Fine-Tuning Language Models on Dutch Protest Event Tweets

Meagan B. Loerakker
Chalmers University of Technology
Netherlands Police
meagan@chalmers.se

Laurens H.F. Müter
Utrecht University
Netherlands Police
l.h.f.muter@uu.nl

Marijn P. Schraagen
Utrecht University
m.p.schraagen@uu.nl

Abstract

Being able to obtain timely information about an event, like a protest, becomes increasingly more relevant with the rise of affective polarisation and social unrest over the world. Nowadays, large-scale protests tend to be organised and broadcast through social media. Analysing social media platforms like X has proven to be an effective method to follow events during a protest. Thus, we trained several language models on Dutch tweets to analyse their ability to classify if a tweet expresses discontent, considering these tweets may contain practical information about a protest. Our results show that models pre-trained on Twitter data, including Bernice and TwHIN-BERT, outperform models that are not. Additionally, the results showed that Sentence Transformers is a promising model. The added value of oversampling is greater for models that were not trained on Twitter data. In line with previous work, preprocessing the data did not help a transformer language model to make better predictions.

1 Introduction

The number of protests across the globe has grown in the last decade (e.g. (Haig et al., 2020)).[1] Public safety can be threatened at these protests when riots can break out. For example, Trump supporters attacked the United States Capitol in Washington, D.C. on January 6, 2021 (Dave et al., 2021). State property was destroyed (repairs exceeding $1.5 million (United States Attorney's Office, 2021)) and several law enforcement officers lost their lives during the riots that followed (United States Senate Committee on Homeland Security & Governmental Affairs, 2021). While these examples of extreme social unrest are generally uncommon, they express a need to forecast these types of events. In the Netherlands, a possibility of large-scale protests exists. An example is the Curfew protests (Dutch: Avondklokrellen) held in 2020 and 2021 across several cities during the Covid-19 pandemic (Moors et al., 2022; COT, 2021). Internationally, an emergence of Covid-related protests at the end of 2020 was observed (van der Zwet et al., 2022). Additionally, people experience more confidence in influencing politics during a protest, compared to voting (Harding et al., 1986; Oliver, 2001), where Kleiner (2018) argues that extremists are likely to voice their opinions through protests.

Previous protests and stricter Covid rules might lead to a divided population over time due to decreased social mobility (Moors et al., 2022). This lack of social mobility is argued to be the source of growing discontent and polarisation in the country (Sandel, 2020). It is reported that these higher levels of affective polarisation have increased in the Netherlands (Harteveld and Wagner, 2023). Since polarisation might lead to more social unrest, the Dutch police are interested in gaining knowledge about the emergence of protests.

Nowadays, it is possible to follow incidents in real time as people increasingly use social media to broadcast live events (e.g. (Shamma et al., 2010)). As a result, X (formerly Twitter) is increasingly being studied as a news reporting platform more than anything else (Weng and Lee, 2011; Petrovic et al., 2013; Phuvipadawat and Murata, 2010). It is observed that disaster-related events are also being reported on X (Imran et al., 2015; Shamma et al., 2010; Thelwall et al., 2011; Williams and Burnap, 2015; Burnap et al., 2014). As an example, Starbird and Palen (2012) describe the Arab Spring protests in 2011 as uprisings of a political nature, where social media was pointed out as having gained a more important role in these types of protests. Subsequently, actors such as governments and policing agencies "aim to understand how events are reported using social media and how millions of online posts can be reduced to accurate but meaningful information that can support

[1] See also http://visionofhumanity.org/reports

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 6–23

decision making and lead to productive action" (Alsaedi et al., 2017, p. 2). Scholars studying social movements have argued that social networks–as established through social media–are fundamental to protest participation (e.g. (Snow et al., 1980; Boulianne et al., 2020)). Alsaedi et al. (2017) used an event detection framework in combination with temporal, spatial and textual content features from X to detect different kinds of events, including disruptive ones and those on smaller scale. Furthermore, they found that their method performs at least as well as using other terrestrial sources. In line with this, social media has been used as a way to warn people of unsafe areas and to spread awareness for disaster relief fundraising (Lindsay, 2011). This power that social media possesses has also been demonstrated during the Haiti earthquake in January 2010, where the awareness raising resulted in 8 million U.S. dollar donations to the Red Cross (Gao et al., 2011). This suggests that understanding the dynamics of social media messaging, especially during high-impact events like protests, are key to timely decision-making.

An advantage to social media analysis is that information about events can be extracted faster than official news reports publicise (Osborne and Dredze, 2014). However, one of the main research challenges in studying civil unrest, is the actual identification of such information in the fasts amount of data (Sech et al., 2020). We propose an analysis approach to recognising such information: classifying messages based on expressions of discontent.

2 Expression of Discontent

A link between discontent and collective protests is described by Somma (2017), where discontent is a negative feeling towards certain aspects of the world, which includes distrust in political authorities, rules, or decisions. Since X is a popular way to motivate people to protest (Doğu, 2019), we aim to detect expressions of discontent in tweets (see Section 5.1.2 for a precise definition). We hypothesise that people expressing discontent are more likely to start protesting.

This paper compares several BERTje, mBERT, Bernice, TwHIN-BERT and Sentence Transformers models fine-tuned to newly annotated datasets of Dutch tweets. We include experiments with the SetFit framework and compare to a logistic regression baseline.

We aim to understand how these models can identify expressions of discontent, and how well they perform on Dutch protest event tweets.

3 Social Media Analysis Challenges

OSINT utilises social media analysis to gain insights into events taking place in the country. However, current social media analysis practices pose several challenges related to privacy and the availability of suitable tools.

3.1 Privacy

In the context of the Dutch police, the OSINT unit aims to predict when and where police forces are needed in case a protest is organised. OSINT must take the GDPR (General Data Protection Regulation) into consideration when predicting these riots (Schermer et al., 2018). For example, the GDPR does not allow for the creation of profiles or monitoring of individuals' anticipation of potential crime or involvement in a riot.

As part of protest prediction, OSINT evaluates tweets according to their sentiment. If a tweet contains expressions of discontent, it is typically deemed as more relevant for analysis. Using machine learning models can result in more objective predictions of discontent. At the same time, OSINT requires models that respect individuals' privacy, as well as obtain insightful predictions. Individual privacy can be respected with models that focus on topics, communities, and sentiments of communities, rather than focusing on individuals. Moreover, the creation of these types of models can aid in the transfer of tacit knowledge within organisations. For example, the creation of manually tailored queries require experience from former protests, hence involving tacit knowledge that is difficult to express due to its non-codified disembodied nature (Howells, 1996; Ribeiro, 2013). Besides, a machine learning model's quality is assessed on its generalisability by evaluating their performance on previously unseen data (Roelofs, 2019; Raschka, 2018), whereas queries remain difficult to generalise due to their usage of specific keywords. Therefore, developing a machine learning model on a given task results in a more efficient prediction process.

3.2 Non-English Data

Despite the availability of numerous efficient and well-designed algorithms, models produced using these algorithms are often trained on the English language. This poses a challenge for organisations

situated in countries where English is not the native language. Baden et al. (2022) discussed three research gaps in the field of Computational Text Analysis Methods (CTAM). One of these research gaps is the focus on the English language, which results in a lack of tools to study other languages. Entities situated in the Netherlands have to deal with Dutch text and information, primarily, for example when analysing social media posts. Hence, there is a need to evaluate how well language models perform on Dutch text, as well as evaluating to what extent fine-tuning a model may influence its performance. Unsurprisingly, the Dutch police and thereby OSINT have to deal with Dutch text and information, primarily. Hence, this calls for a need to evaluate how well language models perform on Dutch text, as well as evaluate to what extent fine-tuning existing models influences performance.

