Predicting Social Unrest using Sentiment Analysis

MSc Data Science: Business & Governance

Master thesis

Ufuk Ozdemir

SNR: 1243192

ANR: 751409

Supervisor: dr. H.J. Brighton

Second reader: dr. Sharon Ong

Tilburg University,

School of Humanities

Tilburg, The Netherlands

December 2018

Abstract

Through the years many events related to politics, economy and demography have led to social unrest. The most recent is the ongoing Yellow Vests movement in France that caused uproar among citizens and lead to heavy demonstrations. The Global Database of Events, Language and Tone (GDELT) collects news related events from all over the world. This thesis investigates to what extent it is possible to make accurate predictions regarding sentiment about non-violent and violent protest events in Europe using the GDELT database. Also is it possible to achieve similar model performance by using smaller datasets. Lastly, the models were tested on unseen data from the Unites States in 2005 to see to what extent the models can be generalized. For this, a time series was used with a sliding window method. The best performing model was the random forests for non-violent and violent protests. Results show that the events from the GDELT are useful in predicting tone of non-violent and violent protests with all examined sizes of train sets. It can be concluded that there is a relationship between news media sentiment and protest events, which can be predicted to some extent.

Table of contents

Chapter 1: Introduction

Through the years political or economic related decisions led to many social unrest events in societies. Social unrest is a broad term that includes protests, strikes, riots or occupation of areas and can be defined as public disturbance caused by a gathering of at least three or more people, in reaction to an event or decision to raise awareness against injustice (Cadena, Kormaz, Kuhlman, Marathe, Ramakrishnan, & Vullikanti, 2015). Social unrest usually involves damage to property, economic loss, and can lead to civilian injury or death. A recent example of social unrest is the 'Gilets Jaunes' (Yellow Vests) movement in France that started on 17 November 2018. The rise of fuel taxes, higher living costs and the unreasonable burden of the government's tax reforms were at the cost of the working and middle classes. The social unrest is still ongoing and led to protests, blocking of traffic, looting stores, vandalism, barricades, 1000 civilian casualties and 4 civilian deaths (Yellow vests movement, 2018) . Such events usually need to be planned or organized in some way. A prime example of how social unrest can be planned or organized are the riots in the United Kingdom (UK) in 2011, which made use of RIM's Blackberry Messenger service to communicate. The features of the Blackberry Messenger (BBM) make it possible to create chatgroups, send locations and send timed messages. This lead to organized riots that started in the Tottenham area of London and spread to Birmingham, Manchester and many other areas afterwards (Benkhelifa, Rowe, Kinmond, Adedugbe, & Welsh, 2014).

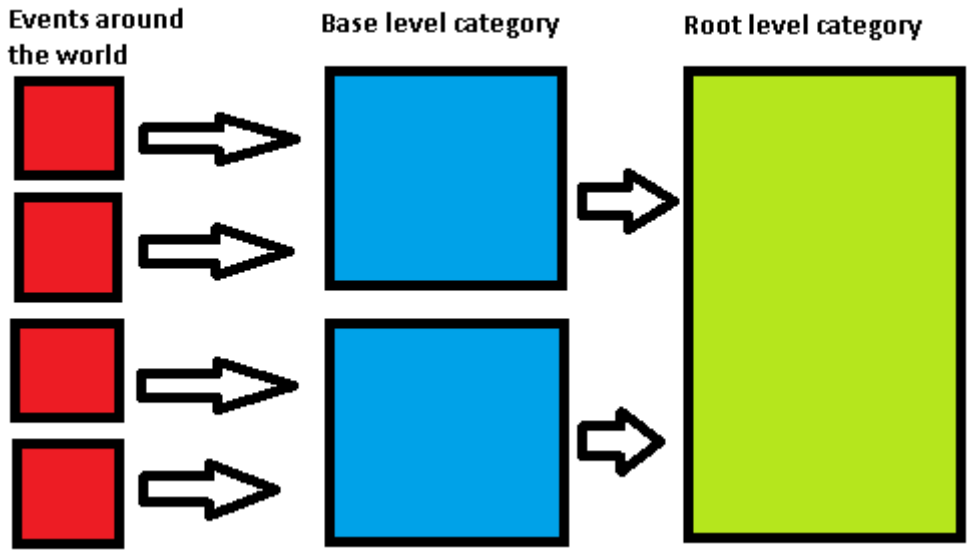
Another tool that is used to plan and organize events are social networks like Twitter, Tumblr and Facebook. These social networks contain an abundance of data (volume), with different data types such as text, images and video (variety), and the data is updated in real-time and changes constantly (velocity). A lot of people share their sentiment (attitudes, opinions, views and emotions) about certain topics on social network platforms, which gives the opportunity to study various social factors and events that take place in society (Galla & Burke, 2018). The sentiment of people towards certain topics change over time and can be used for sentiment analysis to find useful indicators relating to social unrest events (Mbunge, Vheremu & Kajiva, 2015).

The prediction of social unrest associated to political or economic changes can be beneficial for policy makers or decision makers, but also has implications for industrial, governmental or logistic companies. For example, if the protests and riots in France were detected sooner, the damage to property and civilian casualties could have been minimized. A lot of literature is available on predicting social unrest using data from social networks.

Compton, Silva and Macy (2013) tried to predict social unrest using the sentiment from Twitter, by filtering on keys words, dates and locations that associate to social unrest and from this demographics were identified that could be used to predict future unrest events. Similarly, Xu, Lu, Compton and Allen (2014) focused on predicting social unrest from Tumblr, also by using a simple filter-based method to identify posts related to social unrest.

While social networks are popular for sentiment analysis, less has been done with news media such as the Global Database of Events, Language, and Tone (GDELT). The GDELT collects news reports about 300 different types of events from all over the world since 1979. For each event within the GDELT, attributes such as the information about actors, location, sources and tone are identified and stored in a format called Conflict and Mediation Event Observations (CAMEO). The CAMEO has two levels of categories: root level and base level. The root level consists of 20 categories and the base level consists of specific categories that fall under the root level. For example, specific events such as protests, riots and strikes fall under the base level, and together these specific events fall under social unrest, which is the root level (Leetaru & Schrodt, 2013). Figure 1 illustrates the structure of the CAMEO categories.

Figure 1. *Events captured around the world are first put in specific categories (base level), and from there into broad categories (root level).*



Therefore, this thesis aims to explore the relationship between social unrest and tone using the GDELT over a period of 18 years. Galla and Burke (2018) made an attempt in predicting social unrest on State and County level in the United States using the GDELT. The daily

frequency of several categories related to social unrest were used to classify whether such event would occur or not. Qiao et al. (2017) also used the GDELT to predict social unrest in Southeast Asia by classifying the developmental stages first and making predictions based on temporal burst patterns. Additionally, Qiao et al. (2017) mentions that the features in the GDELT have an temporal order which can be used for time series methods. When dealing with time series data, the sequential observations are dependent which means that the order of data has to be taken into account (Mahalakshmi, Sridevi, & Rajaram 2016). To our current knowledge, no other study has been done on social unrest with a time series approach. Therefore, this approach is chosen for the current research.

Both Galla et. al. (2018) and Qiao et al. (2017) focused on the root level of social unrest, which consists of specific events such as protests or riots. Additionally, Galla and Burke (2018) mentioned that using base level categories could give a better representation of social unrest, since the root level category also contains less relevant events. A large change in tone caused by increasing frequency counts of base level events could be an indication that an protest event might happen in future. Therefore, this research will distinguish itself by making use of specific events in the base level category to predict non-violent and violent protest events within Europe (EU). The dependent variable non-violent protest is the tone from the base event *Demonstrate or rally* and the other dependent variable violent protest is the tone from the base event *Protest violently or riot.* From this, the following two research questions are formulated:

RQ 1: *To what extend is it possible to predict tone of non-violent protest events within Europe by using the GDELT?*

RQ 2: *To what extend is it possible to predict tone of violent protest events within Europe by using the GDELT?*

It is possible that protest related events will not occur every day within a country and also because this research focuses on specific events, it might be the case that there is insufficient data. To know whether it is possible to predict generalizable results with smaller sizes of train sets, the following question is formulated:

RQ 3: *To what extent do smaller training datasets give generalizable results?*

Lastly, to see whether the models also perform well on countries outside EU, a new unseen dataset will be downloaded. The unseen data is from the United States (US) 2005. Hence, the final question:

RQ 4: *To what extend can results of the models be generalized to the United States?*

Chapter 2: Prior work

The use of sentiment to predict protest events can be important for decision makers to take action on time to minimize damage. Several studies have investigated predicting social unrest using social networks like Twitter, Facebook, blogs and Tumblr (Korkmaz et al., 2016; Qiao & Wang, 2015; Mbunge, Vheremu & Kajiva, 2015; Mishler, Wonus, Chambers & Bloodgood, 2017). Most of these studies did a sentiment analysis and used the frequency of words to predict whether a social unrest event would take place or not. Additionally, a combination of social networks are usually used to validate and improve performance. However, the use of new media is less compared to social networks, and only two studies are found that use the news media from GDELT as only source to predict social unrest.

