Insights into a New Social Movement: Utilizing Machine Learning to Classify and Analyze #BlackLivesMatter Twitter Advocacy

Sinclaire Schuetze

Submitted in Partial Fulfillment
of the
Prerequisite for Honors
in Data Science
under the advisement of Eni Mustafaraj

April 2023

© 2023 Sinclaire Schuetze

Abstract

Black Lives Matter has been closely linked with Twitter throughout its entire history as a movement. The platform first offered a unique opportunity for BLM to grow because of its novel use of hashtags that enabled widespread conversations surrounding police brutality. The messages span from sharing objective information about the movement to promoting actions like protests and petitions. These forms of activism, who participates, and how #BlackLivesMatter is tied to other social movements is connected with its classification as a New Social Movement – which places emphasis on garnering public support to force change rather than working directly through political institutions. There have been prior analyses of these features of the #BlackLivesMatter movement; however, they have been primarily limited to qualitative and smaller-scale quantitative research. This work aims to study how more advanced data science tools, such as those provided by natural language processing and machine learning, can be applied to existing #BlackLivesMatter research. By using these types of methods, a larger dataset consisting of 21 million tweets can be studied that spans the entire length of the movement from July of 2013 through January of 2022. Traditional machine learning and large language models will are researched, where machine learning models are those that make predictions about data and large language models are a subset of more complex models specifically targeted towards understanding natural language. Both types of models are first assessed on their ability to classify tweets by differentiating between those that are relevant and irrelevant. They are further tested on how well they classify relevant tweets as the following calls to action: within the system, disruptive activism, and encouraging. These classifications of tweets are then used in understanding trends of the overall movement, those by Twitter users, and those related to other movements. The findings include that large language models are more successful in classifying tweets than traditional machine learning methods using metrics such as accuracy, precision, recall, and F1-score. Additionally, tweets overwhelmingly spread encouragement of the movement rather than within the system or disruptive actions; however, disruptive tweets have become more popular than within the system tweets over the course of the movement. Popularity does correspond with how users tweet, with popular individuals and organizations tweeting more about within the system and disruptive calls to action compared to all users. Lastly, the movement is intersectional, but this is mostly related to feminism rather than other social movements. Together, these results support #BlackLivesMatter's title as a New Social Movement and demonstrate the success in using machine learning to answer social science research questions.

Acknowledgements

I first want to thank Professor Eni Mustafaraj for her support not only as my thesis advisor this past year but also as my concentration advisor for the Data Science major. Your mentorship and advice during my time at Wellesley have been unbelievably instructive and invaluable.

I also want to thank my entire thesis committee: Professors Carolyn Anderson, Maneesh Arora, and Cassandra Pattanayak, for all of their guidance and feedback along the way.

Thank you to the Wellesley CAPS Lab members and research assistants who made getting labeled data possible. This thesis truly could not have been completed without your work. Our discussions about the lab's progress were so insightful and always gave me renewed motivation towards this project.

To my friends, thank you for cheering me on after every single page I wrote. And lastly, thank you to my family who has been my biggest support system during my entire college career.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The intense emotions surrounding the acquittal of George Zimmerman in the murder of Trayvon Martin and the surge in popularity of Twitter in 2013 coincided to create an environment suited to a modern wave of social movements. The formation of the #BlackLivesMatter movement found its homebase on Twitter due to the platform's invention of hashtags and Trending Topics that created potential for both national and global impact (Edrington and Lee, 2018). Because of this usage of Twitter, researchers have looked towards the platform for many of their studies regarding the movement. Their questions often ask why and how the platform is being used. The more obvious manifestations of the movement, such as protests, receive much news coverage and other media attention, but Twitter can offer a space for activism that is not as easy to see (Edrington and Lee, 2018). Analyzing the prevalence of all calls to action on Twitter is imperative to understanding the movement's current goals and its potential for tangible change.

This analysis, however, first requires determining the forms of activism contained in tweets. This is a complicated task because of the countless ways people express many forms of activism and the sometimes indirect nature of tweets' messaging. The task of processing millions of tweets can be accomplished through an ensemble of natural language processing tools, machine learning models, and other data science methods. Previous social science research has mostly used more traditional data analysis techniques, but these require

smaller subsets of data (Tillery, 2019). Working with that limited data therefore restricts conclusions to particular accounts or time periods in the dataset, whereas Big Data opens the door to broader conclusions about the overall #BlackLivesMatter movement. This work therefore first explores the research question of which natural language processing techniques and machine learning models best classify the forms of activism contained in #BlackLivesMatter tweets. The paper will then utilize the activism classifications from these models to explore social science questions regarding how different people relate to the movement and how this manifests itself on Twitter.

The following research questions guided this work:

Q1. Can advanced data science methods, such as natural language processing and machine learning, be applied to studies of #BlackLivesMatter? Specific to this work, can models successfully classify forms of activism contained in tweets?