4 Models & Frameworks

De Vries et al. (2019) created BERTje for Dutch text, which outperforms a multilingual BERT model with Dutch training data on word-level tasks. However, De Vries et al. note that it remains unclear how well it performs with tasks on sentence-level, which relates to a model's deeper understanding of different types of information. In general, transformer models like BERT are restricted in their input length. Pascual et al. (2021, p. 2) note that a transformer's complexity 'scales quadratically with the length of their input.'

Bernice is a multilingual RoBERTa language model specifically trained on tweets through a custom tokenizer, which is described as the first BERT model to have been trained on this type of data (DeLucia et al., 2022). Another multilingual model trained on a large Twitter corpus has been released recently: TwHIN-BERT (Zhang et al., 2023). Both models were developed in 2022. The creators of both the Bernice and the TwHIN-BERT models found that they outperform or matches other models' performance on social media data. Therefore, we aim to evaluate how Bernice and TwHIN-BERT perform against other models on a specific task like discontent detection.

SetFit stands for 'Sentence Transformer Fine-Tuning' (Reimers and Gurevych, 2019). Sentence Transformer frameworks use Siamese and triplet network structures to modify pre-trained transformer models (Tunstall et al., 2022) to efficiently derive contextual embeddings for larger units of text such as sentences. SetFit has been used for social media data (Bates and Gurevych, 2023). A characteristic is its relatedness to few- and zero-shot approaches (Tunstall et al., 2022). These approaches have gained traction in the research community as they may prove helpful in domains lacking resources. Few-shot learning (FSL) refers to the principle of learning a task with a limited number of labelled inputs, the 'shots' (Liu et al., 2022). The training data is smaller than normally used to train models. Thus, FSL is relatively data-efficient. SetFit can achieve a high accuracy with few-shot fine-tuning, with a performance comparable to fine-tuned RoBERTa models.[2]

Although Sentence Transformers (ST) models using SetFit show promising results for languages such as German, Japanese and more on classification tasks[2], to our knowledge it has not been tested on Dutch yet. In this work, we test ST models both using regular fine-tuning and using FSL through the SetFit framework.

5 Method

The collected tweets are labelled according to the classes 'No discontent' and 'Expression of discontent'. Then, *mBERT*, *BERTje*, *Bernice*, *TwHIN-BERT*, and multilingual *Sentence Transformer* models are fine-tuned using the labelled datasets from the previous step. A Logistic Regression model is trained to determine baseline performance. We mainly focused on training a Sentence Transformers model without the SetFit framework due to time and resource constraints, as the SetFit framework took substantially longer to train.

The models are evaluated on several metrics. The anonymised data and used code for the models are publicly available at `https://github.com/Meaganium/Detecting-Discontent-in-Dutch-Events`. In summary, we evaluate whether or not there is a difference in models' performance in how well they predict a tweet's expression of discontent.

Table 2 provides an overview of the used models.

5.1 Data Collection

The data consists of Dutch tweets related to protests that took place in the years 2020–2022. Collecting the data for each protest was done in a reactive manner where tweets are downloaded a few days

[2] `https://huggingface.co/blog/setfit`

Protest	Date of collection	Filter keywords	Discontent / Total
Fireworks ban protest (RO)	November 20, 2022	protest †, rotterdam, protesters *, hooligans ‡	600 / 4214
Curfew riots (EI)	January 24, 2021	protest †, curfew §, eindhoven, riots ˆ	383 / 3892
Black Pete (MA)	November 14, 2020	protest †, piet, maastricht	1440 / 10395
Black Lives Matter (AM)	June 1, 2020	protest †, black lives matter, amsterdam	2064 / 6329

Table 1: Datasets related to Dutch protests used in this study. Each dataset is defined by specific keywords used to retrieve relevant tweets. Original Dutch keywords: † demonstratie, * betogers, ‡ relschoppers, ˆ rellen, § avondklok.

after the incident. Since the X API allows downloading historic tweets not older than two weeks, the available data spans between two to three days before and after the incident. On the days of the protests themselves, we extracted the majority of the tweets. The tweets were collected via the X API. A filter was applied to select Dutch tweets only related to protests. Table 1 provides an overview of the specific filters and result set size per protest. Retweets are excluded and since a free version of the API is used, only a subset of the tweets is available.

5.1.1 Preparing the Data

Solely the tweets' contents were used. Any meta information such as geolocation, number of likes, number of retweets, and comments were ignored, as this type of meta information is mostly relevant for the creation of networks rather than determining sentiment. Elements such as hashtags, emojis and punctuation are included in the analysis. From the tweet texts, any personal information was replaced by a placeholder. Username mentions in the tweet were not masked. During the labelling process, off-topic tweets (see Appendix C), tweets containing personal information, duplicates, and auto-generated tweets were removed from the datasets.

5.1.2 Data Labelling

The tweets were labelled according to whether the tweet contained an 'Expression of discontent' (**EOD**). If the tweet included an indication that the corresponding user disagreed with the government's actions, the rioters' actions, or provided a potential reasoning for protesting, the tweet was labelled as EOD. For this labelling process, weekly meetings were held to discuss tweets that were more difficult to label, e.g., due to nuance, sarcasm and jokes. This labelling process was performed by the first and second author with eight other annotators, including university students and police employees. Each dataset was annotated by a different composition of the annotator team. The average inter-annotator agreement across all datasets was around 70%, which is considered respectable, especially for linguistic annotations (Artstein, 2017). The labelling was done in a self-made tool named Tweeti, available at `https://github.com/LMuter/Tweeti`. The labelling process was conducted over a period of 18 months. Table 7 (Appendix B) provides some example annotations.

5.1.3 Test and Training Data

The data is divided into two sets: training (80%) and testing (20%). The training data is used to train model weights and the test set is used to test the models' performance. We used fixed hyperparameter settings for all models. Due to the small size of the training set and spelling variations in tweets, words might not overlap between training and test, impeding direct keyword mapping. This prompts the model to focus on the context of the keyword occurrences instead of the words themselves, which can make the model more flexible. The datasets were imbalanced (Table 1), as they contained substantially more tweets in the 'No discontent' class than the EOD class. Due to this imbalance, we took into consideration four other evaluation metrics (precision (P), recall (R), F1 and Area Under the Curve (AUC)) besides accuracy (ACC), as accuracy will be influenced by how well the majority class can be predicted (Abd Elrahman and Abraham, 2013). In this paper, we report the macro averages of ACC and AUC, and the micro averages of P, R and F1. The micro averages allowed us to gain more insight into, i.e., the distribution of the number of true positives and false positives across the two classes. The metrics were measured per class, as macro-averages are heavily influenced by imbalance.

5.2 Training Phase

We consider several models for our study. As a baseline, a bag-of-words based Logistic Regression (LR) model is trained. Furthermore, pre-trained mBERT, Bernice, TwHIN-BERT, BERTje and Sentence Transformer (ST) models are fine-tuned. Ini-

Model	Hugging Face URI
BERT	bert-base-uncased
mBERT	nlptown/bert-base-multilingual-uncased-sentiment
BERTje	GroNLP/bert-base-dutch-cased
Sentence Transformers (ST)	sentence-transformers/paraphrase-multilingual-mpnet-base-v2
Bernice	jhu-clsp/bernice
TwHIN-BERT-base	Twitter/twhin-bert-base
TwHIN-BERT-large	Twitter/twhin-bert-large

Table 2: Overview of the models used from the Hugging Face platform.

tial experiments were performed using the SetFit framework with an ST model.

