**To Remain Peaceful or Not? Using News to Predicting Protestor Behaviours.**

**Mason Ricci**

***Trinity College Dublin, The University of Dublin***

***https://github.com/riccimason99/Dissertation\_in\_progress***

***riccim@tcd.ie***

**Abstract**

Recent protests in the United States have resulted in the destruction of property, police brutality, violence between protesters, and, in some cases, death. An existing body of literature has successfully forecasted the time and location of protests; however, few attempts have been made to predict the nature of the demonstration. This paper analyzes news articles four days before a protest occurs in an effort to predict whether the gathering will resolve peacefully or not. After implementing several text analysis methods, including embedding models, to relevant news articles preceding one hundred past U.S. protests, I determined that sentiment scores provided the most valuable insight regarding the outcome of a protest. Logistic regression determined that news articles which precede non-peaceful protests have lower sentiment scores than those preceding peaceful protests. Sentiment scores were then applied as features to several machine learning classification algorithms, which correctly classify out-of-sample protests with an F1 score of .63. The findings of this study have significant implications for law enforcement agencies and property owners, allowing them to take early action to prevent dangerous situations and the destruction of property.

**Introduction**

This study uses one hundred protests from across the United States between January 2021 and May 2024 as case studies and examines news articles from up to four days before each protest occurred. The cases include fifty peaceful and fifty non-peaceful protests, which were held for various reasons. Several text analysis methods were applied to each body of text, such as word embeddings, document embeddings, and extracting sentiment scores. Subsequently, the outputs of each text analysis method were used as features in machine learning classification models to detect if a protest was peaceful or non-peaceful. This study aims to identify the text analysis tool and classification algorithm that best predicts out-of-sample cases.

The remainder of this paper is organized into several sections. The first section will consist of a literature review to explore previous efforts in forecasting protests. After that, we will identify all data sources used. This will be followed by an explanation of the text analysis and machine learning models that were implemented, along with their results. The paper will then delve into the study's limitations and considerations for future research. The final sections will summarize key findings and their implications.

**Related Works**

The concept of forecasting conflict and civil unrest using news and social media has been widely used by intelligence communities across the globe for some time (Chadefaux 2014; Leetaru 2011). Over the past decade, this area of study has received increased attention from academia (Cadena et al. 2015; Leetaru 2011; Congyu & Gerber 2017; Goode et al. 2015; Pinckney and RezaeeDaryakenari 2022). Most studies have focused on predicting the time and location of protests and sparsely address whether they will be peaceful or not. This study aims to address this gap in the literature.

This literature review will begin by briefly exploring past efforts to identify the causes of protests. It will then examine academic and private sector attempts at forecasting protests, highlighting several relevant studies. The final section will outline how my research integrates with and contributes to the existing publications.

**Causal Inferences:**

The body of literature that attempts to determine the causation of civil unrest has a far stronger foundation than literature that attempts to forecast protests. Several seminal works have examined historical protests and revolutions using qualitative methods, identifying factors such as economic decline after a period of consistent growth (Davies 1962), the expansion of political opportunities (McAdam 1999), and media attention (McCarthy and Zald 1977) as causal factors. More recent research has utilized quantitative methods, namely logistic regression, to suggest that the modernization of the manufacturing industry (Butcher 2016) and minority marginalization (Thurber 2018) can cause unrest. This body of literature has been essential in paving the way for the more recent exploration of short-term protest forecasting.

**Short-term****Forecasting**:

Unlike theory-based research that explores causality, a significant growing body of literature focuses strictly on predicting the time and location of protests. While multiple publications have demonstrated the ability to predict protests, five articles are particularly relevant to my research due to their methodologies and findings; they will be discussed in further detail.

One of the first studies that attempted to forecast protests was conducted by Leetaru (2011), who found that media sentiment declines significantly before an uprising. My study is inspired by Leetaru's research and builds on his methodologies. Leetaru's paper, however, only identifies media sentiment as a warning signal without forecasting a particular time or location.

Notably, my study is based on the premise that existing methods exist to accurately predict a protest's time, location, and purpose. This assumption is supported by several articles (Bastos 2015; Congyu 2017; Goode 2015; Pinckney 2022; Steinert 2015), with three being explored in detail. (Cadena et al. 2015; Kallus 2014; Ning et al. 2016).