Q2. How has activism shared within the movement changed over time?

Q3. How do forms of advocacy within the #BlackLivesMatter movement vary with the type of Twitter user? Specifically, do the accounts of popular individuals and organizations tweet differently about the movement compared to all users?

Q4. What is the #BlackLivesMatter movement's relationship to intersectionality? Is the movement intertwined with others and has that changed over time?

In this work, I introduce the #BlackLivesMatter and its classification as a New Social Movement in section 2. In section 3, I provide background about the traditional machine learning methods and large language models that will be used. Section 4 reviews previous works regarding #BlackLivesMatter's Twitter presence as well as tweet classification. Section 5 provides an overview of the data I will be using and how it will be labeled with the forms of advocacy. In section 6, I explain how the machine learning methods introduced will specifically be applied to this research. In section 7, I describe the results of using these methods on the dataset to answer research question 1. Section 8 outlines the methods for analyzing the data following classification. Section 9 details the results of these methods that will answer research questions 2, 3, and 4, and section 10 presents

the conclusions that can be drawn from these results as well as limitations and future research.

Chapter 2

Background

This chapter introduces the Black Lives Matter movement and how its use of Twitter has evolved over time. The chapter also defines New Social Movements and their characteristics, which are relevant to how Black Lives Matter has been shaped. This section concludes by outlining the motivation for this study.

2.1 The History of #BlackLivesMatter on Twitter

On February 26, 2012, a Black teenager from Florida named Trayvon Martin was walking home from a convenience store when he was fatally shot by a neighborhood watch volunteer, George Zimmerman. Police initially did not arrest Zimmerman based on the claim of self-defense, and this lack of action sparked national protests and uproar (Lebron, 2017). Zimmerman was later charged with second-degree murder, but he was eventually acquitted of those charges. In response to the acquittal, Alicia Garza started a series of social media posts collectively called "A Love Letter to Black People" where the final post stated:

> "stop saying we are not surprised. that's a damn shame in itself. I continue to be surprised at how little Black lives matter." (Garza, 2016)

Patrisse Cullors, a friend of Garza, had seen her post and in Cullor's own Facebook post wrote:

> "declaration: black bodies will no longer be sacrificed for the rest of the world's enlightenment. i am done. i am so done. trayvon, you are loved infinitely. #blacklivesmatter" (Garza, 2016)

These were the first appearances of the phrase "Black Lives Matter" on social media. Alicia Garza, Patrisse Cullors, and Opal Tometi soon after formed the Black Lives Matter (BLM) movement. They defined the movement as "an ideological and political intervention in a world where Black lives are systemically and intentionally targeted for demise" (Garza, 2016). In 2013, the three organizers began creating online spaces to further establish the movement, including social media pages on Facebook, Tumblr, and Twitter. Although the murder of Eric Garner in July of 2014 sparked national demonstrations, the hashtag did not gain much widespread online presence outside of the spaces created by Garza, Cullors, and Tometi (Lebron, 2017). It was not until the killing of Michael Brown in August of 2014 when the hashtag's use saw exponential growth.

The use of Twitter as a platform was pivotal in this surge of #BlackLivesMatter on social media since Twitter was the first space to incorporate hashtags and Trending Topics. With posts no longer limited to one's close network of friends and family, discourse about trials and protests could be seen locally, nationally, and globally (Edrington and Lee, 2018). This hashtag became the primary avenue for people and organizations to share photographs, live videos, and other details regarding the deaths of Black individuals and calls to action (Edrington and Lee, 2018). From 2016 through 2018, the hashtag saw consistent activity with an average of 15,856 tweets containing it per day (Anderson, 2018). The combination of this widespread online activity and many in-person demonstrations provided the foundation for #BlackLivesMatter to be taken seriously as a social movement.

The movement saw an unparalleled increase in engagement following the killing of George Floyd in May of 2020. In the week following his death, users generated 3.4 million original posts – equivalent to 13% of all Twitter posts, whereas the previous record was 145,631 posts following the deaths of Alton Sterling and Philandro Castile in 2016 (Wirtschafter, 2021). The activity following Floyd's murder has been regarded as a turning point in the movement. Prior to May of 2020, the popularity of #BlackLivesMatter

and related hashtags rose and fell in response to instances of police brutality. However, after May of 2020, this pattern seemed to no longer hold. Throughout 2021, Twitter saw on average 2,500 more posts per day compared to 2020, demonstrating that Black Lives Matter has seen a consistent increase in activity even during periods not directly following an instance of police brutality (Wirtschafter, 2021).