Qiao et al. (2017) build Hidden Markov Models (HMMs) to predict indicators that are related to social unrest events in Southeast Asia. The used root level categories from GDELT are the following: *Disapprove*, *Demand*, *Reject*, *Threaten* and *Protest*. First, the developmental stages of social unrest were identified and after that a prediction was based on temporal burst patterns of these developmental stages. The best performing model was the HMM and for the baseline a logistic regression was used. From the findings it can be concluded that the GDELT dataset does contain useful preceding indicators that reveal the causes or development of future social unrest events.

Galla and Burke (2018) tried to predict the probability of social unrest events in the United States on state and county level. The prediction is based on the frequency of events from different categories in the GDELT. The event categories used from the GDELT are also on root level and almost similar to Qiao et al., (2017), but instead of the *Demand* and *Reject* category, Galla and Burke (2018) used *Coerce* and *Assault*. The reasoning for this was that Galla and Burke (2018) wanted to add a violent aspect of social unrest, while Qiao et al., (2017) mainly focused on the non-violent aspects of social unrest. The model that predicted most accurately on both state and county level is the random forest model. Results show that the GDELT can be used to predict social unrest events one month ahead.

Both Galla et al. (2018) and Qiao et al. (2017) used the root level categories in the GDELT, which also contain less relevant variables for the prediction of social unrest, and with this reason Galla and Burke (2018) suggest that using base level categories might give a better representation in the prediction of social unrest.

To investigate this issue, this study will focus on the base level category of non-violent and violent protest events. Non-violent protests here are represented by their verbal

character only, which can involve shouting or verbal threatening, while violent protests involve property damage, physical injuries and arrests of civilians. Additionally, another difference that this study will add is predicting the *tone* of protests by using the base level category events.

Chapter 3: Data Pre-Processing and Cleaning

3.1 GDELT Database Description

The GDELT is a real time database that updates every 15 minutes and collects news media events from all over the world in over 100 languages since 1979 until present time. Over 300 types of physical activities around the world are recorded in this database which vary from riots and protests to diplomatic exchanges or the georeferenced of a city. From every unique event approximately 60 attributes are captured, which include information about the location, the actors that are involved and the action that has been used. These events then are classified in Conflict and Mediation Event Observations (CAMEO) format, which is an event coding scheme optimized for the study of third party mediation in international disputes (Leetaru & Schrodt, 2013).

3.2 Extracting from Database

From the GDELT event table the following 4 fields are used: SQLDATE, EventBaseCode, AvgTone and ActionGeo_CountryCode. SQLDATE is the date in YYYYMMDD format. EventBaseCode is the base level category, which falls under the root level. For example, the code 1411 (demonstrate or rally for leadership change) falls under the base code 141 (demonstrate or rally, not specified otherwise) and the base code falls under the root code 14 (PROTEST). ActionGeo_CountryCode refers to the location of the event, which is a 2-character FIPS10-4 country code for the location. AvgTone is the average tone of all documents containing one or more mentions of an event. The scores from the AvgTone ranges from -100 (extremely negative) to +100 (extremely positive).

The tone of an event can be seen as an indicator that gives information about the impact of an event. Moreover, the AvgTone only refers to the first news report mentioning the event, which means that the AvgTone score will not be updated when an event occurs in multiple news reports. Also important to mention is that the tone must be interpreted with caution because it is not a measurement of emotion (GDELT, 2015). In this study, the change in tone will be an indicator of protest events that might happen in the future.

Before extracting the data from GDELT, a few filters were applied to get smaller subsets of data. The first filter was set to extract the following base level events: *Disapprove, Criticize or denounce, Complain officially, Demonstrate or rally, Protest violently or riot, Use repression.* Table 1 shows the root level to which these events belong and their description.

Table 1. *The root and base level categories of the CAMEO and the description of the base level events.*

Root level	Base level	Description event
Disapprove	Disapprove, if not specified	Express disapprovals, objections, and complaints if not specified.
Disapprove	Criticize or denounce	Condemn, decry a policy or an action; criticize, defame, denigrate responsible parties.
Disapprove	Complain officially	Written and institutionalized protests, appeals, and all petition drives and recalls.
Protest	Demonstrate or rally, if not specified	Dissent collectively, publicly show negative feelings or opinions; rally, gather to protest a policy, action, or actor(s).
Protest	Protest violently or riot, not specified	Protest forcefully, in a potentially destructive manner, if not specified.
Coerce	Use repression	Actively repress collective actions of dissent by forcing subjugation through crowd control tactics and arrests.

The second filter was set to select 10 European countries, which are shown in Table 2. After applying these filters on a time frame from 1995 to 2013, a dataset was constructed that consisted of 1.033.138 rows and 4 columns.

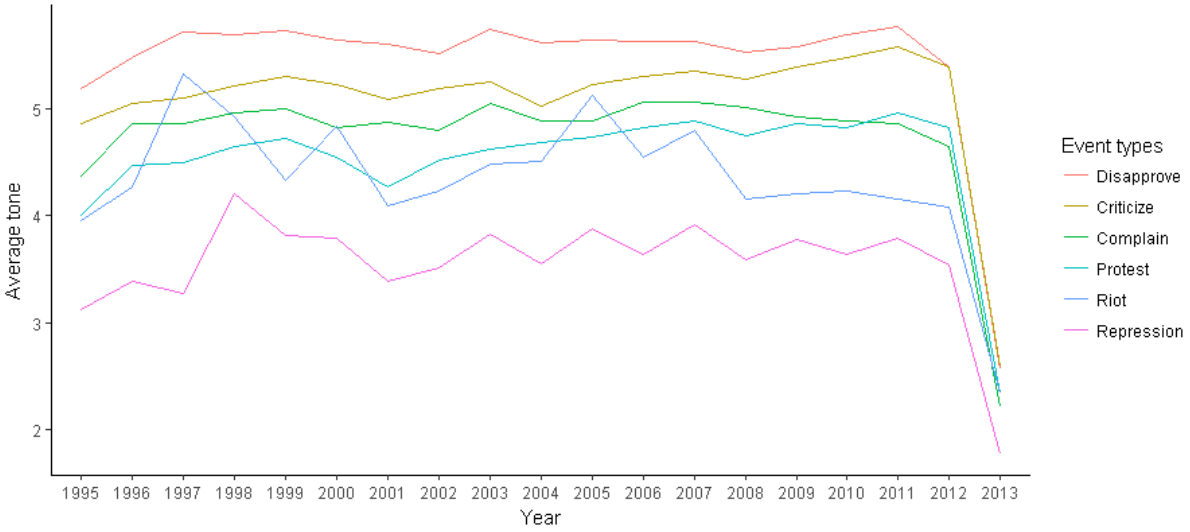
Table 2. *Selected European countries and their 2-character FIPS10-4 code.*

1. Belgium (BE)	6. Netherlands (NL)
2. Ireland (EI)	7. Spain (SP)
3. France (FR)	8. Sweden (SW)
4. Germany (GM)	9. United Kingdom (UK)
5. Italy (IT)	10. Greece (GR)

3.3 Normalizing data

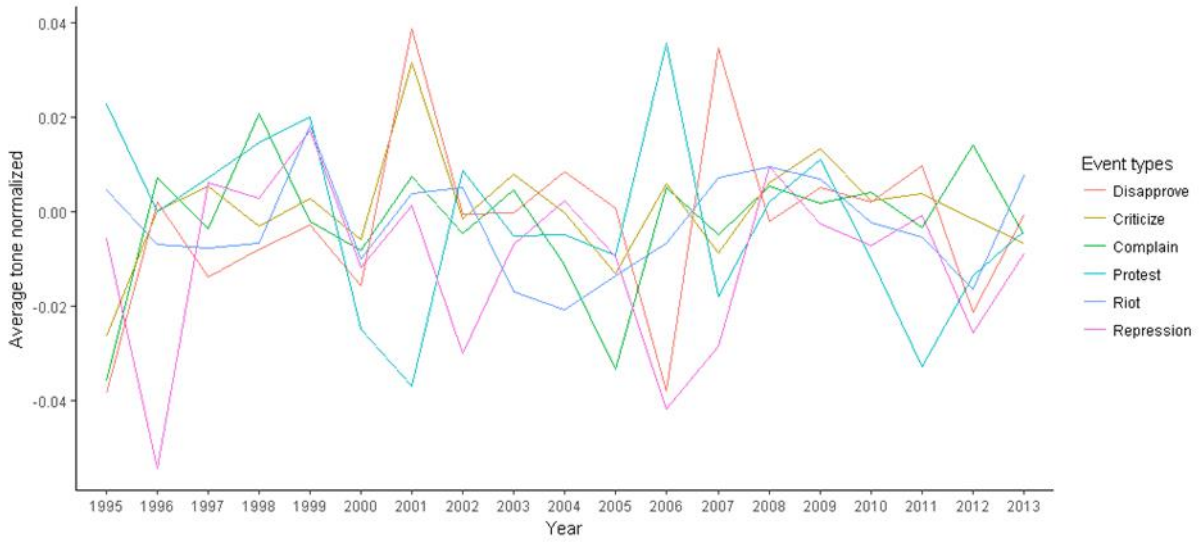
The GDELT's Average Tone (AvgTone) variable declines around the timeframe of 2012-2013. This happens for every selected base level event from the CAMEO. The most likely explanation for this is that GDELT changed the way of scaling tone. This change was also mentioned by Kumar, Benigni and Carley (2016), but the reasoning behind this change in scale remains unclear. Additionally, the average tone score does not fall below 0 while the highest score does not exceed 28. Figure 2 shows the decline of tone around 2012-2013.