One year after Leetaru published his paper, a team of academic and private researchers launched the EMBERS Project, creating an automated 24/7 monitoring and forecasting system to predict protests in several South American countries (Cadena et al. 2015). Using a variety of open-source indicators such as news, social media, and economic trends, EMBERS has successfully predicted several violent and non-violent protests. Importantly, EMBERS has proven its ability to forecast not only the time and location of a protest four days before it occurs but also the reason for the protest. Understanding the cause for a demonstration is crucial for my study, as only news articles relevant to the protest in question are used.

The ability to predict the time, location, and reasons for protests is also supported by Kallus (2014), who used mass open-source data, including new articles, applied to a random forest machine learning classification algorithm to predict a city that will have a violent protest three days in advance. While the number of use cases is relatively small, Kallus' study shows significant promise, and I implemented similar methods in my own research.

A third article that demonstrates the ability to forecast protests is by Ning et al. (2016). In his study, news stream posts are divided by day into "bags." Each bag is transformed into a vectorized representation using the document embedding model Doc2Vec, wich will befurther explained later. Utilizing a "nested structure within the multi-instance learning paradigm," the study accurately predicts protests occurring in Mexico and several South American countries up to five days in advance. My study uses Ning's work as a basis to support the claim that protests can be predicted and draws inspiration from the methods she employed.

While the aforementioned articles have not had total success in forecasting protests, they demonstrate the capability of determining when and why a protest will occur. The three mentioned studies are highly relevant to my area of research due to their topic and methods; however, they strictly forecast the location and time of demonstrations without considering whether the event will become violent or resolve peacefully.

Only one article specifically attempts to predict the violent potential of protests. A 2018 study, also published by Ning et al., demonstrates that it is possible to predict if protests will become violent. His research analyzes various texts preceding both violent and non-violent protests in several South American nations. An ensemble of machine learning algorithms is employed to predict the likelihood of a protest turning violent; however, only a few use cases are observed.

**Contribution:**

1) As mentioned, only Ning et al.’s 2018 piece addresses the same base question as my study and uses some similar methodologies; however, our research has noteworthy differences. My article uses significantly more use cases and specifically identifies news articles as predictors. Furthermore, my study focuses on the U.S., which has notably received less attention in protest forecasting compared to regions such as South America, North Africa, and the Middle East.

2) The study uses several text-processing techniques for classifying news articles and compares their results. Interestingly, I find simple lexicon-based sentiment scores to be the most effective features for machine learning algorithms. This suggests that sentiment analysis should not be overlooked for more commonly used classification methods, such as embeddings.

3) I am pleased to make available the tailored dataset used for this project. While I am not able to provide the full news articles due to copyright reasons, a cleaned and tokenized version of the article is included. The dataset also includes information about the protest which the articles proceed, such as the number of participants, organizations involved, and a short description of the demonstration and its consequences. All information regarding the protest comes from the Armed Conflict Location & Event Data Project, as will be further explained later in the article.

4) Finally, this study will contribute by providing a Doc2Vec model trained on several hundred relevant news articles. To my knowledge, a model trained on such a specific set of text has not previously been publicly available and could be useful to feature researchers. The Doc2Vec model, dataset, and all other reproduction materials have been uploaded to Git Hub (see URL on title page)

**Data Sources**

**Use Cases:**

This study utilizes data from the Armed Conflict Location & Event Data Project (ACLED) to facilitate the identification of use cases. ACLED is an established non-governmental organization that publishes reliable data detailing events such as war and civil unrest. The ACLED pinpoints the location, size and timing of the demonstration and notes if the protest was peaceful. This information will be vital when selecting use cases and gathering relevant news articles.

Four key factors were considered when selecting use cases: whether the protest was peaceful or not, the reason for the protest, the number of participants, and the availability of relevant news articles. Each of these factors will be discussed in detail.