2.2 Overview of New Social Movements

Social movements are defined as "organized yet informal social entities that are engaged in extra-institutional conflict that is oriented towards a goal" with these objectives including both targeting specific policies or breaking down broader systemic issues (Della Porta and Diani, 2006). The New Social Movement (NSM) paradigm emphasizes the branch of social movements targeting common social issues and systemic change (Pichardo, 1997). Whereas traditional social movements of the industrial era were centered around working class individuals, the post industrial 1960's ushered in a shift towards protests of the new middle class, such as during the Women's Rights Movement and Occupy Wall Street. This new middle class included individuals who tended to have greater access to higher education and other resources (Pichardo, 1997).

NSMs' tactics include galvanizing public support and organizing demonstrations rather than working through traditional institutional or political channels (Pichardo, 1997). Many studies about New Social Movements look at the relationship between BLM and Twitter because of how the movement acts as a prime example of attempting to gain public support – the popularity of associated hashtags forces widespread conversations and people to take a stance (Tillery, 2019). This rejection of bureaucratic concepts common to NSMs also manifests itself in #BlackLivesMatter's structure. The movement has not identified a central leader and has instead allowed its supporters to determine how it will next take shape and evolve.

New Social Movements are also distinct in their focus on individuals' expression of identity and how that can become politicized. Most of the literature about social move-

ment theory regards Black Lives Matter as a New Social Movement because of its attention placed on identities, whether that be race, gender, or sexual orientation (Tillery, 2019). The Movement for Black Lives' policy platform directly states:

> "We are intentional about amplifying the particular experiences of racial, economic, and gender-based state and interpersonal violence that Black women, queer, trans, gender nonconforming, intersex, and disabled people face." [1]

Not only do the movement's goals display intersectionality but so does its non-bureaucratic structure. Rather than the male-centered leaders common during the Civil Rights Movement, Black Lives Matter takes a group-centered approach that is inclusive of and driven by all types of members within the movement (MacDonald and Dobrowolsky, 2020).

2.3 Motivation of Research

A significant motivator of this work is to understand how natural language processing (NLP) and machine learning (ML) tools can be applied to study the Black Lives Matter movement on Twitter. These tools will specifically attempt to classify activism promoted by tweets related to the movement and how these and other characteristics pertain to #BlackLivesMatter's classification as a New Social Movement. The exploration of how NLP, ML, and other data science tools can be wielded in the social sciences is imperative because of the potential for access to new areas of study. In this paper, a larger volume of tweets compared to related studies can be included that allow for broader analysis of the entire length of the movement. Additionally, more focus can be paid to specific lenses that are now captured by having access to a greater number of tweets.

[1] 'Policy Platforms', *Black Power Rising*, https://m4bl.org/policy-platforms, (accessed 10 March 2023)

Chapter 3

Overview of Classification Methods

To classify a large dataset like the #BlackLivesMatter tweets, machine learning methods can be utilized. Machine learning algorithms attempt to uncover patterns in data through numerical calculations and use these patterns to predict values for unseen data. When handling textual data such as tweets, additional techniques must convert the text into numbers that can be used within the machine learning algorithms. This chapter will cover these methods and how they interact with one another.

3.1 Machine Learning Models

Machine learning's goal is to create models that imitate the way humans make decisions. This is the technology behind everything from Netflix's movie recommendations to stock market predictions. Within machine learning there are two branches – supervised and unsupervised learning. Unsupervised models learn general patterns within the data without seeing any data prior (Mitchell et al., 2007). A common use case of these models is to cluster the data to understand which data points are most similar to each other. If there is a dataset of artists' songs and features like genre, tempo, length, and mood, an unsupervised model could group these songs to show which artists produce similar music.

Supervised learning uses labeled training data allowing for specific predictions to be made for unseen data (Mitchell et al., 2007). For example, an algorithm could be provided with a dataset containing the following features: home price, number of bedrooms, square footage, and zip code. If the model is trying to predict home price, the information on home price already given would be considered the labeled training data. The model will try to understand how the other features (number of bedrooms, square footage, and zip code) interact to produce home price, and this model can be applied to houses where the price is unknown.

3.1.1 Traditional Machine Learning Methods

There are many types of supervised models, but those relevant to this work are k-nearest-neighbors (kNN), perceptrons, support vector machines (SVMs), and neural networks. With a kNN model, the label of an instance represented as a datapoint in a multi-dimensional vector space is decided by obtaining the mean of the k-closest points (Raschka et al., 2022). Continuing with the home price example, if there are three homes in the same zip code that have similar square footage and number of bedrooms with prices of \$275k, \$300k, and \$310k, the predicted home price for a fourth house with those same characteristics would be \$295k.

Perceptrons instead try to identify a line between classes of data points by iterating through each point and updating numerical inputs known as weight and bias terms. An example of this decision boundary can be seen in figure 3.1 where a line has been found that divides homes that have and have not been sold. Depending on where a new point falls across that line will determine how it is classified. SVMs also identify a decision boundary, but they work by finding the line that maximizes the margin, which is calculated as the distance between the line and data points from all classes. Lastly, a neural network is a network of equations, such as perceptrons, where the output of each node is the input to the next.