5.2.1 Pre-Training Phase

For the implementation of the models, we used the 'text-classification' task in the pipeline in order to be able to assign the EOD label to a tweet. See Table 6 in Appendix A for a more extensive overview of the used Python libraries and functions. in order to make their produced results more comparable with one another, as this task allows for the tokenization of sequences of text, rather than one individual word. These (sub)words are the result of the separated text sequences, representing the tokens.[3] Note that these subword tokenizers partially solve the issue of error mistakes. By declaring this task for BERTje, they perform the same tokenization process, hence making their results comparable.

The Hugging Face tokenizers are used in the pre-training phase. The tokenization process consists of three steps. Firstly, indicators are added to demarcate the start and end of the tweet, signified by the special tokens [CLS] and [SEP], respectively. Secondly, uniformity in the tweet length is ensured by adding [PAD] to short tweets and truncating long tweets. Finally, the converted tokens are assigned IDs, and an attention mask is created.

5.2.2 Pre-Processing the Data for LR

We trained various additional models on a pre-processed version of the datasets in order to evaluate the difference in performance, considering it can provide better results for bag-of-words techniques (Angiani et al., 2016). Pre-processing the data involved lowercasing and lemmatization, and removing all URLs, Dutch stopwords, username mentions, accents on letters and punctuation from the tweet text. Furthermore, only content words like verbs, nouns, adjectives, adverbs and proper nouns were typically included after pre-processing.

[3] https://huggingface.co/docs/transformers/v4.28.1/en/task_summary#sequence-classification

6 Results

To evaluate the results for the EOD prediction, we focused mainly on the results for the EOD label, as this was the minority class for all individual datasets. See Tables 3 and 4 for the prediction results. First, we ran all the models with all four dataset combined, see Table 1. In this round, the models were run with the oversampling technique with the assumption that oversampling the minority class would compensate for the data imbalance. We further investigated the outcomes on the datasets separately by training the best models from the previous round on all the datasets combined with oversampling to identify differences in performance per dataset. After some test runs without oversampling (see Table 12, Appendix D), we observed that oversampling may not produce substantially different results. As such, we ran the models without oversampling the minority class to evaluate the models' sensitivity to data imbalance. Lastly, we did some extra experimentation with the multilingual model mBERT.

6.1 Logistic Regression

First, we evaluated the baseline performance with the Logistic Regression (LR) model. While pre-processing is not always beneficial for deep learning methods (Camacho-Collados and Pilehvar, 2018), for bag-of-words models it is commonly used, hence its inclusion. LR did not perform better than the other models, which is highlighted by the fact that LR has no bold numbers in Table 3.

To evaluate LR's potential further, we experimented with all possible combinations of pre-processing steps as given in Section 5.2.2. See Appendix D, Table 8. The best combination of pre-processing steps was a combination of lowercasing and removal of URLs, username mentions, diacritics and punctuation. This pre-processing combination resulted in similar results (F1: .418) compared to no pre-processing (F1: .422).

				Expression of discontent			No discontent		
Model	**Measure**	**ACC**	**AUC**	**P**	**R**	**F1**	**P**	**R**	**F1**
Bernice	AVG	.870	**.784**	.652	.646	**.649**	**.919**	.921	.920
BERTje	AVG	.867	.745	**.685**	.548	.609	.899	.941	.919
TwHIN-BERT-base	AVG	.859	.782	.631	**.657**	.643	.917	.908	.912
TwHIN-BERT-large	AVG	.828	.609	.600	.261	.248	.856	**.957**	.900
Sentence Transformers	AVG	**.871**	.766	.682	.597	.636	.908	.935	**.921**
Logistic Regression	AVG	.808	.708	.530	.539	.534	.881	.877	.879
Bernice	STD	.004	.009	.018	.024	.012	.005	.008	.003
BERTje	STD	.008	.014	.021	.029	.022	.006	.007	.005
TwHIN-BERT-base	STD	.004	.010	.015	.028	.008	.006	.009	.003
TwHIN-BERT-large	STD	.021	.028	.071	.110	.012	.056	.054	.011
Sentence Transformers	STD	.004	.004	.020	.006	.009	.006	.004	.003
Logistic Regression	STD	.009	.006	.016	.011	.006	.007	.009	.006
Bernice	MIN	**.865**	.772	.623	.619	.632	**.914**	.907	**.917**
BERTje	MIN	.859	.721	**.670**	.497	.609	.890	**.934**	.914
TwHIN-BERT-base	MIN	.853	**.776**	.611	**.639**	**.635**	**.914**	.894	.908
TwHIN-BERT-large	MIN	.810	None	None	None	None	.810	.854	.889
Sentence Transformers	MIN	.864	.761	.656	.589	.625	.899	.931	.916
Logistic Regression	MIN	.797	.704	.516	.523	.528	.875	.871	.870
Bernice	MAX	.875	.793	.667	.679	**.663**	.926	.926	.923
BERTje	MAX	**.877**	.756	**.720**	.570	.629	.905	**.950**	**.926**
TwHIN-BERT-base	MAX	.864	**.800**	.647	.707	.656	**.928**	.918	.916
TwHIN-BERT-large	MAX	.864	.792	.650	**.731**	.628	**.928**	None	.917
Sentence Transformers	MAX	.874	.771	.704	.605	.644	.912	.940	.924
Logistic Regression	MAX	.820	.718	.558	.552	.542	.892	.889	.888
SetFit		.869	.758	.689	.577	.628	.904	.939	.921

Table 3: Comparison between the models for EOD prediction in combination with oversampling of the minority class. The models were run five times, except for SetFit. The averages, standard deviations, minima and maxima values of those rounds are provided. The numbers are rounded, and the best scores for the averages, minima and maxima per metric are in bold. Table 9 in Appendix D provides the results of all the runs.

6.2 Averages, Standard Deviations, Minima and Maxima

We ran the Bernice, BERTje, TwHIN-BERT-base, TwHIN-BERT-large, ST and Logistic Regression models five times to get insight into the range of possible scores they provide. The results of all five runs are provided in Appendix D, Table 9. Of the five runs, we mainly focus on discussing the minima, maxima and averages.

For the minima scores, we observe that the TwHIN-BERT-base overall produces the best scores on EOD (AUC: .776, F1: .635), though Bernice had the highest accuracy (.865).

Similarly for the maxima scores on EOD, TwHIN-BERT-base (AUC: .800) and Bernice (F1: .663) score the best overall. BERTje scored best on accuracy and precision. For both the minima and maxima scores, neither ST nor LR scored highest on a particular metric.

In line with the minima and maxima, we observe that Bernice scores the best on average for the minority class (AUC: .784, F1: .649). Notably, ST scored highest on accuracy for EOD (.871). Universally, we observe that Bernice and TwHIN-BERT-base provide better results compared to TwHIN-BERT-large, BERTje, ST and LR.

6.3 Separate Datasets

In this section, we describe the results for the individual AM, EI, MA and RO datasets. Note that the RO dataset is divided in two, wherein each version (’22 and ’23) was labelled by other annotators. The subsets of the RO dataset overlapped to some degree, but not fully. The annotators of the ’23 version were the same annotators for AM.

6.3.1 With Oversampling

From the former round, we identified that Bernice and TwHIN-BERT-base outperformed the other models with oversampling. Arguably, Bernice performs slightly better than TwHIN-BERT-base due to its average AUC and F1 scores.