Fifty peaceful and fifty non-peaceful protests occurring between January 2021 and May 2024 were selected, creating a balanced and contemporary dataset. The ACLED defines peaceful protests as "non-violent demonstrations, typically involving unorganized actions by society members." This criterion gives a clear standing to designate a peaceful protest; however, the definition of non-peaceful protests is not as straightforward, as it includes demonstrations where isolated incidents occur. For example, if one protester was arrested for disorderly conduct, the event might be labeled non-peaceful even if no other incidences occurred. To address this, I have specified non-peaceful protests to be those that resulted in over 15 arrests, involved the deployment of tear gas, resulted in property destruction, or involved violence between protesters and police. Protests that led to arrests for obstruction, such as blocking traffic, were not categorized as non-peaceful, regardless of the number of arrests. These criteria were set for two reasons. Firstly, they give a clear definition of the term "non-peaceful protest" that is adapted for this study. Secondly, the definition filters irrelevant protests while still allowing enough use cases for a compelling argument to be made.

The study includes protests that occur for various reasons, such as the Gaza conflict, abortion, black lives matter (BLM), labor unions, the environment, the Donald Trump indictment, immigration, LGBTQ+, and Ukraine. An early concern was that the potential for a protest to be peaceful or not might depend heavily on why the protest occurred. For instance, a BLM protest might be more likely to escalate into violence than a labor union protest. In this case, the algorithm model may classify a protest based on the protesters' cause rather than on underlying cues that appear in the text, rendering the model useless for practical purposes. To address this, both peaceful and non-peaceful protests were included for each protest cause. For example, the study included both peaceful and non-peaceful BLM protests, as well as both peaceful and non-peaceful labor protests.

Another crucial factor in selecting use cases was the number of participants in the protest, as studies have shown that smaller protests are less likely to devolve into violence (Ning et al., 2018). All protests with fewer than 100 participants were excluded; this threshold was chosen to remove smaller protests while still providing enough cases to make an effective study. Protest participation was based on ACLED estimates, which are derived from reports by reputable news sources. While accurately gauging protest attendance remains notoriously difficult (Sobolev et al., 2021), ACLED provides the most reliable data available. To maintain consistency, protests lacking participant estimates were omitted from the study.

A final consideration for selecting use cases was the availability of relevant news articles. As mentioned, this study assumes that a protest's where, when, and why are known four days before it occurs. If no relevant news articles were available, the event was not included. The following section details how relevant articles were obtained for each demonstration.

**News Articles:**

Since the 70s, academics have theorized that news reports extend beyond mere fact reporting; they reflect cultural trends that can provide valuable insights (Gerbner and Marvanyi, 1977); however, only in the past two decades have quantitative text analysis methods confounded these claims. With the advent of open-source text analysis methods, news reports have been utilized to forecast various events, including economic trends (Barbaglia, Consoli, and Manzan 2023, 708-719), tourism demand (Park, Park, and Hu 2021, 103273), and oil prices (Li et al. 2016, 1081-1087).

I compiled a dataset of new articles by scraping the LexisNexis News and Business website, which contains publications from diverse press agencies. Articles from local, national, and international press firms were used to avoid bias and construct a more well-rounded dataset. Notable firms used in the study includeThe Santa Fe New Mexican, The New York Times, and The Guardian. The study by Ning et al. (2018) suggests that signs of civil unrest often appear across national media rather than just the city in which the protest occurred, furthering the importance of surveying various local and national press firms.

For each protest, a minimum of four articles were collected. If a protest was reported on before it occurred, those articles were prioritized for use. Alternatively, articles were used that cover prior demonstrations that happened for the same reason as the demonstration in question. For example, the body of news articles preceding the Columbia University pro-Palestine protest includes publications detailing other on-campus pro-Palestine demonstrations. Several articles have shown that past protests are indicators for feature demonstrations (Cadena et al. 2015) (Ning et al. 2018).

To identify relevant articles, I utilized the LexisNexis keyword search feature, which allows for filtering based on the presence or absence of specific words in the publication. For instance, for a healthcare worker protest in Oregon, I searched "protest AND labor AND healthcare AND NOT Pakistan," while for the UC Berkeley pro-Palestine protest, the search terms were "protest AND gaza AND campus." Efforts were made to maintain a consistent structure for keyword searches to prevent bias; however, deviating from the formula to achieve better search results or exclude irrelevant articles was occasionally necessary. Additionally, LexisNexis provides the option to filter based on the date range, enabling the selection of articles from the desired time frame. After scraping the text, I combined the articles from the four days before the protest into one "document"; each document was given one of two labels, 0 for peaceful and 1 for non-peaceful.