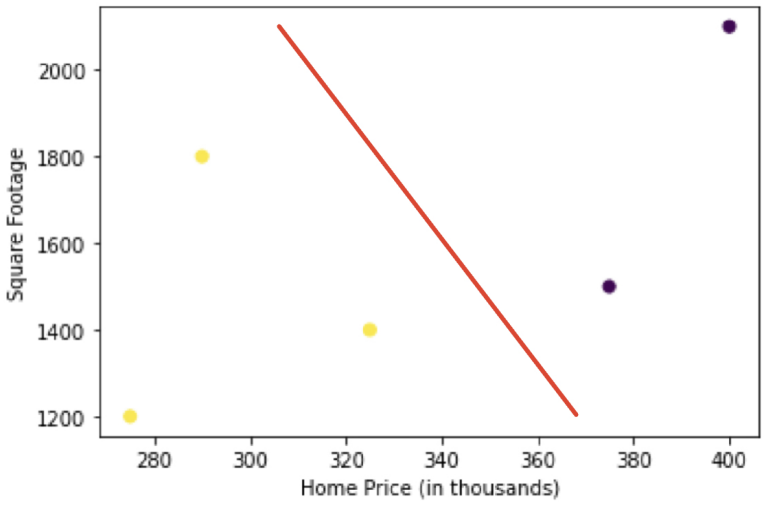


Figure 3.1: Example of decision boundary used in explanation of traditional machine learning methods. The yellow points are homes that have sold and the purple points are homes that have not sold.

3.2 Natural Language Processing

The machine learning models presented in the previous section all require numeric data; however, some data, such as tweets, is textual. Transforming text data into numeric data that can be used by these types of models requires the use of natural language processing – a collection of computational techniques used to represent human language. These include both pre-processing methods as well as text-specific machine learning models.

3.2.1 Pre-processing Techniques

One goal of NLP is to ensure that the meaning of words is maintained when converting text data to numeric data while also keeping efficiency in mind. Yet, to maintain this meaning, what words are contained in a document must first be understood. Tokenization is the process of splitting sentences into individual words since these units can more easily be assigned meaning. This process typically involves identifying where spaces and punctuation are and using these as demarcations of the beginnings and ends of words (Kannan et al., 2014).

Once words have been identified, additional techniques remove or change words to allow

the machine learning model to focus on more important information, therefore improving its efficiency. Stopword removal extracts words from the document that by themselves do not contain much meaning such as articles, pronouns, and prepositions. Stemming involves removing the ends of words to transform them into their base word. Lemmatizing is similar to stemming but instead transforms words to their root form. With stemming and lemmatizing, since these variations of words contain practically the same meaning, this aids in model efficiency by reducing the number of words a model needs to store (Kannan et al., 2014).

Method	Text
Original	I wrote a research paper for my senior year of college on African American history and the National Mall. It was a reminder that our country isn't truly free unless all its citizens are treated as equals #BlackHistoryMonth #BlackLivesMatter
Stopword Removal	I wrote research paper ~~for my~~ senior year ~~of~~ college on African American history ~~and the~~ National Mall. It ~~was a~~ reminder ~~that our~~ country ~~isn't~~ truly free unless ~~all its~~ citizens ~~are~~ treated ~~as~~ equals #BlackHistoryMonth #BlackLivesMatter
Stemming	I wrote a research paper for my senior year of colleg~~e~~ on African American histori and the Nation~~al~~ Mall. It wa~~s~~ a remind~~er~~ that our countri isn't truli free unless all it~~s~~ citizens are treat~~ed~~ as equal~~s~~ #BlackHistoryMonth #BlackLivesMatt~~er~~
Lemmatizing	I wr~~i~~te a research paper for my senior year of college on African American history and the National Mall. It be a reminder that our country be not truly free unless all its citizens be treat~~ed~~ as equals #BlackHistoryMonth #BlackLivesMatter

Table 3.1: Examples of natural language pre-processing techniques. The first row shows the original, example tweet. The following rows show deleted or changed words and letters in red or using strikethroughs. The second row demonstrates stopword removal, which removes common words. The third row demonstrates stemming which changes the end of words to their most common form. The fourth row demonstrates lemmatizing which changes words to their root form.

Usage of these techniques can be seen in table 3.1 using an example tweet. The stopword removal deletes words such as "of", "was", and "its". Stemming changes ends

of words like “history” to ”histori” and removes the -er from “reminder”. Lemmatizing changes verbs likes “was” to “be” and “wrote” to “write”.