As shown in Table 4, when running Bernice and TwHIN-BERT-base on the separate datasets, we observe that the models perform worst on the EI dataset on EOD. The MA dataset was the second worst performing dataset.

Bernice produced the highest scores for the ACC (.893) and P (.754) metrics on the RO dataset. Though, the AM dataset was observable the best performing dataset, with TwHIN-BERT-base pro-

ducing the highest scores for AUC (.791), R (.732) and F1 (.712) on this dataset.

6.3.2 Without Oversampling

In Table 10 are the results provided by running BERTje, Bernice, TwHIN-BERT-base and TwHIN-BERT-large on all separate datasets without oversampling the minority class.

The EI dataset performs substantially worse compared to the other datasets. The AM dataset also again outperforms the other datasets, with BERTje producing the highest P (.765), and Bernice producing the best AUC (.806), R (.760) and F1 (.731) on the EOD class.

The poor results on the EI dataset can be partially attributed by the labelling process. Whereas the AM dataset was solely labelled with the EOD class, the EI and MA datasets were labelled with substantially more classes, with only tweets labelled as EOD or 'No discontent' retained and all other tweets removed from the data. The usage of more labels poses greater opportunity for disagreements among annotators, hence affecting the quality of the labels. Moreover, the proportion of EOD tweets is substantially lower for these two datasets.

6.3.3 Oversampling vs Non-Oversampling

Besides the observations previously mentioned, it is noteworthy that the oversampling technique does not always guarantee better results for some models. For some models, oversampling has more added value compared to others.

For example, oversampling on the EI dataset did not help for the Bernice model on some metrics, including AUC, P, R and F1.

Furthermore, we observed that the oversampling was ineffective for the TwHIN-BERT-base model on the ACC and P metrics. This was the case for all datasets RO23 on ACC. Although this finding may suggest that oversampling provides less added value for the models trained on Twitter data due to their higher performance in general, this observation warrants further investigation.

6.4 Notable Results

In this section, we describe some noteworthy additional results we found.

6.4.1 TwHIN-BERT-large

The TwHIN-BERT-large model was unable to converge in some runs on the data. This is likely due to the small size of the datasets, whereby the model is unable to fine-tune all its parameters due to its large size. When TwHIN-BERT-large was able to configure all its parameters, it produced good results. As shown in Table 9, run 1 provided the best R score (.731) across all runs and models. Run 5 also configured correctly, with some notable results being the ACC and P.

6.4.2 BERTje

To investigate if the pre-trained data influences the results, we trained a BERTje model based on a Dutch text corpus with a variety of settings. The extra results for BERTje can be found in Appendix D, Tables 8 and 11.

For BERTje, we experimented with preprocessing to evaluate whether our results are in line with previous work (e.g. (Camacho-Collados and Pilehvar, 2018; Kurniasih and Manik, 2022; Alzahrani and Jololian, 2021)). For each metric, except recall in the 'No discontent' class, preprocessing lowered BERTje's performance. Especially the recall (.091) and F1 (.153) scores were particularly poor in the minority class. This is explained by BERT's use of contextual information, like punctuation, morphology and sentence structure.

To find out if there is a difference in performance between annotators for the same dataset, we trained BERTje on the two RO dataset versions. We found a noticeable difference for all metrics.

Using all datasets provided the best score compared to using the datasets separately for EOD on precision (.863), though recall was poor (.328). This shows that providing BERTje with more data, despite the imbalance, will result in the model classifying a large number of items with the minority class correctly whilst still missing quite a large portion of tweets to label as 'discontent'. This shows that BERTje can identify strong markers in the tweets that suggest discontent, as long as it is given a sufficient amount of data. At the same time, the low recall score would suggest that identifying discontent is still a nuanced task, meaning that these nuances make it difficult to define all concrete markers of discontent. Generally, it is questionable if it is possible to capture this complex notion with a language model using short social media messages.

6.4.3 SetFit

We ran SetFit once, and it did not produce better results over Bernice and TwHIN-BERT-base (see Table 3). SetFit also takes substantially longer to

				Expression of discontent			No discontent		
Model	**Data**	**ACC**	**AUC**	**P**	**R**	**F1**	**P**	**R**	**F1**
Bernice	AM	.826	.781	.745	.664	.702	.857	.898	.877
Bernice	EI	.888	.574	.368	.182	.243	.915	**.966**	**.940**
Bernice	MA	.866	.639	.508	.328	.399	.901	.950	.925
Bernice	RO22	**.893**	.768	.622	.595	.608	.935	.941	.938
Bernice	RO23	.884	.765	**.754**	.573	.652	.906	.957	.931
TwHIN-BERT-base	AM	.813	**.791**	.694	**.732**	**.712**	.873	.851	.862
TwHIN-BERT-base	EI	.886	.653	.412	.364	.386	.931	.943	.937
TwHIN-BERT-base	MA	.868	.665	.514	.386	.441	.908	.943	.925
TwHIN-BERT-base	RO22	.864	.769	.510	.638	.567	**.939**	.901	.920
TwHIN-BERT-base	RO23	.866	.749	.677	.560	.613	.901	.938	.919

Table 4: Comparison between the models Bernice and TwHIN-BERT-base for EOD prediction across all four datasets with oversampling of the minority class. The numbers are rounded.

train than the other models. Due to these constraints, we did not experiment with SetFit further. However, the underlying ST model did produce reasonable results when oversampling the minority class for both the AM and MA datasets. Therefore, future work with less restrictions regarding time and resources could explore SetFit's potential further.

6.4.4 Multilingual BERT: mBERT

Since multilingual models are known to perform well on monolingual tasks (Rust et al., 2021), we experimented shortly with the multilingual version of the BERT model: mBERT (see Appendix D for extra results). Oversampling the minority class in the AM dataset produced reasonable scores for AUC (.758), recall (.646) and F1 (.677). However, this introduces a performance reduction for the 'No discontent' class of around .08. Notably, mBERT scored particularly well on precision (.833) for a pre-processed dataset, while ACC, R and F1 came out relatively low.

The mBERT results are not in line with previous work where monolingual models outperform multilingual models (e.g. (De Vries et al., 2019; Rust et al., 2021)), but they are not totally unexpected given the substantial amount of English words and phrases used in Dutch social media. Similar to other models, combining oversampling and including emojis did not improve the results compared to solely applying oversampling. However, we suggest future work to take this multilingual nature of social media messaging into consideration through analyses based on the principle of code-switching (e.g. (Das and Gambäck, 2014)), like Language Identification (see (Aguilar et al., 2020; Barman et al., 2014; Khanuja et al., 2020; Molina et al., 2019; Solorio et al., 2014)).

6.5 T-test Results

To gain insight into how different the models perform compared to one another, we conducted two-way t-tests on the average F1 of five runs. Table 5 provides a full overview of the t-test results. However, note that the t-test results for comparison between TwHIN-BERT-large and other models were influenced by the fact that TwHIN-BERT-large could not compute several runs. Naturally, runs that were not completed successfully were excluded from the tests.

From the t-tests, we find that Bernice's, TwHIN-BERT-base's, and the ST's F1 scores are significantly different from logistic regression ($p < .001$). Furthermore, we found that BERTje's F1 score was significantly difference compared to TwHIN-BERT-base's and logistic regression with $p < .01$.