**Methodologies**

Determining how to represent text as features is of considerable importance; the next sections detail several methods used to numerically represent documents and how useful each output was when applied to machine learning classification algorithms.

**Sentiment Analysis:**

Sentiment analysis was conducted using VADER (Valence  Aware  Dictionary for sEntiment Reasoning), a simple yet effective open-source sentiment analysis tool. VADER uses a lexicon-based approach, meaning it references a "dictionary" of words in which each word is scored on a positive or negative range (Liu 2010, 627-666).  Negations and degree modifiers are also accounted for; as an example, "very bad" would be given a more negative score than "bad," and "not bad" would be given a positive sentiment score (Chauhan et al., 2018: p. 487). After exposing a document to VADER, positive, negative, neutral, and "compound" scores are returned. Compound signifies the document's overall score, from -1 being the most negative to 1 being the most positive. Recall that a document consists of the complete body of articles preceding the protest in question.

VADER has been selected for this project because it is considered the gold standard of lexicon-based sentiment analysis tools, having been validated by and often outperforming human raters (Bonta, Kumaresh, and Janardhan 2019, 1-6). While VADER was designed for conducting sentiment analysis on social media posts, it has been shown to be effective in extracting sentiments from news articles (Agarwal 2020; Shapiro, Sudhof, and Wilson 2022) and for using such sentiments as features in machine learning classification models (Gomes et al. 2013, 1-6).

Logistic Regression:

Before applying sentiments as features to machine learning classifiers, binary logistic regression was conducted to identify if a relationship exists between negative sentiments and the potential that a protest will be non-peaceful. The dependent variable of the regression was peaceful or non-peaceful. The independent variable was the negative sentiment score, a continuous variable taking any values between 0 and 1; values closer to 1 signify articles with more negative sentiments, and 0 with no negative sentiment. A constant intercept is included with the independent variable, as recommended when using the statsmodels library in Python (Seabold and Perktold 2010, 94). As noted in the literature review, it is believed that protests can influence or trigger one another, violating the logistic regression assumption of independent observations. However, since this model is intended solely to identify correlations rather than establish causation, this issue is not a significant concern.

The regression results display a positive relationship between the "negative" variable and the potential for non-peaceful protest. In simple terms, articles preceding non-peaceful protests tend to have more negative sentiments than those preceding peaceful protests. Figure 1 displays that, while close, the average negative sentiment score of articles preceding non-peaceful protests is higher than those preceding peaceful protests. While the coefficients of the regression are not statistically significant (p-value  .07), enough of a relationship exists to warrant moving forward with using sentiments as features in classification algorithms.

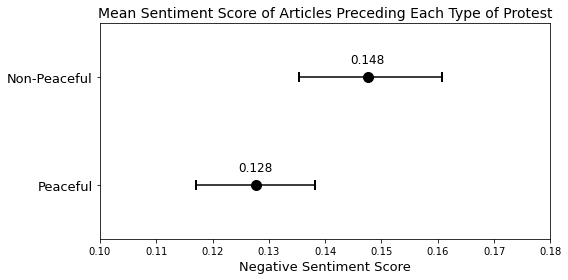
****

Figure 1: The dot displays the average sentiment scores for all articles in that category, and the black lines are confidence intervals. Note that higher values on the x-axis correspond to more negative sentiments.

Classification with Sentiment Scores:

After identifying a correlation between negative sentiments and protest outcomes, I applied sentiment scores as features in machine learning classification algorithms. The data was divided into training (80%) and test  (20%), with model selection, hyperparameter tuning and, feature selection being determined based on the results of k-fold cross-validation within the training dataset using three folds. Several classification algorithms were tested, including K Nearest Neighbors, Logistic Regression, and Linear Discriminant Analysis. Each model's effectiveness was measured using the F1 metric, as it provides an unbiased representation of false positives and false negative classifications. Given that the dataset is balanced with 50 peaceful and 50 non-peaceful use cases, the baseline performance for the model would be an F1 score of .50.