3.2.2 Vectorization and Word Embeddings Techniques

Once pre-processing techniques are implemented, models then use vectorization or word embedding methods to transform text into vectors of numbers. Whereas vectorization is this general process of capturing word relationships through vectors, word embeddings are a type of vectorization that explicitly convert each word into its own vector that is supposed to capture the word’s meaning. Term Frequency-Inverse Document Frequency (TF-IDF) is a vectorization method that attempts to gauge how relevant a word is to a set of documents. The formula for calculating TF-IDF is:

$$TF \times log(N/df)$$

TF involves counting the number of times a word is contained in a document. This number is then divided by the total number of words in that document which together calculate the term frequency (Ramos et al., 2003). The log of the total number of documents divided by the number of documents containing the word calculates the inverse document frequency (IDF). The two components (TF and IDF) are then multiplied together. This overall weights words more heavily that appear many times in a limited number of documents. Words that are common in every document (such as the stopwords previously mentioned) will have lower weights to indicate that they are not as meaningful (Ramos et al., 2003).

Document Number	Tweet
1	Say her name #BreonnaTaylor
2	The story of #BreonnaTaylor breaks my heart
3	Say their name #BreonnaTaylor #AhmaudArbery
4	I run for #AhmaudArbery

Table 3.2: Examples of tweets used in TF-IDF explanation. These demonstrate how words like #BreonnaTaylor have lower TF-IDF values because they are present in all of the sentences in the document.

In the tweets in table 3.2, the term frequency (TF) of "say" in document 1 would be 0.25 since it is one out of four words present in the sentence. The inverse document frequency (IDF) would be 0.693 – calculated by taking the natural log of the number of documents divided by the number of documents which contain "say." Multiplied together, this results in a TF-IDF for "say" in document 1 of 0.173. #BreonnaTaylor in document 2 would have a TF of 0.14 since it is one out of seven words. Its IDF would be 0.288, obtained by taking ln(4/3), and multiplying these together its TF-IDF would be 0.04. This lower TF-IDF score indicates that #BreonnaTaylor holds less unique meaning since it is present in more documents. TF-IDF could be calculated for each word in a document and combined into a vector providing a numeric representation of that sentence.

Compared to TF-IDF, Word2Vec is a word embedding method that is better at capturing associations and dependencies between words using Continuous Bag of Words (CBOW) and skip-gram architectures, meaning that words like "queen" and "king" would be given similar values (Church, 2017). The first architecture, CBOW, predicts a target word from a group of context words. For example, with the tweet "I wrote a research paper for my senior year of college on African American history and the National Mall," if we want to predict senior, the context words could be "paper", "for", "my", and "year". The second architecture, Skip-gram, oppositely tries to predict the words before and after a given input word by predicting probabilities, so it would instead try to predict the words "my" and "year" if given "senior." Another method, GloVe, is similar to Word2Vec but also captures global document context rather than just local context by creating a co-occurrence matrix of all words in the document, which essentially maintains a record of which words appear together (Pennington et al., 2014).

Bidirectional Encoder Representations from Transformers (BERT) is a large language model that provides pre-trained language model word representations (Devlin et al., 2019). BERT is unique in its ability to understand the specific context within which a word is situated. BERT does this by attaching a unique word embedding dependent on the differing context. Yet, when trying to derive embeddings for entire sentences rather than

words, BERT's results lagged behind other embedding methods, like Word2Vec and GloVe, which led to the creation of Sentence-BERT (sBERT). sBERT is instead pre-trained on entire sentences ("Sentence-BERT", 2019).

3.2.3 Unsupervised Methods

The numeric vectors obtained from word and sentence embeddings can be utilized as inputs for machine learning models. Unsupervised learning relevant to this research includes two types of topic modeling - latent dirichlet allocation (LDA) and BERTopic. Both of these methods intend to discover common themes across documents. After converting the document to its vectors using a method such as TF-IDF, LDA topic modeling groups these documents based on statistically significant words (Blei et al., 2003). The method determines and outputs these statistically significant words by using Bayseian inferences to estimate the chances of the words occurring again.

BERTopic is very similar to LDA topic modeling in the goal of its model; however, its purpose is to create more understandable topics by understanding the context of words in a sentence. BERTopic first converts each document to its BERT embeddings, reduces the dimensionality of these embeddings, clusters these representations, and finally extracts topics using class-based TF-IDF (c-TF-IDF) ("BERTopic", 2022). c-TF-IDF's goal is to supply all documents within a class with the same class vector which it does by combining all documents from one topic into a singular document. c-TF-IDF then uses the number of classes instead of the number of documents in its calculations. This creates a matrix of similar words for each class and BERTopic extracts the highest probability words to create its topic models. The base output of BERTopic is similar to LDA topic modeling where it outputs the most relevant words per topic; however, there is a variation of BERTopic that conducts hierarchical topic modeling. This provides an understanding of which outputted topics are most similar and potential sub-topics.