Figure 2. *The yearly average tone for each event category before normalization. Around 2012-2013 the average tone declines.*



To resolve the issue of the declining tone data after the year 2012, a normalization of data was applied. First the grand mean of each year and each event category was calculated. After that, the grand mean was subtracted from the daily tone corresponding to each same year and each event category. Aggregating the daily tone to yearly for each event category gives the following results, which can be seen in figure 3.

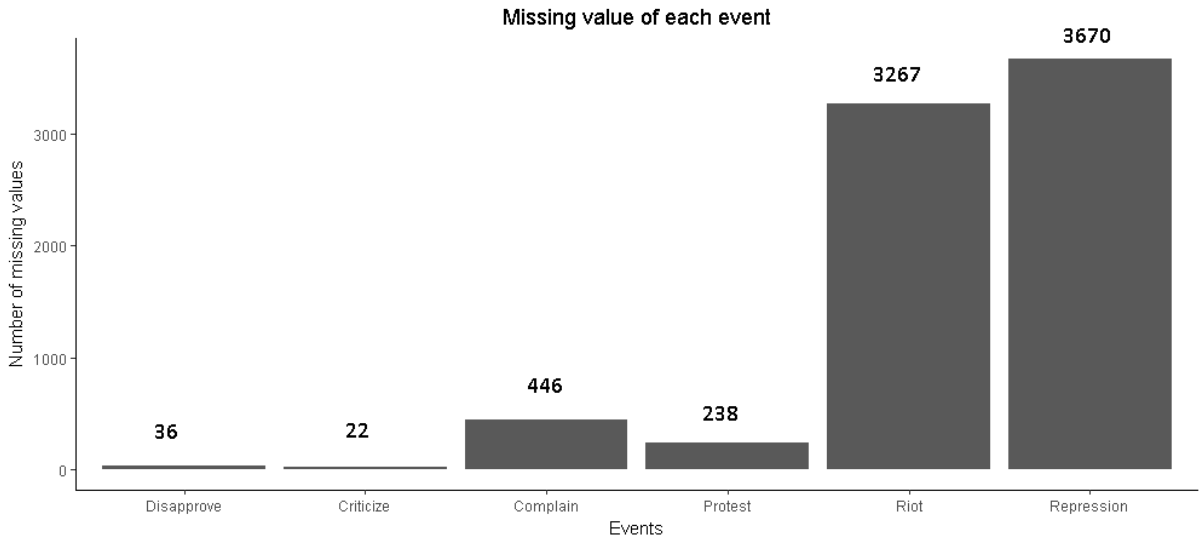
Figure 3. *The yearly average tone for each event category after normalization.*



3.4 Missing values

Not every country had an event happening on each day, which means that there will be some gaps in the data. If each event was aggregated for each country, then the missing values could not be replaced, because there were consecutive missing values of almost 2 weeks. This is resolved by aggregating on the event categories instead, which still gave a missing value total of 7679. Table 4 summarizes the missing values for each event after aggregation.

Figure 4. *The missing values for each event after aggregating on the event categories.*



To resolve the problem after aggregating on events, the missing tone value of the previous day and the next day are taken, added together and finally divided by 2, which yields a mean

value for the missing day. This was done for both the tone and the frequency for each event category. Table 3 illustrates how the missing values are replaced.

Table 3. *Replacing missing values by the mean of the adjacent observations.*

Date event	Missing tone data	Missing tone data replaced
2008-01-02	2.584	2.584
2008-01-03	*NA*	2.381
2008-01-04	2.178	2.178

3.5 Transformation of categorical features and data aggregation

The events selected from the CAMEO are categorical features that need to be converted into continues values. The dataset had1.033.138 instances after extraction from the GDELT. A single day in the dataset could for example have 10 protest related events and 6 repression related events. New columns have been created for each event that contain the frequency of daily events by aggregating on days, which led to 41640 instances. This study will use weeks to predict the tone of protests, which means that the daily frequency of events and tone need to be aggregated to weekly data. This is done for each year, from 1995 to 2013 and resulted into 1008 instance. Each year has 52 weeks here.

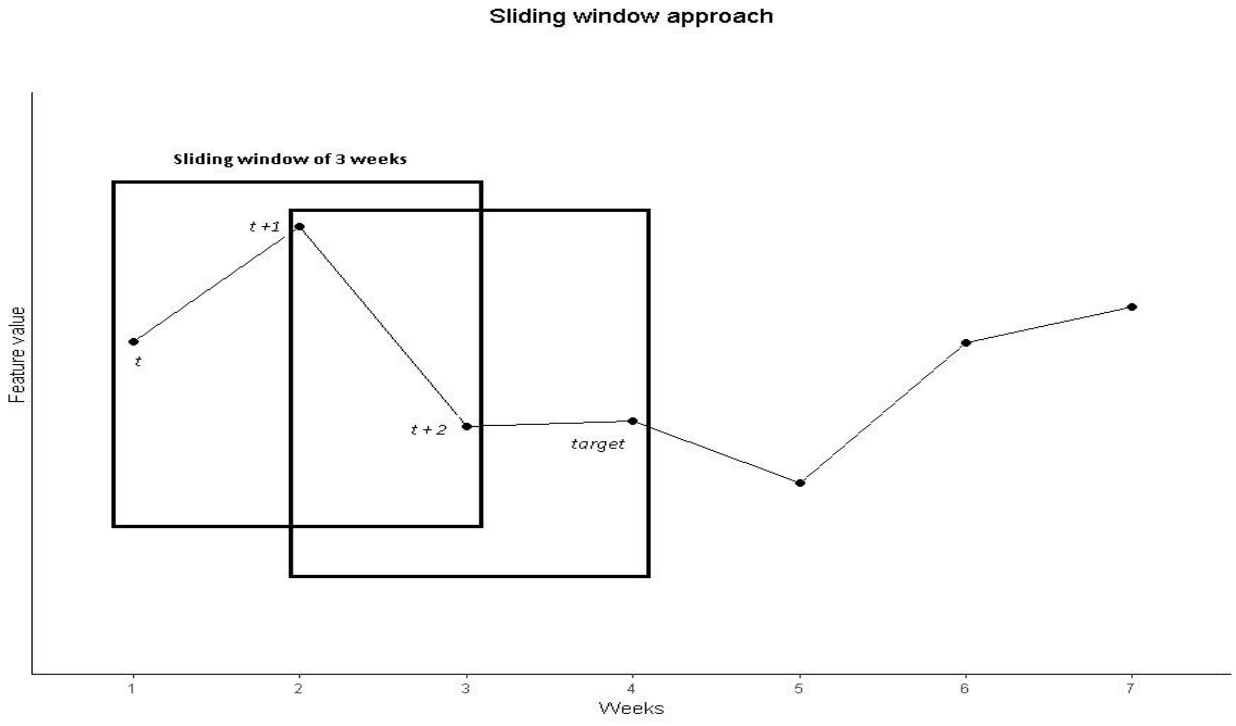
Chapter 4: Experimental Setup

4.1 Sliding window method

The features in the dataset consist of weekly intervals from the year 1995 to 2013. For time series the order in data is important because the observations from the past and the future are dependent on each other. With the sliding window method, a training set will be constructed for each experiment. The idea is that the sliding window goes over a data sequence of lets say for the event *Protest*, and extracts features like the mean and variance.

Adding these new features can enhance the prediction performance of the outcome, but it can also increase the dimensionality of the data and lead to overfitting (Noorian, Moss & Leong, 2014). The sliding window has a fixed length of 3 weeks, in which the first week is represented as *t*, the second as $t + 1$ and third week as $t + 2$. Then the window slides from left to right and shifts one week ahead, keeping the second ($t + 1$) and third ($t + 2$) week, while adding the predicted week (*target*), and this repeats itself for all weeks. The extracted features from the tone and events are discussed later but are summarized for now as following: mean, variance, range, minimum, maximum and the median. Figure 5 illustrates how the sliding window approach works.