Of the model and feature combinations tested, the Gaussian Naive Bayes algorithm with compounds as the sole feature returned the best results. The model achieved an average F1 score of .69 across the three folds in the training set. When applied to the remaining 20% of the data, the model achieved an F1 score of .63. This score, significantly higher than what could be achieved by chance, demonstrates that the sentiments of articles can be used to determine if a protest will become violent. Figure two displays a balance between false negative and false positive classification, suggesting that the model is generalizable and balanced.

A blue squares with white text

Description automatically generated

Figure 2: Confusion matrix of Gaussian Naive Bayes algorithm on the test dataset using compound sentiment score as the sole feature.

**Word and Document Embeddings:**

Despite significant efforts to use embeddings to represent news articles, these features did not prove as effective as sentiment scores in classifying documents. The remainder of this section briefly explains what word embeddings are and how the text was processed before application to embedding models. The following two sections detail the processes used for Word2Vec and Doc2Vec models, as well as the results when the embeddings were used as features for text classification.

Embeddings are vector representations of words or documents. Unsupervised models, such as Word2Vec, are trained on large corpora of text and use neural networks to create multi-dimensional vector representations of words. This approach allows for a more nuanced understanding of language as models grasp the subtleties and relationships between words beyond simple frequency counts.

While the traditional Bag of Words (BoW) method has been useful in its own right, it fails to capture word order and relationships between similar words. For example, all synonyms are treated as entirely separate words, leading to a loss of meaningful information (Jurafsky and Martin 2020, 100-105).  Embeddings capture semantic meaning by placing similar words closer together in the vector space. For example, the words "king" would be close to "queen" and "royalty," reflecting their similarity (Mikolov et al. 2013, 3111-3119).

Before implementing the embedding model, all documents are pre-processed. This involved the removal of URLs, hyperlinks, numbers, punctuations, and stop words such as "the" and "and" using the NLTK stopword dictionary (several other words were removed as necessary, such as "said"). Stemming was also performed to transform a word to its root. Stemming helps normalize words, treating variations of a word as a single item; for example, the words "running," "runner," and "ran" can all be reduced to the root word "run." While stemming is not expressly necessary, a study by Truşcă (2019, 496-503) has shown that the process produces better embeddings.

It is also essential that the model ignores words that appear only a few times in the entire training set. Unexpected words can lead the model to represent those words without as much context, leading to unpredictable results. Words that appear at least five times are filtered out of the dataset; this threshold has been shown to be particularly effective when classifying news articles (Dogru et al., 2021).

Word Embedding using Word2Vec:

When extracting word embeddings, it is recommended to use a model that has been trained on text similar to the documents which the model will be applied to. This enables the model to capture domain-specific language and context (Asudani, Nagwani and Singh, 2023, 10345-10425). While several embedding models exist, Google's Word2Vec model is best suited for the purposes of this study as it has been extensively trained on Google News articles.

Word2Vec models cannot be directly applied to an entire document and fed into a machine learning algorithm, as each document would have a different length depending on the number of words. To resolve this, I used the weighted average method, which involves calculating the average of all the vectors in the document and generating a single vector (Mikolov et al., 2013). Averaging has shown surprisingly good results; however, weighted averages take the process a step further. Weighted averages assign more significance to certain words based on

TF-IDF (Term Frequency - Inverse Document Frequency). This process assigns a weight to each word based on its frequency compared to the total number of words in a document (Djaballah et al., 2019). While this method is useful, it does not consider word order, potentially causing the model to miss important context.

Word2Vec has two main training algorithms: Continuous Bag of Words (CBOW) and skip-gram. CBOW predicts the vector for the target word based on surrounding words, whereas skip-gram predicts the surrounding words based on the target. Both approaches were tested, with the final model using CBOW.

Several hyperparameters for the Word2Vec model were tested based on recommendations by Rong (2014). The window parameter, which specifies the number of context words considered for each target word, was found to be most effective with a window size of five. For document vectors to be used as features in machine learning classification models, they must be of uniform size; hence, determining the appropriate vector size is crucial. Limiting the vector size might result in the loss of important information, whereas excessively large vectors could lead to overfitting. After reviewing studies by several other academics (Dogru et al., 2021; Kim et al., 2015) and experimenting with multiple sizes to observe their performance in classification models, I decided to set the vector length to 300.