These results support our previous findings that the models trained on Twitter data (Bernice and TwHIN-BERT) report better prediction of EOD, as Bernice and TwHIN-BERT-base perform significantly different from the baseline (logistic regression). Notably, BERTje and ST also perform significantly different from the baseline, suggesting that these models also have the potential to provide reasonable results on the data.

7 Discussion

In this paper, we aimed to identify how future NLP models can be improved in order to provide better predictions for social media text. Our work provides an overview of several language models' performances on Dutch tweets for the prediction of *Expression of Discontent*.

Whether someone expresses discontent is dependent on human interpretation, thus complicating the identification process of parameters that determine tweet sentiment. Moreover, human annotators may

	Bernice	BERTje	TwHIN-BERT-base	TwHIN-BERT-large	Sentence Transformers
BERTje	.022*				
TwHIN-BERT-base	.250	.006**			
TwHIN-BERT-large	.030*	.038*	.030*		
Sentence Transformers	.029*	.028*	.170	.032*	
Logistic Regression	7.4E-06***	.002**	5.4E-06***	.068	2.8E-05***

Table 5: Overview of the t-test results between the F1 scores of the models' predictions on all of the four datasets combined. Asterisks denote p-values: * $p < .05$, ** $p < .01$, *** $p < .001$.

consider other kinds of information subconsciously when labelling a tweet for discontent. This claim is supported by results we found when training models on subsets of the RO dataset labelled by two different annotator teams.

The results showed that the models trained on Twitter data, namely TwHIN-BERT and Bernice, performed best. Pre-processing did not improve the results for any model. This highlights the importance of using models that have been pre-trained on similar types of data for event prediction, which is in line with a review conducted by Zimbra et al. (2018). They found that the average accuracy for sentiment analysis on Twitter data was 61%, and that state-of-the-art approaches performed similarly, with accuracies routinely below the 70%. However, they did find that domain-specific approaches performed better by 11%, which is an average increase in performance we did not achieve. For all datasets combined, Bernice and TwHIN-BERT-base achieved average scores ranging from .63 to .65 for precision, recall and F1 on the EOD class, though for both classes (EOD and 'No discontent') the average accuracy and AUC scores were substantially higher, ranging from .78 to .87.

Surprisingly, we found that the Sentence Transformers models perform on par with Bernice and TwHIN-BERT, despite not being a pre-trained model on Twitter data. Additional results showed that mBERT, a multilingual model, performed better than BERTje. This may be because social media users tend to lend words from other languages, including English. Furthermore, mBERT is trained on a larger corpus of text compared to BERTje. This indicates that the selection of a specific dataset to pre-train a language model is one of the main indicators to acquire a greater return on prediction performance.

Lastly, we found that oversampling provides substantial benefits for smaller datasets, like EI and MA in our work, whereas the benefit is limited for larger ones, like AM in our work. Furthermore, when combining all datasets together, the benefit of oversampling was also limited. However, the issue of highly imbalanced datasets cannot be fully solved with oversampling, which was observed in the results. In line with this, some models gain more benefit from oversampling than others. In particular, oversampling had the least added value for the models trained on Twitter data, potentially due to their relatively high base performance.

All in all, the results indicate that the identification of discontent in social media text is a feasible approach to filtering relevant to irrelevant messaging, given that the appropriate language models are chosen. The ability to accurately filter the data provides opportunities for more efficient extraction of a variety of information relevant to entities like the police and OSINT, including locations, dates and time stamps.

7.1 Limitations

First, in all of the used datasets, the number of 'No discontent' tweets outnumbers the number of discontent tweets with a ratio of around one to five. In order try to make up for this limitation, we used the widely used oversampling technique named Synthetic Minority Oversampling TEchnique (SMOTE) (Chawla et al., 2002) for the discontent tweets. However, SMOTE has limitations, including misclassification of the majority class, resulting in negative effects for the model's overall balance (Puntumapon and Waiyamai, 2012).

Second, some publicly available tweets may have been removed by the corresponding users since the tweets have been extracted via the X API, which may reduce the reproducibility of the study. Besides that, it is possible that some of the results were unsatisfactory partially due to the switch-ups in the annotator teams. The compositions in the annotator teams may have resulted in some inconsistencies in the labelling process.

Third, we did not conduct an error analysis in this work. Therefore, future work that aims to

build upon this paper should consider aiming to gain more insights into, i.e., whether the degree of false positives for a specific dataset correlates with the (perceived) difficulty of the annotation task. However, to support such an error analysis, we propose follow-up studies to report more details on the inter-annotator agreement.

Fourth, being able to predict a protest's location, date, time or size is also of interest to OSINT and the Dutch police, especially in times of higher affective polarisation and social unrest. In this work, we did not explore the extraction of such information from the tweets, presenting an opportunity for future work.

Lastly, the practice of combining all four datasets may be flawed. Some protests may have been more extreme in terms of the events that took place, hence (indirectly) influencing how the annotators interpret discontent per protest. Therefore, some datasets may capture a limited, or even a different, meaning of expression of discontent, given that the datasets were labelled for different protests and/or with more labels, affecting classification performance.

8 Ethics Statement

Ethical approval to conduct this study, including approval for the collection and annotation of the datasets, was acquired from the appropriate local institutional review boards and ethics committees. To minimise potential privacy issues, we excluded direct and indirect personal identifiers from the data, including names and locations. In line with the GDPR guidelines, the data has been anonymised by hashing usernames and mentions.

Besides the focus on Dutch text, it is desirable for high-impact applications, like those used in medical practice and law enforcement, to work with models and algorithms that have low false negative rates, due to potential societal and ethical complications that arise with false positives. For example, it is unethical and socially undesirable to inaccurately label a person's social media message along the lines of 'high-risk' or 'negative'. Therefore, in this work, we focused on the optimisation of the precision metric, as this indicates lower false positives. We encourage future work to put low false positive rates at the forefront in the evaluation of models' performance.

Furthermore, we followed the European Data Protection Board (EDPB) guidelines to assess the risks and potential impacts of the data.[4] These guidelines were followed in order to minimise potential risks for individuals' freedoms, and to use the data in a lawful and transparent manner.

For future work, we provide several suggestions on how to use social media data in an ethical manner. First, ethical data assessment methodologies should be used before the analysis is conducted in order to evaluate potential conflicts with (public) values and to minimise social disruption. We recommend using approaches like 'De Ethische Data Assistent' (DEDA, 'The Ethical Data Assistent') from Schäfer et al. (2022). Second, when conducting social media analysis, the focus should be on groups rather than individuals, so that privacy is ensured and the results remain 'superficial' in nature. As previously mentioned, the GDPR emphasises that *monitoring* and *profiling* is not allowed, even in the context of anticipating crimes and riots. Therefore, social media analyses for research purposes should emphasise the recognition of general trends, sentiments and events instead, as presented in this paper.

Acknowledgements

This work was supported by the Swedish Research Council, award number 2022-03196.

References

Shaza Abd Elrahman and Ajith Abraham. 2013. A review of class imbalance problem. *Journal of Network and Innovative Computing*, 1:332–340.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, page 1803–1813.

Nasser Alsaedi, Pete Burnap, and Omer Rana. 2017. Can we predict a riot? Disruptive event detection using Twitter. *ACM Transactions on Internet Technology (TOIT)*, 17(2):1–26.

Esam Alzahrani and Leon Jololian. 2021. How different text-preprocessing techniques using the BERT model affect the gender profiling of authors. *3rd International Conference on Machine Learning & Applications (CMLA 2021)*, pages 1–8.