3.2.4 Deep Learning and Large Language Models

Deep learning methods, relative to models such as those previously mentioned, are more complex neural networks that aim to truly replicate brain-like processes, and those specific to understanding human language are known as large language models (LLM). One example of a LLM is BERT, whose word embeddings were mentioned in section 3.2. BERT can be used as a model for a multitude of tasks, such as next sentence prediction or masked language modeling, where a hidden word in a sentence is predicted. A benefit of these large language models is that they can be fine-tuned to work on more specific tasks and datasets, which grants BERT the capability to work on a variety of tasks. Fine-tuning a model includes using the pretrained weights as starting points, which minimizes the amount of additional work that needs to be done. Another small neural network layer is added to tailor the model to the necessary task. The specific model that will be tested in this work is fine-tuning DistilBERT (Sanh et al., 2020). This model is a distilled version of the BERT model previously mentioned. Distilling includes compressing the model so that the smaller model (DistilBERT) more efficiently replicates the behavior of the larger model (BERT).

3.3 Metrics

Supervised classification models can be evaluated by comparing the output of the model to the true values of labeled data. The first step in evaluating classifiers is creating a confusion matrix. This is a matrix that compares numbers of actual and predicted values per class. Specifically, in binary classification, the values correctly predicted to be 1's are labeled as true positives, the values incorrectly predicted to be 1's are false positives, the values correctly identified as 0's are true negatives, and the values incorrectly predicted to be 0's are false negatives. For example, in the confusion matrix shown in table 3.3, there are 100 values total with 35 belonging to class 0 and 65 belonging to class 1. 30 of the class 0 values are true negatives with 5 being predicted as false positives, and 55 of the class 1 values were true positives with 10 being predicted as false negatives. Ideally, there

would be very low values for the cells where the truth value does not equal the predicted value.

		Predicted Value		
		0	1	Total
True Value	0	30	5	35
	1	10	55	65
	Total	40	60	100

Table 3.3: Example of confusion matrix with binary model. The matrix displays the true values versus the predicted values. These values (true positives, true negatives, false positives, and false negatives) can be used in the calculation of other metrics.

Additional metrics that can be used include accuracy, precision, recall, and F1-score, whose formulas can be found in table 3.4. Accuracy is the number of correctly identified points divided by the total number of points. In the example above, the accuracy would be 0.85 since 85 of the values were correctly predicted. Accuracy works well when the dataset is balanced – meaning there are equal numbers of points in all classes. In this example, there are far more 1 values, so if the 1's are predicted correctly at a higher rate or the 0's are predicted at an especially low rate, the accuracy score will be artificially high because it is not demonstrating what it has truly learned about 0's (Zhou et al., 2021).

Therefore, having class-specific metrics is especially important. Precision is similar to accuracy but is calculated for each class by taking the total number of true positives divided by the total of true positives and false positives. From above, the precision for class 1 would be 0.92 by taking 55 divided by 60. This demonstrates what percentage of all positives predicted are truly positive. Another metric that accounts for class imbalance is recall. Instead of dividing by the total number of points predicted to be in class 1, it divides by the number of points actually in that class. Rather than dividing by 60, it would be divided by 65, resulting in a recall of 0.84. Lastly, F1-score combines precision and recall by taking their product divided by their sum and multiplying by two. The F1-score for this example would be

$$2 \times \frac{0.7728}{1.76} = 0.8792$$

For multi-class classification, these metrics can be calculated for each class or they can be aggregated via either micro or macro-averaging. Macro-averaging involves taking the mean of the metric calculated per class, whereas micro-averaging averages the sum of all required true positive, false positive, or false negative values, depending on the metric, for all classes.

Metric	Formula
Accuracy	$\frac{true positives + true negatives}{true positives + false positives + true negatives + false negatives}$
Precision	$\frac{true positives}{true positives + false positives}$
Recall	$\frac{true positives}{true positives + false negatives}$
F1-Score	$2 \times \frac{precision * recall}{precision + recall}$
Micro-averaging (Using precision with 3 classes)	$\frac{TP1 + TP2 + TP3}{TP1 + FP1 + TP2 + FP2 + TP3 + FP3}$
Macro-averaging (Using precision with 3 classes)	$\frac{Precision1 + Precision2 + Precision3}{3}$

Table 3.4: Equations used to calculate metrics. These metrics can be used to determine how well machine learning models are classifying data. Micro- and macro-averaging can be used with accuracy, precision, recall, and F1-score, but the table shows an example when used with precision.

3.4 Conclusion

An understanding of both traditional machine learning models and large language models and how they interact with pre-processing techniques is imperative as much of the previous literature regarding the use of data science methods in the social sciences utilize these same methods. The work in this research will also use these same methods as starting points for creating a classifier unique to the set of #BlackLivesMatter tweets.

Chapter 4

Related Works

This chapter will provide an overview of machine learning research relevant to creating natural language models using Twitter data – especially when concerned with social movements. Additionally, it will explore how Twitter data related to the #BlackLivesMatter movement has been analyzed previously with and without machine learning tools.

