**	mine	**488**	ridicul
89	ignor	**188**	law	**289**	bodi	**389**	forget	**489**	singl
90	kid	**189**	sorri	**290**	pictur	**390**	woman	**490**	door
91	deci	**190**	folk	**291**	lo	**391**	futur	**491**	annoi
92	mike	**191**	phone	**292**	drive	**392**	san	**492**	skin
93	fire	**192**	amaz	**293**	admiralti	**393**	trust	**493**	workout
94	believ	**193**	check	**294**	cuz	**394**	movi	**494**	wine
95	america	**194**	littl	**295**	arrest	**395**	polit	**495**	tryna
96	free	**195**	rememb	**296**	tho	**396**	wor	**496**	coupl
97	beauti	**196**	late	**297**	build	**397**	brother	**497**	fall
98	car	**197**	nation	**298**	onc	**398**	ticket	**498**	exist
99	birthdai	**198**	dead	**299**	sign	**399**	justiceformikebrown	**499**	repeat
100	they'r	**199**	cool	**300**	trndnl	**400**	colleg	**500**	thousand

Appendix C.2

Complete list of all multi-grams used for the violent tweet classification model

1	thousand	41	juri.grand.juri	81	job.tweetmyjob	121	social.media	161	peac.protest	201	job.alert
2	sleep.sleep	42	mad.black	82	newyork.job	122	peopl.white	162	tear.ga	202	realli.peopl
3	sleep.sleep.sleep	43	peopl.black.black	83	newyork.ny.job	123	white.kill	163	admiralti.occupycentr	203	protest.block
4	nadi.nadi	44	peopl.black	84	ny.job	124	peopl.chang	164	occupycentr.umbrellarevolut	204	peopl.act
5	nadi.nadi.besara	45	peopl.mad.black	85	fuck.polic	125	chang.love	165	life.life	205	fuck.shit
6	nadi.nadi.nadi	46	ppl.riot	86	citi.hall	126	plea.love	166	peopl.dai	206	whi.fuck
7	gt.gt.lt	47	black.fridai	87	peopl.love	127	black.kill	167	meant.protect	207	chang.chang
8	gt.lt	48	live.matter	88	black.live	128	peopl.tri	168	system.fail	208	newyork.ny
9	gt.lt.fuck	49	lt.lt	89	everi.dai	129	occupyhk.umbrellarevolut	169	system.meant	209	dai.ago
10	gt.lt.lt	50	love.love	90	peopl.realli	130	protest.riot	170	system.meant.protect	210	X14th.broadwai.fergusonoakland
11	gt.lt.presid	51	black.black.crime	91	peopl.shit	131	polic.offic	171	system.protect	211	broadwai.fergusonoakland
12	fuck.fuck.cancer	52	black.crime	92	peopl.fuck	132	mong.kok	172	peopl.system	212	anti.occupi
13	lie.lie.lie	53	black.peopl	93	understand.peopl	133	peopl.stai	173	white.racist	213	understand.whi
14	white.cop.black	54	union.squar	94	talk.shit	134	stai.safe	174	topic.trndnl	214	peopl.feel
15	nadi.nadi.tan	55	god.bless	95	peopl.onli	135	peopl.protest	175	trend.topic	215	protest.cop
16	fuck.cancer	56	london.london	96	peopl.racist	136	peopl.walk	176	trend.topic.trndnl	216	riot.gear
17	lie.lie	57	hate.hate	97	peopl.polic	137	protest.walk	177	burn.flag	217	peopl.stand
18	black.black.black	58	san.francisco	98	peopl.noth	138	girl.becau	178	polic.polic	218	hong.kong
19	fuck.cancer.cancer	59	lo.angel	99	peopl.peopl	139	protestor.polic	179	pic.u2014	219	shit.fuck
20	fuck.cancer.fuck	60	nigga.nigga	100	peopl.white.peopl	140	occupycentr.hongkong	180	polic.protestor	220	notjustferguson.rt
21	million.fuck.cancer	61	black.live.matter	101	peopl.talk	141	polic.protest	181	alert.trend	221	nigga.shit
22	black.black	62	fuck.nigga	102	healthcar.job	142	protest.broadwai	182	trend.alert	222	everi.time
23	nadi.nadi.mi	63	justic.system	103	nur.job	143	mike.brown	183	trend.alert.trend	223	cop.peopl
24	nadi.nadi.rico	64	whi.alwai	104	live.world	144	protest.brown	184	trend.trend	224	hate.peopl
25	fight.fight	65	black.fuck	105	job.job	145	love.miss	185	love.shit	225	washington.dc
26	black.mad.black	66	peopl.black.peopl	106	job.job.tweetmyjob	146	justic.peac	186	american.flag		

27	white.kill.black	**67**	cop.black	**107**	love.plea	**147**	protest.peac	**187**	burn.american		
28	white.white.black	**68**	kill.black	**108**	follow.love	**148**	protest.polic	**188**	burn.american.flag		
29	fuck.fuck	**69**	white.black	**109**	chicago.il	**149**	protest.peopl	**189**	X14th.fergusonoakland		
30	grand.juri	**70**	peopl.whi	**110**	causewai.bai	**150**	citi.center	**190**	sustain.notosofitelday85n		
31	mad.black.black	**71**	whi.peopl	**111**	polic.ga	**151**	mani.peopl	**191**	post.photo		
32	black.peopl.black	**72**	darren.brown	**112**	becau.fuck	**152**	whi.doe	**192**	chang.peopl		
33	whi.whi	**73**	michael.brown	**113**	black.matter	**153**	dai.dai	**193**	cop.kill		
34	white.white	**74**	block.freewai	**114**	peopl.becau	**154**	protest.freewai	**194**	freewai.protest		
35	hei.ho	**75**	whi.black	**115**	peopl.care	**155**	brown.famili	**195**	walk.freewai		
36	hei.ho.supremaci	**76**	cop.shoot	**116**	reason.whi	**156**	realli.fuck	**196**	peopl.justic		
37	hei.ho.white	**77**	black.white	**117**	happi.birthdai	**157**	X14th.broadwai	**197**	justic.serv		
38	fuck.fuck.fuck	**78**	trend.trndnl	**118**	law.enforc	**158**	peopl.live	**198**	peopl.freewai		
39	white.peopl	**79**	fuck.hate	**119**	black.cop	**159**	protest4th	**199**	polic.line		
40	grand.juri.juri	**80**	peopl.scare	**120**	peopl.mad	**160**	polic.car	**200**	atlanta.ga		