Various machine learning classification models were used in an attempt to achieve the best out-of-sample results, including Linear Discriminant Analysis, SVM and Random Forest. Despite all efforts, the Word2Vec mode performed poorly, with the best out-of-sample F1 score of approximately 0.54 using the Support Vector Machine classifier. Recall that the baseline score was .50, while sentiments achieved a score of 0.64. Further model details can be found in the reproduction materials, which have been uploaded to GitHub.

Document Embeddings using Doc2Vec:

Considering the inherent shortcomings and poor performance of the weighted average Word2Vec models, I transformed articles into document embeddings in an effort to achieve better results. The Doc2Vec extends the Word2Vec model by generating vector representations for entire documents rather than individual words. While Word2Vec predicts a word's vector based on its sentence context, Doc2Vec captures the semantic meaning of an entire document and returns a fixed-length vector. Importantly, unlike Word2Vec averaging, Doc2Vec considers the order of words within a document (Truşcă et al., 2019). This capability makes Doc2Vec particularly effective for tasks like document classification, where grasping documents' broader context and content is crucial (Le and Mikolov 2014, 1188-1196).

While Doc2Vec offers significant advantages over Word2Vec, it is important to note that far fewer pre-trained models are available. This scarcity, combined with the fact that many existing models are based on outdated versions of the Gensim library and lack thorough documentation, compelled me to train my own model. The Doc2Vec model employed in this study was trained using the training portion of the dataset, which consists of approximately 500 news articles. While this dataset is considered small by most standards, it is bolstered by the fact that all documents are of the same type and cover similar topics, enabling the model to be more specialized (Quoce, 2014). Furthermore, research by Leu et al. (2016) demonstrates that Doc2Vec models trained on relatively small datasets (e.g., 400 articles) can still produce satisfactory results. The model will be made available on GitHub with all other reproduction materials.

To build its vocabulary, the model will iterate over all documents a prescribed number of times; iterations or "epochs" are typically set between ten and twenty when training on large datasets. For smaller datasets, it can be useful to use more iterations to help build context with the drawback of potentially having an overfit model (Le and Mikolov 2014, 1188-1196). Forty iterations were used to build the vocabulary for the model used in this study.

As with the Word2Vec model, several hyperparameters were configured based on Rong's (2014) recommendations. Similar to Word2Vec models, two algorithms exist to produce document embeddings. Both the Distributed Bag of Words (DBOW) and Distributed Memory (DM) methods were tested; the final model uses DBOW with a window of five and a vector size of three hundred. A range of machine learning classification models were employed to optimize out-of-sample performance, with the best results coming from the Support Vector Machine classifier.

Despite efforts to use embeddings from various adaptations of my Doc2Vec model as features in different machine learning algorithms, the out-of-sample F1 score was only approximately 0.56, marginally higher than the weighted average Word2Vec score of 0.54.

Why Poor Results?

The poor performance of machine learning models using document and word embeddings compared to sentiment scores was unexpected. It appears that embeddings derived from articles preceding peaceful protests share language and themes similar to those preceding non-peaceful protests, making it difficult for the models to distinguish between them. Figure three identifies that using embeddings as features leads to the model overclassifying articles as non-peaceful.  In any case, the sentiments of these articles provide significant insights that the embedding models cannot capture. Additional research is needed to determine the exact reason for this; however, I will leave that for future studies.

A blue squares with white text

Description automatically generated

Figure 3: Confusion matrix displaying the classification of articles when using word and document embeddings as features.

**Shortfalls and Considerations**

While this paper has demonstrated extensive efforts to use cutting-edge methods and address various theoretical challenges, it is essential to acknowledge areas where the study could be improved for transparency and to guide future research.

A theoretical improvement to the study would involve changing the target variable of the machine learning models from binary to categorical. Currently, the target variable takes on a value of 0 for peaceful and 1 for non-peaceful and does not consider different levels of non-peaceful protest. Ideally, several categories would exist, each with its own qualifications, allowing the model to predict the degree of violence or destructiveness of a protest. To implement this improvement, a significantly larger number of use cases would need to be observed, spanning further back in time.