4.1 Machine Learning Models and Social Movements

4.1.1 Unsupervised Text Classification

Topic modeling has been utilized in the cases of other New Social Movements with a large presence on Twitter. Lee and Jang explore the popular topics and patterns of tweets within the #StopAsianHate (SAH) movement following the March 2021 shooting in Atlanta (Lee and Jang, 2021). This study used held-out likelihood to determine the appropriate number of topics and then used Structural Topic Modeling (STM) to obtain the topics themselves which consisted of a list of ten words. Lee and Jang then combined words within these lists to determine the meanings behind these topics. They determined the most prominent themes included recognizing the increase in violence against Asian-Americans and the need to build a sense of community during this time (Lee and Jang, 2021).

Tong similarly worked with topic modeling in the context of tweets regarding the SAH

movement as well as the BLM movement but instead used Latent Dirichlet Allocation (LDA) for topic modeling (Tong et al., 2022). A coherence score evaluated the LDA model by classifying the themes as interpretable or not. Tong tried varied numbers of topics ranging from 5 to 40 by intervals of 5. For #BlackLivesMatter, the coherence score determined the best number of topics was 15. Five tweets from each theme were sampled and human coders read through them to gain an understanding of the theme. Dominant topics included movement slogans, key characters (protestors, celebrities, politicians, etc.), and keywords clearly indicating emotional responses (Tong et al., 2022).

4.1.2 Supervised Text Classification

There has been much research regarding the use of natural language processing and machine learning in the analysis of social movements' use of social media. A study by Mishra, et al. created a sentiment classifier of tweets related to the following social causes: cyberbullying, concussions in the NFL, and LGBTQ+ issues (Mishra et al., 2014). The authors collected 1,500 tweets related to these topics and utilized four labels: enthusiastic/supportive, enthusiastic/non-supportive, passive/non-supportive, passive/supportive. They used a Linear Support Vector Machine with the following features: num. of emoticons, num. of URLS, num. of mentions, num. of hashtags, word features, num. of double quotes, and length of tweets. This classifier achieved a 79% accuracy for classifying enthusiastic vs. passive tweets and 77% accuracy for classifying supportive vs. non-supportive tweets (Mishra et al., 2014). Their research provides features that could also be beneficial in classifying forms of activism in this work.

A study by Qi, et al. attempts to find a model that best classifies tweets by state-level policy makers by their policy agenda topic (Qi et al., 2017). They investigated support vector machines, convolutional neural networks, and long short-term memory networks. For pre-processing, they looked at both removing and not removing stop words and used tf-idf vectorization on word and character-level n-grams. They utilized a new data augmentation method using word embedding which works by making k copies of a

tweet and replacing a selected word in the tweet with the k most similar words. This improved results by about 2% because it increased the dataset size and accounted for imbalances in the data (Qi et al. 2017). The convolutional neural network provided the best results with an F1 score of 78.3% (Qi et al., 2017).

Similar to the methodology in this work, Zeyad El-Zanaty focuses on hate speech on social media by classifying tweets as offensive or not offensive and then further classifying the offensive tweets into more specific categories (El-Zanaty, 2019). These categories include targeted insult: a post containing an insult or a threat to an individual, group, or others and untargeted: a post containing non-targeted profanity and swearing. The study also works to determine the target of these targeted insults (individuals, groups, and others). For pre-processing, El-Zanaty used each of the combinations of stop word removal, lemmatization, and stemming. The study then tried three different vectorization techniques: word2vec, fastText, and the pre-trained gloVe model. He then used GridSearchCV to determine the hyperparameters to be contained in the model. The models tested include KNN, Naive Bayes, SVM, Decision Trees, Random Forest, Logistic Regression, and MLP. The best F1-score of 0.683 and an accuracy of 0.85 came from removal of stop words followed by lemmatization with the use of Naive Bayes (El-Zanaty, 2019).

The focus of the study by Srivatsa was zero-shot characterization of tweets using an approach named "Designed Contextual Questions and Templates" to understand the presence of advocacy, disinformation, and propaganda (Srivatsa et al., 2022). Srivatsa constructed prompting questions that humans would typically use to understand a tweet and fine-tuned a pretrained GPT-2 model using Hugging-Face implementation. They conducted experiments using prompts in Yes/No, multiple choice question, general commonsense reasoning, and larger-context commonsense reasoning formats. This fourth model included providing a concept, its definition, and a few examples. Even with this given information, the model failed to provide any meaningful classifications, which indicates that task-specific supervised training may be required (Srivatsa et al., 2022).

4.2 Studies of #BlackLivesMatter

4.2.1 General Data Analysis

A study by Freelon, et al. conducted a network analysis of 40.8 million tweets to determine what types of people are tweeting the most about #BlackLivesMatter and what the tweets are primarily about (Freelon et al., 2016). The authors also conducted interviews with users to obtain more anecdotal information about their findings. Overall, they found that Black youth discuss police brutality in a very different way compared to activists.