Giulio Angiani, Laura Ferrari, Tomaso Fontanini, Paolo Fornacciari, Eleonora Iotti, Federico Magliani, and Stefano Manicardi. 2016. A comparison between

[4] https://edpb.europa.eu/sites/default/files/consultation/edpb_guidelines_202008_onthetargetingofsocialmediausers_en.pdf

preprocessing techniques for sentiment analysis in Twitter. In *Proceedings of KDWeb 2016*.

Ron Artstein. 2017. Inter-annotator agreement. *Handbook of Linguistic Annotation*, pages 297–313.

Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken van der Velden. 2022. Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, 16(1):1–18.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23. Association for Computational Linguistics, Doha, Qatar.

Luke Bates and Iryna Gurevych. 2023. Like a good nearest neighbor: Practical content moderation with Sentence Transformers. *arXiv preprint 2302.08957*.

Shelley Boulianne, Karolina Koc-Michalska, and Bruce Bimber. 2020. Mobilizing media: Comparing tv and social media effects on protest mobilization. *Information, Communication & Society*, 23(5):642–664.

Pete Burnap, Matthew Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. 2014. Tweeting the terror: Modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining*, 4:1–14.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 40–46. Association for Computational Linguistics.

Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357.

COT. 2021. Een machteloos gevoel: Leerevaluatie naar aanleiding van de ongeregeldheden in Den Bosch op 25 januari 2021. *AON*, pages 1–27.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed Indian social media text. *Proceedings of the 11th International Conference on Natural Language Processing*, page 378–387.

Dhaval Dave, Drew McNichols, and Joseph Sabia. 2021. Political violence, risk aversion, and non-localized disease spread: Evidence from the US capitol riot. Technical report, National Bureau of Economic Research.

Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. *arXiv preprint 1912.09582*.

Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark Dredze. 2022. Bernice: A multilingual pre-trained encoder for Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6191–6205. Association for Computational Linguistics.

Burak Doğu. 2019. Environment as politics: Framing the Cerattepe protest in Twitter. *Environmental Communication*, 13(5):617–632.

Huiji Gao, Geoffrey Barbier, and Rebecca Goolsby. 2011. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, 26(3):10–14.

Christian S Haig, Katherine Schmidt, and Samuel Brannen. 2020. The age of mass protests: Understanding an escalating global trend. `https://www.csis.org/analysis/age-mass-protests-understanding-escalating-global-trend` (accessed: 2 February, 2024).

Stephen Harding, David Phillips, and Michael Patrick Fogarty. 1986. *Contrasting values in Western Europe: Unity, diversity and change*. Macmillan Publishing Company.

Eelco Harteveld and Markus Wagner. 2023. Does affective polarisation increase turnout? Evidence from Germany, The Netherlands and Spain. *West European Politics*, 46(4):732–759.

Jeremy Howells. 1996. Tacit knowledge. *Technology Analysis & Strategic Management*, 8(2):91–106.

Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4):1–38.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 3575–3585.

Tuuli-Marja Kleiner. 2018. Public opinion polarisation and protest behaviour. *European Journal of Political Research*, 57(4):941–962.

Aliyah Kurniasih and Lindung Parningotan Manik. 2022. On the role of text preprocessing in BERT embedding-based DNNs for classifying informal texts. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 13(6):927–934.

Bruce Lindsay. 2011. Social media and disasters: Current uses, future options, and policy considerations. Technical report, Congressional Research Service Washington, DC.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 35:1950–1965.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2019. Overview for the second shared task on language identification in code-switched data. *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, page 40–49.

Hans Moors, Lea Klarenbeek, Emily Berger, Michel Dückers, Menno van Duin, Gijs Kist, Marte Luesink, Tess Schrijvenaars, and Mary van der Wijngaart. 2022. 'Avondklokrellen': Lokale dynamiek in een mondiale crisis: Analyse van de voedingsbodem van de ordeverstoringen in vier Noord-Brabantse steden. *Technology Analysis & Strategic Management*.

J Eric Oliver. 2001. *Democracy in suburbia*. Princeton University Press.

Miles Osborne and Mark Dredze. 2014. Facebook, Twitter and Google Plus for breaking news: Is there a winner? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 611–614.

Damian Pascual, Sandro Luck, and Roger Wattenhofer. 2021. Towards BERT-based automatic ICD coding: Limitations and opportunities. *Proceedings of the BioNLP 2021 workshop*, pages 54–63.

Sasa Petrovic, Miles Osborne, Richard McCreadie, Craig Macdonald, Iadh Ounis, and Luke Shrimpton. 2013. Can Twitter replace Newswire for breaking news? In *Proceedings of the 7th International AAAI Conference on Web and Social Media*, volume 7 of *ICWSM 2013*, pages 713–716.

Swit Phuvipadawat and Tsuyoshi Murata. 2010. Breaking news detection and tracking in Twitter. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 120–123. IEEE.

Kamthorn Puntumapon and Kitsana Waiyamai. 2012. A pruning-based approach for searching precise and generalized region for synthetic minority over-sampling. In *Advances in Knowledge Discovery and Data Mining: 16th Pacific-Asia Conference, PAKDD 2012, Part II 16*, pages 371–382. Springer.

Sebastian Raschka. 2018. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint 1811.12808*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 3982–3992.

Rodrigo Ribeiro. 2013. Tacit knowledge management. *Phenomenology and the Cognitive Sciences*, 12:337–366.

Rebecca Roelofs. 2019. *Measuring generalization and overfitting in machine learning*. Ph.D. thesis, University of California, Berkeley.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? On the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135. Association for Computational Linguistics.

Michael Sandel. 2020. *The tyranny of merit: What's become of the common good?* Penguin Books: UK.

Mirko Tobias Schäfer, Aline Franzke, Danique van der Hoek, Marjolein Krijgsman, Iris Muis, Julia Straatman, Redmar Fransen, and Sammy Hemerik. 2022. De ethische data assistent handleiding: Inventarisatie van ethische kwesties rond data projecten bij overheden. *Utrecht Data School, Utrecht University*, pages 1–42.

Bart Schermer, Dominique Hagenauw, and Nathalie Falot. 2018. Handleiding Algemene verordening gegevensbescherming en Uitvoeringswet Algemene verordening gegevensbescherming. `https://www.rijksoverheid.nl/onderwerpen/privacy-en-persoonsgegevens/documenten/rapporten/2018/01/22/handleiding-algemene-verordening-gegevensbescherming` (accessed: 2 February, 2024).

Justin Sech, Alexandra DeLucia, Anna Buczak, and Mark Dredze. 2020. Civil unrest on Twitter (CUT): A dataset of tweets to support research on civil unrest. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 215–221. Association for Computational Linguistics.

David Shamma, Lyndon Kennedy, and Elizabeth F Churchill. 2010. Tweetgeist: Can the Twitter timeline reveal the structure of broadcast events? *CSCW Horizons*, 26.

David Snow, Louis Zurcher Jr, and Sheldon Ekland-Olson. 1980. Social networks and social movements: A microstructural approach to differential recruitment. *American Sociological Review*, 45:787–801.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72. Association for Computational Linguistics, Doha, Qatar.

Nicolás Somma. 2017. Discontent, collective protest, and social movements in Chile. *Malaise in Representation in Latin American Countries: Chile, Argentina, and Uruguay*, pages 47–68.

Kate Starbird and Leysia Palen. 2012. (How) will the revolution be retweeted? Information diffusion and the 2011 Egyptian uprising. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 7–16. Association for Computing Machinery.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2011. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, pages 1–14.