Another limitation is that this study assumes all protests' time, location and reason have been accurately predicted, as any practical uses are nullified if these factors are unknown. While academics have demonstrated a significant capability to forecast spontaneous protests, there are still areas for improvement. Furthermore, this approach to protest prediction would likely be ineffective in nations with limited free speech, as news articles might not convey true sentiments or fail to report on events entirely.

A methodological improvement to the study would be to use a more thoroughly trained Doc2Vec model. Currently, document embedding representation of text is one of the cutting-edge methods for text classification; however, few pre-trained models are openly available. While this study contributes a model trained on approximately 500 news articles, further training would enhance its robustness and benefit feature researchers.

Although considerable attention was given to tuning classification models, including referencing various studies, no systematic method, such as grid search, was employed for hyperparameter tuning. Consequently, it is likely that better results could be achieved, particularly when using embeddings as features, as their high dimensionality would lend itself to more complex classification models.

A final improvement to the study would be to incorporate both sentiment scores and embeddings as features in the classification algorithm. This approach would necessitate weighting the sentiment scores more heavily to ensure the multi-dimensional vectors produced by embeddings do not overshadow them. By doing so, the model could leverage the strengths of both features, potentially enhancing predictive results.

**Key Findings and Conclusion**

In this study, I implement several text analysis and classification methods and demonstrate the ability of news sentiments to predict whether a protest will be peaceful or not. I empirically evaluated the effectiveness of these methods using a case study of one hundred U.S. protests.

Surprisingly, word and document embeddings were less effective than sentiments, achieving out-of-sample F1 scores of .54 and .56, respectively, compared to sentiments, which attained an F1 score of .63. This finding highlights that sentiment analysis captures crucial information that embedding models may overlook, and should be considered as a tool when conducting document classification. It also underscores an important but often overlooked aspect of applied machine learning: simpler approaches sometimes outperform complex methodologies and advanced models.

Anticipating if a protest will be peaceful is invaluable for property owners, private citizens, and law enforcement. This foresight allows property owners to protect possessions and businesses, potentially reducing damage and financial loss. Private citizens can use this information to assess personal risks and make informed decisions about their activities. Law enforcement agencies benefit by allocating resources more effectively, planning strategic responses, and potentially communicating with protest organizers to increase safety. Ultimately, accurate forecasting of protest behaviour enhances community preparedness, decreases the likelihood of violence, and promotes a more coordinated approach to managing public demonstrations.

It is important to note that this study is not intended to serve as a standalone forecasting tool. Rather, it highlights news sentiments as a valuable warning for anticipating protest behaviour. This paper lays the groundwork for further research, which aims to develop more comprehensive tools to forecast protests both in the United States and globally. In future work, I plan to incorporate additional factors, such as social media activity, to enhance model performance.

Bibliography

Agarwal, Arul. "Sentiment analysis of financial news." In *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 312-315. IEEE, 2020.

Asudani, Deepak Suresh, Naresh Kumar Nagwani, and Pradeep Singh. "Impact of word embedding models on text analytics in deep learning environment: a review." *Artificial Intelligence Review* 56, no. 9 (2023): 10345-10425.

Barbaglia, Luca, Sergio Consoli, and Sebastiano Manzan. "Forecasting with economic news." *Journal of Business & Economic Statistics* 41, no. 3 (2023): 708-719.

Bastos, M. T., Mercea, D., & Charpentier, A. "Tents tweets and events: The interplay between ongoing protests and social media." *J. Commun*., vol. 65, no. 2, pp. 320-350, 2015.

Bonta, Venkateswarlu, Nandhini Kumaresh, and Naulegari Janardhan. "A comprehensive study on lexicon based approaches for sentiment analysis." *Asian Journal of Computer Science and Technology 8, no. S2* (2019): 1-6.

Butcher, Charles, and Isak Svensson. "Manufacturing dissent: Modernization and the onset of major nonviolent resistance campaigns." *Journal of Conflict Resolution* 60, no. 2 (2016): 311-339.

Chadefaux, Thomas. "Early warning signals for war in the news." *Journal of Peace Research* 51, no. 1 (2014): 5-18.