Figure 5. *Illustration of a 3-week sliding window approach, in which t, t+1, and t+2 are summed and divided by 3 to predict the target.*



4.2 Feature extraction and selection

In order to get more data, the feature values of the tone and events will be used to construct additional features. The tone that will be predicted in this study is from the event feature *Protest* and *Riot*. The event features that will be used to predict protest tone are: *Disapprove, Criticize, Complain, Protest, Riot,* and *repression.* In total there are 6 event features from which the tone of Protest and Riot will be predicted. The additional features will extracted by applying the sliding window method, as mention in section 4.2 and illustrated in figure 5.

From each feature, the following additional features will be extracted: mean, variance, range, minimum, maximum and the median. This is done by extracting at the time point t, $t + 1$, and $t + 2$ for each window of 3 weeks. This results in 6 additional features for every event feature, which is a total of 36 additional features and 6 additional features for every 2 tone feature, which is a total of 12 additional features. Thus, 48 additional features are added to current dataset, which consists of 1008 instances and 54 features.

The reasoning for constructing extra features is that additional information extracted from the original data can improve the prediction performance in the experiments (Guyon & Elisseeff,, 2003). Therefore, expectation is that these additional features contributed in predicting tone of non-violent and violent protest events. The selection of the features is done automatically by the algorithms themselves and the way the they do this is described in section 4.6.

4.3 Baseline

To determine the performance of the algorithms, a baseline is needed for comparison. This baseline will act as a threshold and once this threshold is exceeded by the other algorithms, it will mean that they performed well (Brownlee, 2016). The persistence algorithm will be applied here, because time series data is being used. This is done by taking the tone value at a previous time step ($t - 1$) and used to predict the outcome for the next time step (t). For example, this study takes the first 3 weeks ($t - 1$, $t - 2$ and $t - 3$) of tone data to predict the 4^{th} week (t).

4.4 Evaluation criteria

This study predicts numbers based on the used models in each experiment, which will have errors in different sizes. The Mean Squared Error (MSE) will be used to evaluate the performance of the models. The way the MSE does this is by taking the distances from the points of the regression line and squaring them. A lower MSE indicates that a model is performing better. In order to have an indication of whether a model performs better than the baseline, the MSE of the model should be lower than the MSE of the baseline.

4.5 Algorithms

This section will describe the reasoning for the used algorithms. According to Wolpert and Macready (1997), no algorithm performs the best in every task or situation, which means that no algorithm fits on every type of problem. It may be the case that an algorithm performs well in one particular situation and poorly in another. For this reason, a several regression algorithms have been selected to predict the tone of protest events. The five algorithms selected for this study are: Linear Regression, Ridge, Lasso, Elastic Net and Random Forests Regression.

A *linear regression* model is used to find a relationship between a dependent variable and one or more continues predictors. The values of the dependent variable can be estimated from the observed predictor values. The main idea of linear regression is to find the line that best fits the data, which is done by using the least squares approach to minimize the residuals sum of square errors (Schneider, Hommel & Blettner, 2010).

The *ridge regression* is a shrinkage method that penalizes the size of the coefficients with L2-regularization, which minimizes the residual sum of squares. These coefficients are defined by the complexity parameter, where a higher value causes more shrinkage on the coefficients that makes the model less complex. The advantage of ridge regression over Ordinary Least Squares (OLS) lies in the bias-variance trade off. The OLS minimizes the sum of squared residuals of its estimates, which are often unbiased but they do have large variance that affects the prediction performance negatively. This is why the ridge regression allows for more bias in order to reduce the variance, which might lead to lower MSE (Hoerl & Kennard, 1970; James, Witten, Hastie & Tibshirani, 2013). While the ridge regression shrinks the coefficients and is more stable, it does not set the coefficients to zero and makes the interpretation of the model more difficult (Tibshirani, 1996).

This problem is resolved by using the *lasso regression* by putting a constraint on the sum of the absolute values of the model parameters, which has to be less than the fixed value.

The lasso penalizes the coefficients by shrinking the coefficients and this leads to estimates towards zero. Only the variables that still have a non-zero coefficient after the feature selection process, will be selected to be a part of the model. The main goal of this process is to minimize the prediction error (Fonti & Belitser, 2017).

Another shrinkage method is the *elastic net* that is a combination of both L1 and L2 regularization, which also minimizes the sum of squared error. The difference is that the L1 regularization adds a constraint on the sum of the absolute values of the coefficients, while the L2 regularization adds a constraint on the sum of squares of the coefficients (Zou & Hastie, 2005).

The last method used is the *random forests regression*, which Galla and Burke (2018) also used in their study to predict social unrest. The random forests is a form of ensemble learning technique that can be used for either classification or regression problems. It consists of multiple decision trees in which every tree node specifies a condition on one feature and splits the dataset in two parts. The way the random forest minimizes the MSE is by selecting the best split at the tree node. Also the random forest performs a feature selection and provides feature importance that will help in understanding which features have a significant impact. The selection of importance is done with the Mean Decrease Impurity (MDI), in which a higher MDI indicates a higher importance of a feature (Galla & Burke, 2018; Breiman, 2001).

Chapter 5: Experiments

5.1 Experiment 1: Predicting non-violent protest tone

The goal in the first experiment is to predict the tone values of non-protests events 3 weeks ahead using the sliding window method. All event features (see Table 1) will be used and their additional features that were extracted from them. The data will be split into a training set that consists of 813 (80%) instances and the test set consists of 195 (20%) instances, while keeping the order of time which is a requirement for the time series. All the regression models will be using the sliding window method to learn to predict 3 weeks ahead by taking the *t, t* + 1 and *t* + 2 to predict the *target*, as illustrated in figure 5. The target here is the tone value of non-violent protests after an interval of 3 weeks. Parameter tuning of the models is done automatically by the R-packages caret, RandomForest and Glmnet, in which the best parameter is selected.

5.2 Experiment 2: Predicting violent protest tone

The same procedure is applied here as in question 1, but this time for violent protest events.

5.3 Experiment 3: Predicting with less train data

Not everyday protest related events will happen within a country, which means that there will be less data to make predictions. Also because the protest related events that this study examines are specific, which means that it is likely that there will be insufficient data. This is why experiment 3 will also predict the tone value of non-violent and violent protests using the same procedure as question 1 and 2, but with smaller train sets. The baseline of the 20% test set from question 1 and 2 will be used to compare model performance. First, a train set of 60% (605 instances) will be used and tested on the 20% test set. The next train set consists of 40% (403 instances) data and will also be and tested on the 20% test set. The purpose of this experiment is to see whether smaller train sets give generalizable results on the same test set.

5.4 Experiment 4: Generalize models on US data

The models have been tested on EU countries to predict the tone value of protest events, now the question remains whether the models can be generalized to other countries. For this a new unseen dataset is downloaded from the GDELT database, which contains the tone of non-violent and violent protests in the US from the year 2005. Pre-processing is done in the same way as previous data and gave 52 instances, which represent the weeks in a year. This unseen

data will act as the test set upon which the train set (80%) in question 1 and 2 will be tested. Additionally, the same models will be used and also here the parameter tuning is done automatically by the R-packages.