United States Attorney's Office. 2021. One year since the Jan. 6 attack on the Capitol. https://www.justice.gov/usao-dc/one-year-jan-6-attack-capitol.

United States Senate Committee on Homeland Security & Governmental Affairs. 2021. Examining the U.S. Capitol attack: A review of the security, planning, and response failures on January 6. https://www.hsdl.org/?view&did=854959.

Jianshu Weng and Bu-Sung Lee. 2011. Event detection in Twitter. In *Proceedings of the 5th International AAAI Conference on Web and Social Media*, volume 5, pages 401–408.

Matthew Williams and Pete Burnap. 2015. Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. *British Journal of Criminology*, 56(2):211–238.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. TwHIN-BERT: A socially-enriched pre-trained language model for multilingual tweet representations. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 5597–5607. Association for Computing Machinery.

David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. 2018. The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation. *ACM Transactions on Management Information Systems (TMIS)*, 9(2):1–29.

Koen van der Zwet, Ana Barros, Tom van Engers, and Peter 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(68):1–11.

A List of Used Python Libraries

Processing Step	Libraries
Pre-processing	re
	string
	pandas
	sklearn (TfidfVectorizer)
	nltk (word_tokenize, stopwords)
	spacy (nsubj, VERB)
	WordCloud
	datasets (Dataset, DatasetDict)
Training BERT models	codecs
	tqdm
	datasets (concatenate_datasets, load_dataset, Dataset, DatasetDict)
	pandas
	numpy
	sklearn (f1_score, roc_auc_score, accuracy_score, train_test_split)
	torch
	transformers (BertTokenizerFast, AutoTokenizer, AutoModelForSequenceClassification, TrainingArguments, Trainer, EvalPrediction, pipeline)
Training SetFit frameworks	sentence_transformers (CosineSimilarityLoss)
	setfit (SetFitModel, SetFitTrainer)
EOD Prediction	pycm
	emoji
	matplotlib
Additional testing	random

Table 6: Overview of the Python libraries used to train the models.

B Example Tweets and their Corresponding Label

Translated tweet from Dutch to English	**Label**	**Annotators' reasoning for the given label**
Half of the hooligans in #Rotterdam were underage!!!! Where are the parents????	EOD	* Usage of the word 'hooligans'. * Usage of 4 exclamation marks. * Indirect expression of discontent towards the parents of the hooligans, implying that they did not raise their kids correctly.
Only 3 wounded in Rotterdam from last night's riots? It is time for the police to take some shooting lessons...	EOD	* The 'Only 3 wounded [...]' subsentence has a sarcastic tone. * Suggesting that police officers should take shooting lessons, implies that the user wants the police to shoot at rioters and succeed.
The Austrian Baudet could not accompany the anti-vaccin protest. He was so ill from the corona-virus that he is staying at the hospital.	No discontent	* Without additional contextual information, it is unclear from the tweet itself who is meant with 'The Austrian Baudet'. * The tweet is too descriptive in order to determine the user's intent with certainty.
Has someone already called themselves in for the torn off finger ? #Rotterdam	No discontent	* A potential expression of discontent towards people who light fireworks. * Too unclear what is meant with a torn off finger.

Table 7: Overview of some example tweets with their corresponding label, including the reasoning used by the annotators to assign the 'Expression of Discontent' (EOD) or the 'No discontent' class. Although tweet examples 1 and 3 were relatively easy to label, tweet examples 2 and 4 were more difficult, causing annotators to have differing opinions on how to interpret the nuances in the text.

C Annotation Rules

A tweet was considered relevant or 'on-topic' for the EOD classification if:

1. The tweet refers to the protest for which it was scraped;
2. The tweet contains expressions of indignation towards the corresponding protest;
3. The tweet contains first-person observations of a protest and includes explicit disdain for the situation;
4. The tweet uses slurs, slang, and other inflammatory words to describe the opinions and actions of others (e.g. protesters, government);
5. The tweet uses expressive symbols like capital letters and punctuation (e.g. exclamation marks) to express their disdain towards the situation at hand;
6. The tweet shows support for the incitement of violence (towards any person or groups of people).

A tweet was considered irrelevant for the EOD classification, hence given 'No discontent', if:

1. The tweet contains solely observations regarding the situation at hand or the general public;
2. The tweet contains the person's own opinion, but the person highlights the perspectives from both sides, e.g., the protesters and the government;
3. The tweet contains expressions of confusion, e.g., towards what and why the protests are happening;
4. The tweet seems to contain sarcasm but it could be interpreted in multiple ways;
5. The tweet includes discussions about the topic at hand whereby the protest is used to support one's non-inflammatory opinions.

A tweet was excluded from the dataset, hence considered 'off-topic', if:

1. The tweet refers to a different protest for which it was scraped;
2. The tweet is a response to another tweet potentially related to the protest, but the content of the considered tweet does not refer to the protest;
3. The tweet contains signs of discontent towards parties relevant in protests (e.g. police, protesters), but it is not explicitly concerning the protest for which it was scraped.

D Extra Results

						Expression of discontent			No discontent		
Model	**Data**	**FT**	**PP**	**ACC**	**AUC**	**P**	**R**	**F1**	**P**	**R**	**F1**
LR	ALL	N	Y	.867	.629	.577	.293	.389	.890	.964	.925
LR Best PP Step †	MA	N	Y	.872	.642	.611	.318	.418	.893	.966	.928
LR	MA	N	N	.867	.646	.570	.335	.422	.895	.957	.925
BERT	MA	Y	N	.861	.541	.647	.091	.159	.866	**.992**	.924
BERTje	MA	Y	Y	.855	.537	.489	.091	.153	.865	.984	.921
BERTje	MA	Y	N	.882	.672	.664	.376	.480	.902	.968	.934
BERTje	ALL	Y	N	.862	.658	**.863**	.328	.475	.862	.988	.920
BERTje Emojis	MA	Y	N	.885	.672	.687	.372	.483	.901	.971	.935
BERTje Overs & Emojis ‡	MA	Y	N	.886	.686	.681	.405	.508	.906	.968	.936
BERTje Oversampling	AM	Y	N	.791	.755	.691	**.653**	.672	.835	.858	.846
BERTje Oversampling	EI	Y	N	**.902**	.553	.391	.127	.191	.918	.980	**.948**
BERTje Oversampling	MA*	Y	N	.899	.752	.667	.549	.602	**.929**	.956	.942
BERTje Oversampling	MA	Y	N	.876	.706	.592	.467	.522	.913	.945	.929
BERTje '22 Oversampling	RO	Y	N	.861	.633	.667	.294	.408	.876	.971	.921
BERTje '23 Oversampling	RO	Y	N	.816	.660	.571	.395	.467	.856	.924	.889
mBERT	MA	Y	N	.877	.698	.603	.446	.513	.910	.950	.930
mBERT	AM	Y	N	.799	.720	.777	.506	.613	.804	.933	.864
mBERT	AM	Y	Y	.779	.684	.833	.407	.547	.769	.960	.854
mBERT Oversampling	AM	Y	N	.798	**.759**	.711	.646	**.677**	.835	.872	.853
mBERT Overs & Emojis	AM	Y	N	.803	.752	.746	.606	.668	.823	.899	.859
ST EN	MA	Y	N	.855	None	None	None	None	.855	None	.922
ST EN	MA	N	N	.812	.526	.227	.124	.160	.862	.929	.894
ST EN Oversampling	AM	Y	N	.745	.708	.598	.605	.602	.815	.810	.812
ST EN Oversampling	MA	Y	N	.836	.725	.447	.570	.501	.924	.881	.902

Table 8: Comparison between the models from fine-tuning (FT) or not (Y and N, respectively), pre-processing (PP) the data or not (Y and N, respectively) for the prediction type *Expression of Discontent* (EOD). Some models were given a particular focus, e.g. emojis and oversampling the minority class. Highest scores on accuracy, precision, recall, F1 and AUC are in bold. The numbers are rounded. Abbreviations 'AM', 'EI', 'MA', 'RO' and 'ALL' stand for the Black Lives Matter (Amsterdam), curfew riots (Eindhoven), Black Pete (Maastricht), fireworks ban protest (Rotterdam) and all four datasets, respectively. Notes: † LR was run with the combination of pre-processing steps that provided the best results, and ‡ BERTje was run by combining the focus on emojis with oversampling. By default, MA refers to a subset of the MA dataset, though MA* refers to the full dataset. When a model is marked with 'emojis', we run the model on a subset of the MA dataset solely containing tweets with at least one emoji. This subset was around 12% of the original dataset's size.