Chauhan, V. K., Bansal, A., & Goel, A. (2018). Twitter sentiment analysis using vader. *International Journal of Advance Research, Ideas and Innovations in Technology (IJARIIT),* 4(1), 485-489.

Davies, James C. "Toward a theory of revolution." *American Sociological Review* (1962): 5-19.

Djaballah, Kamel Ahsene, Kamel Boukhalfa, and Omar Boussaid. "Sentiment analysis of Twitter messages using word2vec by weighted average." In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS),* pp. 223-228. IEEE, 2019.

Dogru, Hasibe Busra, Sahra Tilki, Akhtar Jamil, and Alaa Ali Hameed. "Deep learning-based classification of news texts using doc2vec model." In *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA*), pp. 91-96. IEEE, 2021.

Gomes, Helder, Miguel de Castro Neto, and Roberto Henriques. "Text Mining: Sentiment analysis on news classification." In *2013 8th Iberian Conference on Information Systems and Technologies (CISTI*), pp. 1-6. IEEE, 2013.

Jurafsky, Daniel, and James H. Martin. *Speech and Language Processing*. 3rd ed. Pearson, 2020.

Kim, Donghwa, Deokseong Seo, Suhyoun Cho, and Pilsung Kang. "Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec." *Information Sciences* 477 (2019): 15-29.

Lau, Jey Han, and Timothy Baldwin. "An empirical evaluation of doc2vec with practical insights into document embedding generation." *arXiv preprint arXiv*:1607.05368 (2016).

Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." *In International Conference on Machine Learning*, pp. 1188-1196. PMLR, 2014.

Liu, Bing. "Sentiment analysis and subjectivity." *Handbook of Natural Language Processing* 2, no. 2010 (2010): 627-666.

McAdam, Doug. \*Political Process and the Development of Black Insurgency, 1930-1970. *University of Chicago Press*, 1999.

McCarthy, John D., and Mayer N. Zald. "Resource mobilization and social movements: A partial theory." *American Journal of Sociology* 82, no. 6 (1977): 1212-1241.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space." *Proceedings of the International Conference on Learning Representations (ICLR*), 2013, 3111-3119.

Ning, Yue, Sathappan Muthiah, Huzefa Rangwala, and Naren Ramakrishnan. "Modeling precursors for event forecasting via nested multi-instance learning." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 1095-1104. 2016.

Ning, Yue, Sathappan Muthiah, Naren Ramakrishnan, Huzefa Rangwala, and David Mares. "When do Crowds turn Violent? Uncovering Triggers from Media." In \*2018 *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 77-82. IEEE, 2018.

Park, Eunhye, Jinah Park, and Mingming Hu. "Tourism demand forecasting with online news data mining." *Annals of Tourism Research* 90 (2021): 103273.

Seabold, Skipper, and Josef Perktold. "Statsmodels: econometric and statistical modeling with python." *SciPy* 7, no. 1 (2010).

Shapiro, Adam Hale, Moritz Sudhof, and Daniel J. Wilson. "Measuring news sentiment." *Journal of Econometrics* 228, no. 2 (2022): 221-243.

Sobolev, Anton, M. Keith Chen, Jungseock Joo, and Zachary C. Steinert-Threlkeld. "News and geolocated social media accurately measure protest size variation." *American Political Science Review* 114, no. 4 (2020): 1343-1351.

Steinert-Threlkeld, Z. C., D. Mocanu, A. Vespignani, and J. Fowler. "Online social networks and offline protest." *EPJ Data Sci.*, vol. 4, no. 1, pp. 19, 2015.

Thurber, Ches. "Ethnic barriers to civil resistance." *Journal of Global Security Studies* 3, no. 3 (2018): 255-270.

Truşcă, Maria Mihaela. "Efficiency of SVM classifier with Word2Vec and Doc2Vec models." In *Proceedings of the International Conference on Applied Statistics*, vol. 1, no. 1, pp. 496-503. 2019.

Urologin, Siddhaling. "Sentiment analysis, visualization and classification of summarized news articles: a novel approach." *International Journal of Advanced Computer Science and Applications* 9, no. 8 (2018).