Chapter 6: Results

6.1 Results experiment 1: Predicting non-violent protest tone

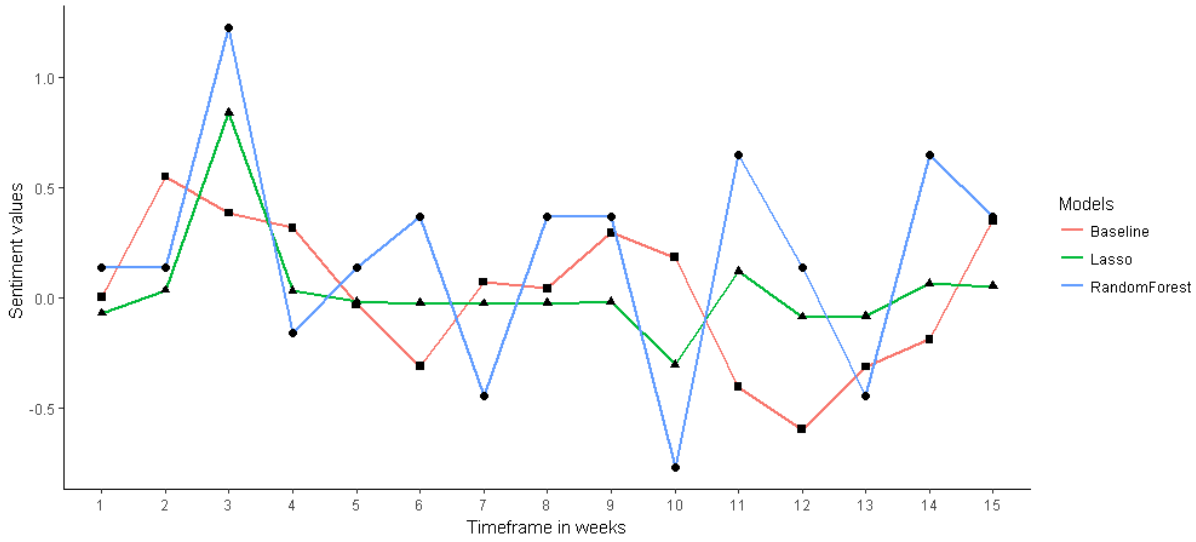
The goal of this experiment was to predict the tone value 3 weeks ahead for non-violent protests. Five models were used for this purpose and the proportions of train set were 80% and the test set was 20%. The performance of the models are evaluated by comparing the MSE with the baseline, in which a lower MSE indicates a better model performance. The results for the non-violent protests in experiment 1 show that all models outperform the baseline in the test phase. The best performing model on the test phase is the random forests with an average MSE of 0.01, which is a large difference of 0.36 (97%) compared to the baseline. The second best model is the lasso with an average MSE of 0.12, and the difference compared to the baseline is 0.25 (68%). Noticeable is that the lasso seems to perform less with 0.07 MSE in the test phase compared to train phase. The performance of the remaining models compared to the baseline vary with a MSE score between 0.11 (62%) to 0.23 (30%), which is a small to medium difference. These findings indicate that the event features from the GDELT are useful to predict the tone of non-violent protests 3 weeks ahead well, with a train set of 80%. Table 4 represents all models and their performance on the train and test set.

Table 4. *Average model performance of non-violent protests on 80/20% train/test sets.*

Non-violent protest events		
Models	**80% train MSE**	**20% test MSE**
Baseline	0.36	0.37
Linear	0.30	0.25
Ridge	0.15	0.14
Lasso	0.05	0.12
Elastic Net	0.28	0.26
Random Forest	0.04	0.01

Figure 3 shows the visualization of the best performing models predicting the first 15 weeks for the tone of non-protest events. The best models are compared with the baseline on a 20% test set.

Figure 3. *Predictions of the first 15 weeks in the 20% test set for non-violent protests tone. The two best performing models are compared with the baseline.*



6.2 Results experiment 2: Predicting violent protest tone

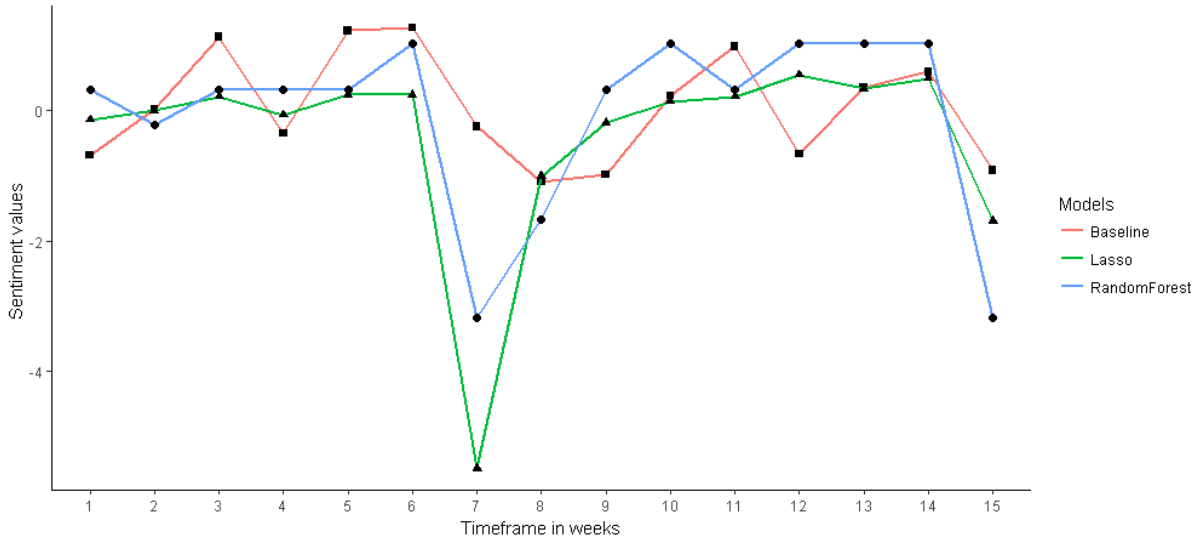
The results experiment 2 also show that all models outperform the baseline in the train and test phase for the violent protests. The best performing models in the test phase are random forests and lasso with an average MSE of 0.41 (70%) and 0.71 (48%). Compared with the baseline, this a difference in MSE score of 0.95 for the random forest and 0.65 for lasso. The performance of the remaining models compared to the baseline vary with a MSE score between 0.09 (6%) and 0.56 (41%), which is a small to medium difference. Also here can be concluded that the event features used from the GDELT are useful to predict the tone of violent protests 3 weeks ahead, with a train set of 80%. Table 5 contains all MSE scores for the train and test phase for violent protest events.

Table 5. *Average model performance of violent protests on 80/20% train/test sets.*

Violent protest events		
Models	**80% train MSE**	**20% test MSE**
Baseline	1.38	1.36
Linear	1.00	1.23
Ridge	0.54	0.80
Lasso	0.36	0.71
Elastic Net	1.02	1.27
Random Forest	0.33	0.41

The visualization of the best performing models predicting the tone of violent protests for the first 15 weeks can be seen in figure 4. Also here the best models are compared with the baseline on a 20% test set.

Figure 4. *Predictions of the first 15 weeks in the 20% test set for violent protests tone. The two best performing models are compared with the baseline.*



6.3 Results experiment 3: Predicting with less train data

Earlier it is already mentioned that protest related events do not occur every day, which means that there will be less data to make predictions from. For example, not having enough data to predict certain weeks ahead could be an issue for logistic companies. Not being able to detect these protest events means that logistic companies cannot redirect their cargo in time to prevent unexpected losses. Therefore, the goal in experiment 3 was to see if less train data could give similar results on the 20% test set from question 1 and 2, for non-violent and violent protests. The used train set proportions are 60% and 40%. The procedure was the same as experiment 1 and 2, but with smaller train sets.

The results for the non-violent protests show that all models outperform the baseline on the test phase with a 60% train set. This is also the case with a 40% train set. The random forests and lasso are the best performing models when the 60% and 40% train set is being used. The average MSE score of the random forest on the 60% train set is 0.05 and on the 40% set 0.07. For the lasso the average MSE on the 60% train set is 0.12 and on the 40% set 0.14. Compared to the baseline, the random forests perform 86% better on the 60% train set and 81% on the 40% set. This suggest that even with smaller train sets generalizable model performance can be achieved, which means that the tone non-violent protests can be predicted with less data. Table 6 represents all models MSE's scores of non-violent protests for the 60% and 40% train sets, which are used on the 20% test set from question 1.

Table 6. *Model performance when train set is 60% and 40% on the 20% test set of non-violent protests.*

	Non-violent protest events	
Models	**60% train set**	**40% train set**
Baseline	0.37	0.37
Linear	0.28	0.28
Ridge	0.16	0.15
Lasso	0.12	0.14
Elastic Net	0.30	0.29
Random Forest	0.05	0.07

Figure 5 shows the best performing models in predicting the first 15 weeks for the tone of non-protest events. The best models are trained on a train set proportion of 60% and are compared with the baseline of the 20% test set.