				Expression of discontent			No discontent		
Model	**Run**	**ACC**	**AUC**	**P**	**R**	**F1**	**P**	**R**	**F1**
Bernice	1	.875	.792	.667	.659	**.663**	.922	.925	.923
Bernice	2	.870	.772	.646	.619	.632	.917	.925	.921
Bernice	3	.865	.793	.623	.679	.649	.926	.907	.917
Bernice	4	.869	.776	.662	.625	.643	.914	.926	.920
Bernice	5	.868	.785	.660	.649	.654	.917	.920	.919
BERTje	1	.860	.721	.674	.497	.572	.890	.944	.916
BERTje	2	.859	.745	.673	.555	.609	.895	.934	.914
BERTje	3	.872	.756	.689	.570	.624	.905	.941	.923
BERTje	4	**.877**	.754	**.720**	.558	.629	.904	**.950**	**.926**
BERTje	5	.864	.749	.670	.561	.611	.900	.935	.917
TwHIN-BERT-base	1	.853	.777	.620	.651	.635	.914	.903	.908
TwHIN-BERT-base	2	.861	.781	.637	.650	.643	.916	.911	.914
TwHIN-BERT-base	3	.858	**.800**	.611	.707	.656	**.928**	.894	.911
TwHIN-BERT-base	4	.864	.778	.647	.639	.643	.915	.918	.916
TwHIN-BERT-base	5	.860	.776	.637	.639	.638	.914	.913	.913
TwHIN-BERT-large	1	.820	None	None	None	None	.820	None	.901
TwHIN-BERT-large	2	.830	.792	.551	**.731**	.628	**.928**	.854	.890
TwHIN-BERT-large	3	.810	None	None	None	None	.810	None	.895
TwHIN-BERT-large	4	.815	None	None	None	None	.815	None	.898
TwHIN-BERT-large	5	.864	.753	.650	.576	.611	.906	.929	.917
ST	1	.864	.761	.694	.589	.637	.899	.934	.916
ST	2	.870	.764	.656	.596	.625	.912	.931	.921
ST	3	.871	.766	.665	.601	.631	.912	.932	.922
ST	4	.874	.771	.690	.605	.644	.911	.937	.924
ST	5	.873	.767	.704	.594	.644	.906	.940	.923
LR	1	.820	.718	.532	.552	.542	.892	.884	.888
LR	2	.797	.704	.522	.542	.532	.875	.866	.870
LR	3	.812	.706	.558	.523	.540	.875	.889	.882
LR	4	.804	.706	.516	.541	.528	.882	.871	.877
LR	5	.805	.704	.523	.534	.529	.879	.875	.877

Table 9: Comparison between the models for EOD prediction with oversampling of the minority class. The models are run five times in order to get insight into the range of the possible scores. The numbers are rounded, and the best scores per metric are in bold.

				Expression of discontent			No discontent		
Model	**Data**	**ACC**	**AUC**	**P**	**R**	**F1**	**P**	**R**	**F1**
BERTje	AM	.808	.755	**.765**	.600	.673	.822	.910	.864
Bernice	AM	.823	**.806**	.703	**.760**	**.731**	.885	.852	.868
TwHIN-BERT-large	AM	.637	None	None	None	None	.637	None	.778
TwHIN-BERT-base	AM	.817	.782	.721	.687	.704	.858	.877	.868
BERTje	EI	.901	None	None	None	None	.901	None	.948
Bernice	EI	.902	.616	.513	.260	.345	.923	.973	.947
TwHIN-BERT-large	EI	.901	None	None	None	None	.901	None	.948
TwHIN-BERT-base	EI	**.909**	.591	.625	.195	.297	.918	**.987**	**.951**
BERTje	MA	.867	.595	.519	.222	.311	.888	.968	.926
Bernice	MA	.872	.723	.527	.519	.523	**.925**	.927	.926
TwHIN-BERT-large	MA	.865	None	None	None	None	.865	None	.928
TwHIN-BERT-base	MA	.874	.644	.559	.328	.413	.901	.960	.930
BERTje	RO22	.885	.680	.639	.397	.489	.908	.964	.935
BERTje	RO23	.869	.761	.677	.587	.629	.907	.935	.920
Bernice	RO22	.882	.718	.594	.491	.538	.920	.946	.933
Bernice	RO23	.877	.760	.717	.573	.637	.905	.947	.926
TwHIN-BERT-large	RO22	.894	.725	.663	.491	.564	.921	.960	.940
TwHIN-BERT-large	RO23	.811	None	None	None	None	.811	None	.896
TwHIN-BERT-base	RO22	.875	.675	.575	.397	.469	.907	.953	.929
TwHIN-BERT-base	RO23	.866	.728	.704	.507	.589	.892	.950	.920

Table 10: Comparison between the models Bernice, BERTje, TwHIN-BERT-base and TwHIN-BERT-large for the EOD prediction across all four datasets without oversampling the minority class. The numbers are rounded, and the best scores per metric are in bold.

				Expression of discontent			No discontent		
Model	**Run**	**ACC**	**AUC**	**P**	**R**	**F1**	**P**	**R**	**F1**
Bernice	1	**.878**	**.788**	.668	.647	**.657**	**.923**	.929	**.926**
TwHIN-BERT-base	1	.858	.778	.630	**.648**	.639	.915	.909	.912
BERTje	1	.865	.728	.715	.504	.591	.889	.952	.919
BERTje	2	.868	.744	.699	.544	.612	.898	.945	.921
BERTje	3	.875	.776	.681	.619	.649	.914	.933	.924
BERTje	4	.869	.717	.757	.469	.579	.884	.964	.923
BERTje	5	.872	.718	**.776**	.469	.585	.885	**.968**	.925
BERTje	AVG	.870	.737	.726	.521	.603	.894	.952	.922

Table 11: Comparison between the models Bernice, BERTje and TwHIN-BERT-base for EOD prediction without oversampling. For BERTje, five runs were completed in order to get insight into the range of the possible scores. The numbers are rounded, and the best scores are in bold.

				Expression of discontent			No discontent		
Model	**Data**	**ACC**	**AUC**	**P**	**R**	**F1**	**P**	**R**	**F1**
BERTje	AM	.814	.770	.762	.639	.695	.834	.901	.866
BERTje	EI	.901	None	None	None	None	.901	None	.948
BERTje	MA	.868	.678	.568	.410	.476	.904	.947	.925

Table 12: Test runs with BERTje for EOD prediction without oversampling on three separate datasets, namely AM, EI and MA. The numbers are rounded.