Figure 5. *Predictions of the first 15 weeks based on 60% train set for non-violent protests tone. The two best performing models are compared with the baseline from the 20% test set.*

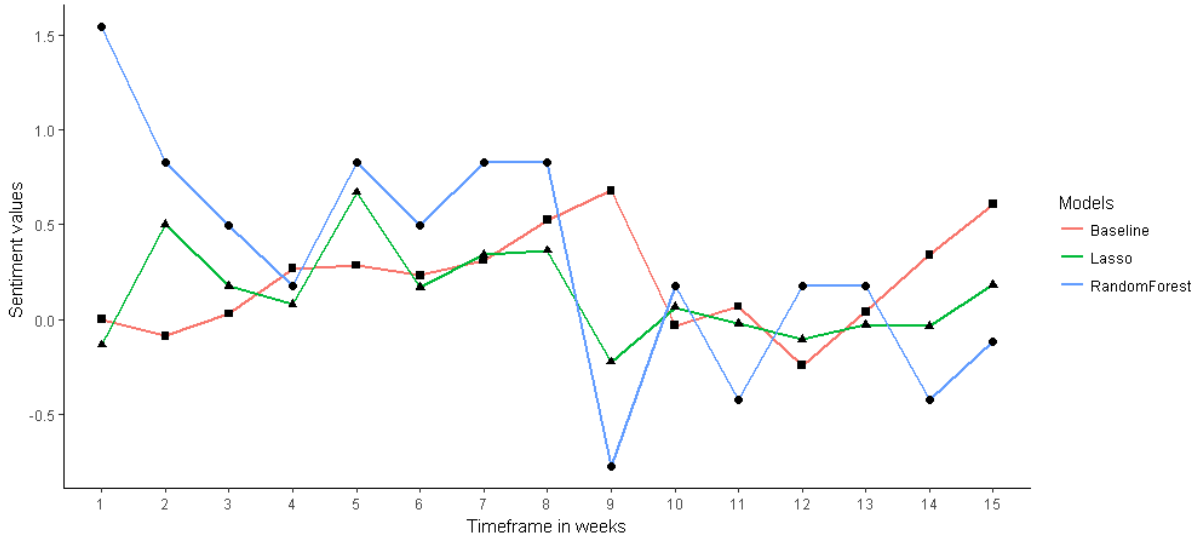
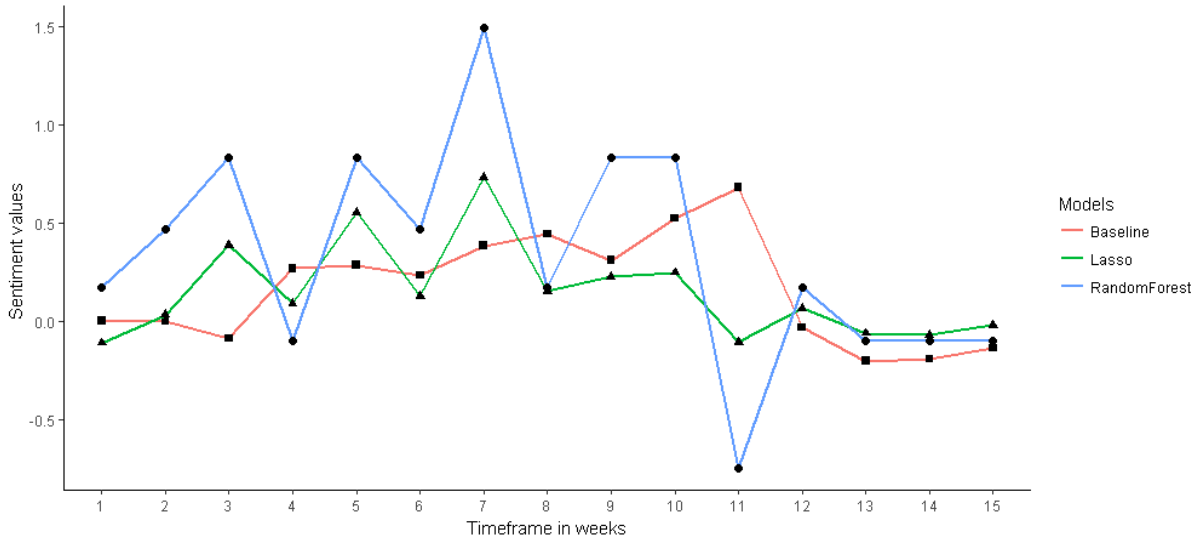


Figure 6 shows the best performing models in predicting the first 15 weeks for the tone of non-protest events. The best models are trained on a train set proportion of 40% and are compared with the baseline of the 20% test set.

Figure 6. *Predictions of the first 15 weeks based on 40% train set for non-violent protests tone. The two best performing models are compared with the baseline from the 20% test set.*



The results for the violent protests show that all models outperform the baseline on the test phase as well with a 60% train set. This is also the case with a 40% train set. Also here the random forests and lasso are the best performing models when the 60% and 40% train set is

being used. The average MSE score of the random forest on the 60% train set is 0.47 and on the 40% set 0.39. For the lasso the average MSE on the 60% train set is 0.49 and on the 40% set 0.42. Compared to the baseline, the random forests perform 65% better on the 60% train set and 71% on the 40% set. Surprising is that the random forests performs 11% better on a 40% train set, which was not the case on non-protests. This suggests that violent protests might be predicted better than non-violent protests with smaller train sets. Also here the conclusion is that the tone can be predicted with less data.

Table 6. *Model performance when train set is 60% and 40% on the 20% test set of violent protests.*

	Violent protest events	
Models	**60% train set**	**40% train set**
Baseline	1.36	1.36
Linear	1.07	1.03
Ridge	0.63	0.55
Lasso	0.49	0.42
Elastic Net	1.08	1.05
Random Forest	0.47	0.39

Figure 7 shows the best performing models in predicting the first 15 weeks tone of protest events. The best models are trained on a train set proportion of 60% and are compared with the baseline of the 20% test set.

Figure 7. *Predictions of the first 15 weeks based on 60% train set for violent protests tone. The two best performing models are compared with the baseline from the 20% test set.*

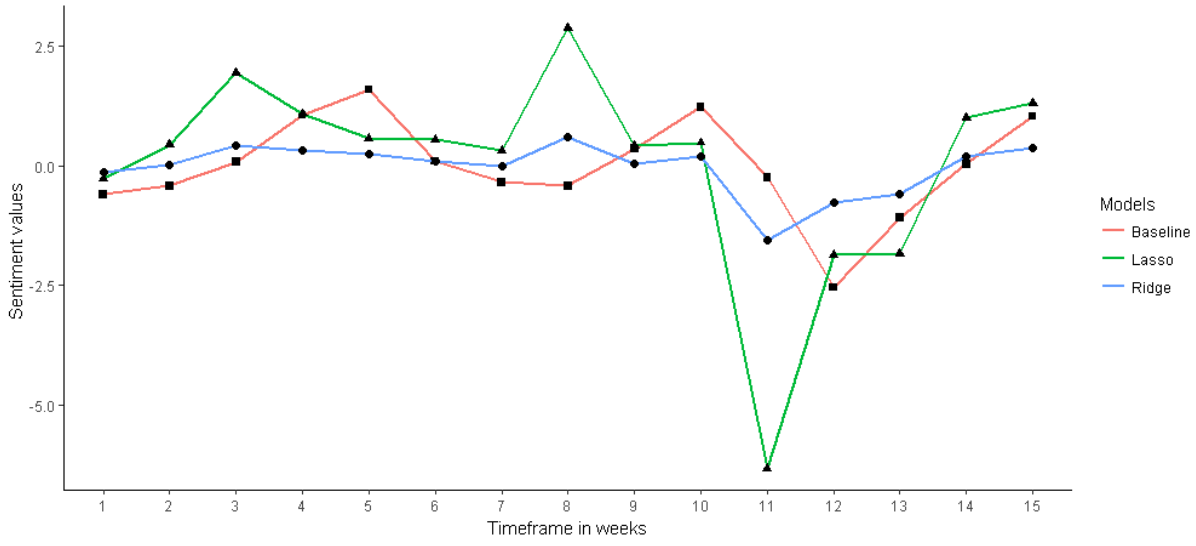
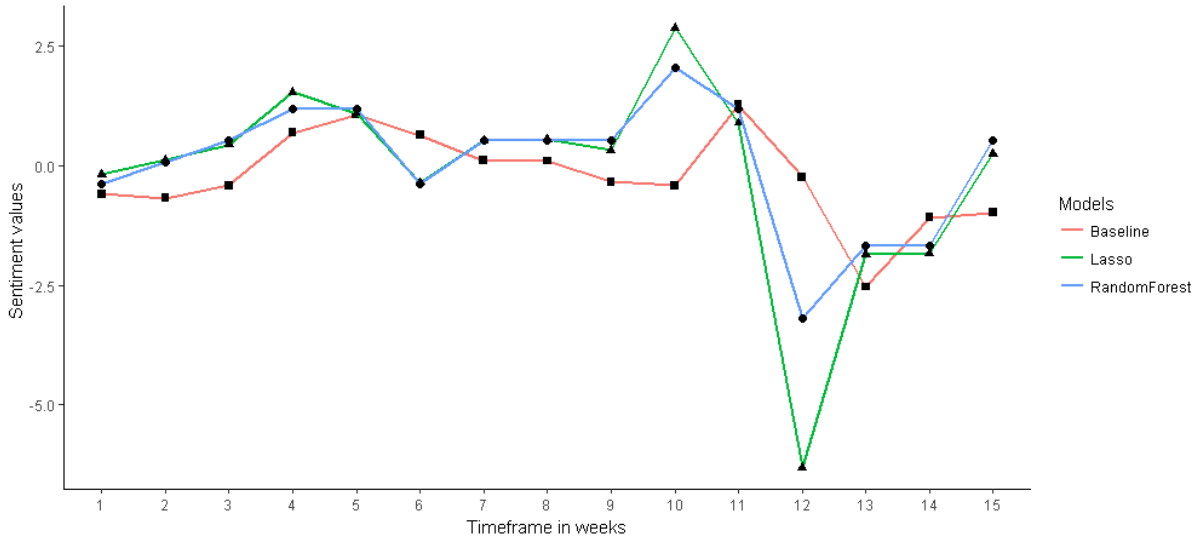


Figure 8 visualizes the best performing models in predicting the first 15 weeks tone of protest events, but this time the models are trained on a train set proportion of 40% and are compared with the baseline of the 20% test set.

Figure 8. *Predictions of the first 15 weeks based on 40% train set for violent protests tone. The two best performing models are compared with the baseline from the 20% test set.*

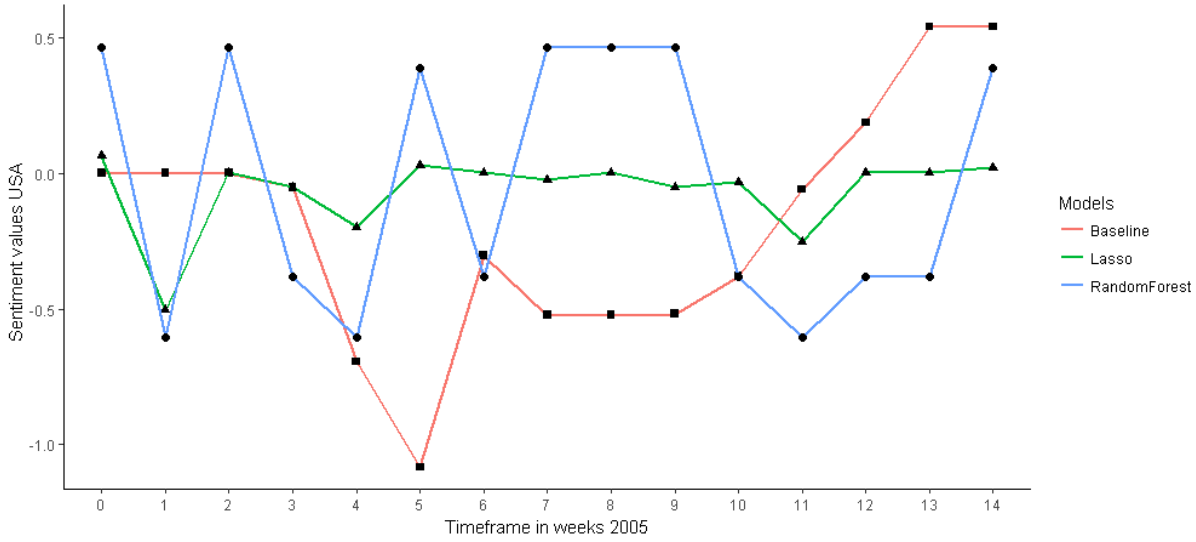


6.4 Results experiment 4: Generalize models on US data

The main goal in experiment 4 is to test the generalizability of the models in predicting the tone of protest events, but this time on US data. The prediction is based on the data of 2005. However, only non-violent protest events will be used, because the violent protest variable had not enough data and replacing the gaps of time was not possible. Some gaps were of 5 consecutive days, which would be hard to replace and also might bias the results. For this reason only the non-violent protest events will be used. The results for the tone value of non-violent protests in the US show that all models perform poorly compared to the baseline. The used features and train set (80%) were the same as in question 1. The best performing model in this case was the random forests with a average MSE score 0.50. However, the fact still remains that the train set from EU does not generalize well on non-violent protests in the US. From these finding it can be concluded that the event features from EU are not good in predicting the tone of non-violent protests in the US.

Figure 9 shows the two best performing models in predicting the first 15 weeks tone of non-protest events in the US. The used train set is from question 1 tested on unseen data from the US 2005.

Figure 9. *Predictions of the first 15 weeks of non-violent protests tone from the United States 2005 data. The models are trained on the train set of question 1 and do not outperform the baseline.*



Chapter 7: Discussion

The early detection of protest events can help the government or companies to minimize escalating actions or financial loss by taking measures. For example, if the Yellow vests movement in France was detected earlier, the damage to properties and civilian casualties could have been minimized. A several studies have already examined the prediction of social unrest and the study of Galla and Burke (2018) is the most recent one. However, they have used a classification approach to predict social unrest, while this study uses time series with sliding window method to predict. Galla and Burke (2018) mentioned in their study that examining the base level categories of the CAMEO might give more insight to specific categories of events. To build further upon predicting social unrest and to fill this gap, this study chose to focus on more specific types of social unrest. Since non-violent and violent protests occur more often in EU, it gave the idea to examine whether it possible to predict these and led to the following research questions in this study:

RQ 1: *To what extend is it possible to predict tone of non-violent protest events within Europe by using the GDELT?*

RQ 2: *To what extend is it possible to predict tone of violent protest events within Europe by using the GDELT?*

RQ 3: *To what extent do smaller training datasets give generalizable results?*

RQ 4: *To what extend can results of the models be generalized to the United States?*

7.1 Discussion experiment 1: non-violent protests

The goal of the first experiment was to predict the tone value of non-violent protest events within EU. A large change in the tone value would act as an indicator for possible non-violent protest events in the future. The sliding window was used to predict 3 weeks ahead and the best model was random forests that performed 97% better than the baseline on a train set proportion of 80%. In the study of Galla and Burke (2018) also the random forests performed the best, although they used a different approach in predicting. This suggests that the random forests are suitable algorithms in predicting social unrest, although the question remains if they are well suited for other social factors. The results in experiment 1 showed that the event features from the GDELT could be useful in predicting the tone value for non-violent protests.

However, Galla and Burke (2018) did not mentioned that the use of specific events also needs to deal with more missing values, since some events do not occur every day. The raw data was first aggregated on days and after that on event categories, which resulted in lots of missing values for some events. The two events which had the most missing values were *Riots* and *Repression.*

One explanation for this is that riots are characterized by violent forms of actions, such as damaging property or physical fights. Repression is actively suppressing certain collective actions by forcing subjugation through crowd control tactics and arrests, which usually happen during riots. Because these two events seem to relate to each other and do not occur frequently, might explain the cause why the missing values for these are large.

7.2 Discussion experiment 2: Violent protests

The goal of experiment 2 is the same as question but this time to predict the tone value for violent protests events within EU. Also here, all models outperformed the baseline with random forest being the best model and performing 70% better. However, one noticeable thing here is that the random forests perform better on the train phase, which performs 41% better than in the test phase. This is the case for all models, which might be connected to the missing values of events that relate to violent protests. As mentioned in section 7.1, the events *Riots* and *Repression* had the most missing values. From the results in experiment 2 can be concluded that the event features from the GDELT could be useful in predicting the tone value for violent protests as well.

7.3 Discussion experiment 3: Predicting with less train data

It is already mentioned that some events had missing values due to that certain events do not occur every day. Sometimes policy changes can cause a uproar in society and lead to massive protests and riots, such as recently in France. If these social unrest events could have been foreseen, then the damage could have been minimized. However, there was not much data available from the time of policy changes to social unrest. Therefore, the goal of experiment 3 was to examine whether it was possible to get generalize results with the use of less data. For this train set proportions of 60% and 40% were used and tested on the 20% test set from question 1 and 2. All models outperform the test set baseline for non-violent protests on both the 60% and 40% train sets. The random forests perform the best with both train set sizes, with a performance of 86% on the 60% train set and 81% on the 40% train set. This suggest that smaller train set decrease the performance of the models, but are still able to predict the

tone value.

Further, also all models outperform the test set baseline for violent protests on both the 60% and 40% train sets. Again the random forests perform the best with both train set sizes, with a performance of 65% on the 60% train set and 71% on the 40% train set. However, expected was that the smaller the train data becomes, the less the performance of the models would be. Surprisingly, the random forests performed 6% better with 20% less train data. This might be explained due to the missing values that are now affecting the performance of the models. From the findings it could be concluded that a 60% train set is still able to predict the tone of violent protests using the feature events from the GDELT. However, the results from the 40% train set must be interpreted with caution, because the results might be biased due to missing values that start to affect the performances of the models.

7.4 Discussion experiment 4: Generalize models on US data

The purpose of experiment 4 was to examine whether the models used in experiment 1 and 2 would give similar predictions on a different population. For this a new unseen dataset is downloaded from the GDELT database, which contains the tone of non-violent and violent protests in the US from the year 2005. However, due to many missing values the violent protests tone was dropped and only the non-violent protests tone was used. The train set from question 1 was used on the unseen US data test set, in which all models perform poor when compared to the baseline. The best model here was random forests as well, although it did not outperform the baseline. These findings suggest that the models trained on the event features from the EU do not generalize well on the US. This means that the event features are not useful to predict the tone value of non-violent protests of populations outside the EU.

The explanation for this result is that is could be the case that not many protest activities occurred during the year 2005. To validate this statement, the list of civil unrests on Wikipedia was used (List of incidents of civil unrest, 2018). Two major events were found that caused an uproar in society in 2005. The first one is that after the hurricane Katerina, civilians started to massively looting stores in order to survive. The second major event is the National Socialist Movement (NSM) in which a protest was planned against African-American gang activity in the North End of Toledo. This led to riots, vandalism, looting and 12 injured civilians. It could be that remaining events during the year 2005 were not as impactful as these two which led to the poor performance of the models. When the normalized tone data for non-violent protests is visualized, it also shows 2 peaks around

August and October, which seem to correspond to the two major events that happened in 2005.

7.5 Limitations and future research

The current study has encountered a several limitations that could play a role in the results. The first limitation is the change of average tone between the years 2012 and 2013, in which the observed values suddenly drop to negative values. This was also mentioned in the study of Kumar, Benigni and Carley (2016), but the reasoning behind this change in scale remains unclear. To deal with this problem the data was normalized by first calculating the grand mean for each event and year, then the grand mean is subtracted from the corresponding events and years.

Another limitation is that this study used specific variables to predict the tone of protests. The CAMEO categories in the GDELT database has 20 categories of events on root level and these consist of base level categories, which are more specific events. By choosing from these specific events this study limited itself by having less data and led to more missing values. Some variables had missing values for few consecutive days which is problematic when using time series. To resolve this problem and minimize the loss of information, a specific R package (*ImputeTS*) was used to automatically calculated and replace the missing gaps. However, it is questionable how the R package calculates missing gaps when there are consecutive missing days.

Furthermore, this study originally had the intentions to examine how the tone would change for each country within EU, but once the data was aggregated on event types per country the missing day gaps were too large. Some of these gaps were larger than 2 weeks, therefore it is almost impossible to replace these missing values, without being biased. The reason for these gaps can be explained by specific events that do not occur every day within a country or are mentioned every day on the news, which is why the GDELT will have missing values for those days.

Despite of having limitations in this study, it also provides opportunities for future research. Since this is the first study that has done a time series with a sliding window on the GDELT data to predict protests, future researchers can be recommend to focus on the root level of the CAMEO categories because selecting specific events is prone to missing values. This way the change in tone of countries can be compared, while having no or a small amount of missing values when the data is aggregated. Since the root level provides enough variables, the missing day gaps should be relatively small.

Another thing that can be considered to examine in future research is how the width of the sliding window can affect the prediction performance. For this study a sliding window with a width of 3 weeks was used to predict the next one. It could be interesting investigate whether the prediction performance will change when a shorter or longer width of weeks are used.

Chapter 8: Conclusion

This study examined how protest related events in the CAMEO contribute to the change in tone of protest events within EU, and for this a time series with a sliding window method was applied. The GDELT database was used and event data from 1995 to 2013 was extracted. The first and second question concerned whether it was possible to predict the tone values of non-violent and violent protests based on selected events that related to protests. The best model for both was the random forests. Findings show that the selected events (disapprove, criticize, complain, protest, riot and repression) can indeed be used to predict the tone value of non-violent and violent protests 3 weeks ahead. Question three examined to what extend it was possible get generalizable results using smaller train sets. With train sets of 60% and 40% the random forests performs the best for both non-violent and violent protests. Results show that events from the GDELT can be useful to predict the tone value, but interpretation should be done with caution on the 40% train set part. Lastly, the results show that the models were not generalizable on the US tone data of 2005 for non-violent protests. This could be explained that not many events in 2005 had a large impact on the change of tone. Only 2 major events happened in 2005 that led to two peaks in the change of tone.

References

Benkhelifa, E., Rowe, E., Kinmond, R., Adedugbe, O. A., & Welsh, T. (2014). Exploiting Social Networks for the Prediction of Social and Civil Unrest: A Cloud Based Framework. *2014 International Conference on Future Internet of Things and Cloud*, 565-572. doi:10.1109/FiCloud.2014.98

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. doi:10.1023/A:1010933404324

Brownlee, J. (2016). How to Make Baseline Predictions for Time Series Forecasting with Python. Retrieved from https://machinelearningmastery.com/persistence-time-series-forecasting-with-python/

Cadena, J., Korkmaz, G., Kuhlman, C. J., Marathe, A., Ramakrishnan, N., & Vullikanti, A. (2015). Forecasting social unrest using activity cascades. *PLoS ONE*, *10*(6). doi:10.1371/journal.pone.0128879

Compton, R., Lee, C., Xu, J., Artieda-Moncada, L., Lu, T., Silva, L. D., & Macy, M. (2014). Using publicly visible social media to build detailed forecasts of civil unrest. *Security Informatics*, *3*, 1-10. doi:10.1186/s13388-014-0004-6

Compton, R., Lee, C., Lu, T., Silva, L. D., & Macy, M. (2013). Detecting future social unrest in unprocessed Twitter data: "Emerging phenomena and big data". *2013 IEEE International Conference on Intelligence and Security Informatics*, 56-60.

Fonti, V., & Belitser, E. N. (2017). Paper in Business Analytics Feature Selection using LASSO.

Galla, D., & Burke, J. (2018). Predicting Social Unrest Using GDELT. *Machine Learning and Data Mining in Pattern Recognition*. doi:10.1007/978-3-319-96133-0_8

GDELT (2015). *GDELT data format codebook V2.0.* Retrieved from http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3(3), 1157-1182. Retrieved from: http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf

Hoerl, A. E., & Kennard, R. W. (1970) Ridge Regression: Biased Estimation for nonorthogonal Problems. Technometrics, 12, 55-67.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with applications in R (Vol. 103). New York: Springer.

Korkmaz, G., Cadena, J., Kuhlman, C. J., Marathe, A., Vullikanti, A. K., & Ramakrishnan, N. (2016). Multi-source models for civil unrest forecasting. *Social Network Analysis and Mining*, *6*, 1-25.

Kumar, S., Benigni, M., & Carley, K. M. (2016, September). The impact of US cyber policies on cyber-attacks trend. In Intelligence and Security Informatics (ISI), 2016 IEEE Conference on (pp. 181-186). IEEE. Retrieved from: http://www.casos.cs.cmu.edu/publications/papers/2016ImpactofUSCyber.pdf

Leetaru, K., & Schrodt, P. A. (2013, April). GDELT: Global data on events, location, and tone, 1979–2012. Paper presented at the ISA Annual Convention. Retrieved from http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf

Li, L., Noorian, F., Moss, D. J., & Leong, P. H. (2014). Rolling window time series prediction using MapReduce. *Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference,* 13-15

List of incidents of civil unrest. (2018). Retrieved December 15, 2018, from https://en.wikipedia.org/wiki/List_of_incidents_of_civil_unrest_in_the_United_States #2000 %E2%80%932009

Mahalakshmi, G., Sridevi, S., & Rajaram, S. (2016). A survey on forecasting of time series data. 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), 1-8.

Mbunge, E. (2017). A Tool to Predict the Possibility of Social Unrest Using Sentiments Analysis - Case of Zimbabwe Politics 2017 - 2018.

Mishler, A., Wonus, K., Chambers, W., & Bloodgood, M. (2017). Filtering Tweets for Social Unrest. *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, 17-23.

Qiao, F., & Wang, H. (2015). Computational Approach to Detecting and Predicting Occupy Protest Events. *2015 International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI)*, 94-97.

Qiao, F., Li, P., Zhang, X., Ding, Z., & Cheng, J. (2017). Predicting Social Unrest Events with Hidden Markov Models Using GDELT.

Schneider, A., Hommel, G., & Blettner, M. (2010). Linear regression analysis: part 14 of a series on evaluation of scientific publications. Deutsches Arzteblatt international, 107(44), 776-82.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society.* Series B (Methodological), 58(1), 267-288. Retrieved from http://www.jstor.org/stable/2346178

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. IEEE transactions on evolutionary computation, 1(1), 67-82.

Xu, J., Lu, T., Compton, R., & Allen, D. (2014). Civil Unrest Prediction: A Tumblr-Based Exploration. *SBP*.

Yellow vests movement. (2018.). Retrieved December 6, 2018, from https://en.wikipedia.org/wiki/Yellow_vests_movement

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*. Series B: Statistical Methodology, 67(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Appendices and Supplementary Materials

Appendix 1.

Packages used in R-Studio (version 1.1.447) for pre-processing and analyzing the data.

R-package	**Task description**
GDELTtools	Extracting data from the GDELT database.
Dplyr, Tidyr	Manipulating the data in the right shape.
ImputeTS	Replacing the missing values.
Ggplot2	Creating visualization graphs.
Caret, Glmnet, RandomForest	Building models and analyzing.