Edrington looked at the official Black Lives Matter twitter account (@Blklivesmatter) to determine what message strategies BLM activists use to build a relationship with the public (Edrington, 2022). Edrington examined 450 tweets from May 2018 through May 2019 and specifically coded tweets by how they aligned with Burke's three identification strategies: sympathy, anthesis, and unawareness. The study found that 33% used sympathy strategies, 16% used antithesis strategies, and 27.4% used unawareness strategies while 23.6% did not fit into any category (Edrington, 2022).

The paper that has directly influenced this thesis is a study by Tillery in which the tweets of six social movement organizations (SMOs) were analyzed to understand the role Twitter plays in contributing to and advancing the conversations related to BLM (Tillery, 2019). Human coders manually labeled all 18,078 tweets to determine whether these tweets were supposed to mobilize, inform, or express emotions about the recent events relating to the movement. Tillery found that most tweets fell into the informative and emotional categories. These tweets were also then used to understand the various frames contained within the movement - how BLM relates to gender, sexuality, race, and others. Tillery found that the most popular frame focused on individual rights rather than intersectionality with gender, LGBTQ+, or other issues. The last research question focuses on calls to action and the proportion of tweets that encourage disruptive or violent action. Overall, the tweets largely encouraged systemic change rather than individual disruptive actions, and no tweets ever called for violent action (Tillery, 2019).

Another study by Bonilla and Tilery focuses even more on this idea of frames con-

tained in #BlackLivesMatter tweets (Bonilla and Tillery, 2020). They used a 2019 survey experiment to study people's relationship to the movement when they belong to communities such as LGBTQ+, Black Nationalists, and Feminists. The authors found that these frames were not salient to subjects, but they also did not depress mobilization.

4.2.2 Studies Containing Machine Learning and Natural Language Processing

As the popularity of natural language processing tools has grown over the past decade, more of these models have started to be applied to social science questions like those regarding #BlackLivesMatter. One such study includes an analysis by Saad Badaoui which utilized sentiment analysis to understand the feelings regarding #BlackLivesMatter (Badaoui, 2020). He looked at 600,000 tweets and first conducted polarity analysis to determine the amount of positive, negative, and neutral sentiment observed in these tweets. He found that over time, the proportion of these emotions was relatively evenly distributed between the three categories (Badaoui, 2020). He also analyzed the subjectivity of tweets and found that most of the tweets are objective.

Another sentiment analysis study is that by Ankita, et al. which proposes a Convolutional Neural Network-Long Short Term Memory (CNN-LSTM) for sentiment analysis of #BlackLivesMatter tweets (Ankita et al., 2022). The tweets, coming from both Minnesota and Washington D.C., were converted to lowercase, lemmatized, and had unnecessary keywords removed. Sci-kit learn's count vectorizer was then used to tokenize the tweets. The chosen sentiments were trust, surprise, anticipation, anger, fear, and disgust with the first three belonging to the non-hateful category and the last three belonging to the hateful category. The model was trained on 18 epochs using softmax pooling and a ReLU activation function, and its performance was compared to the following models: Random Forest, CNN, LSTM, BiLSTM, BERT base, and BERT large. The proposed CNN-LSTM model achieved a 94% accuracy on tweets and had a recall ranging from 36.8 to 74.1 for each of the sentiment categories. The study found that people largely tweeted positive content

to encourage one another with 48% tweets from D.C. and 54% of tweets from Minnesota expressing trust (Ankita et al., 2022).

4.3 Conclusion

The studies utilizing machine learning and natural language processing provide a foundation for which the methods of this paper can be established on. Additionally, the background surrounding the Black Lives Matter movement provided by these works guide the research questions and avenues of data analysis once the machine learning methods have been implemented. The next section will begin to describe how these techniques will be utilized in the context of understanding calls to action within #BlackLivesMatter.

Chapter 5

Data

This chapter will introduce the dataset that will be used throughout this work. There is a raw dataset containing the entire set of collected #BlackLivesMatter tweets as well as a data sample that will be used with the machine learning methods. Additionally, the call to action labels will be described.

5.1 Raw Data

21,415,397 tweets have been collected by Wellesley College's CAPS Lab. These tweets were obtained using a set of 39 hashtags regarding the #BlackLivesMatter movement such as #sayhername, #ferguson, and #abolishthepolice and are from almost the entire length of the movement – from 07-12-2013 through 01-06-2022. These tweets are all unique, meaning there are no retweets or quote tweets included. As seen in figure 5.1, the tweets obtained in this dataset follow exactly with the trends shown by other studies and their analysis of the number of tweets about the movement per year. The #BlackLivesMatter movement had a slow start in 2013 with only 28,400 tweets, but grew exponentially in 2014 with a total of 278,290 tweets. The number of tweets declines between 2017 through 2019 but again drastically increases in 2020 to a total of 1,026,260 tweets. The number of tweets again drops suddenly in 2021.

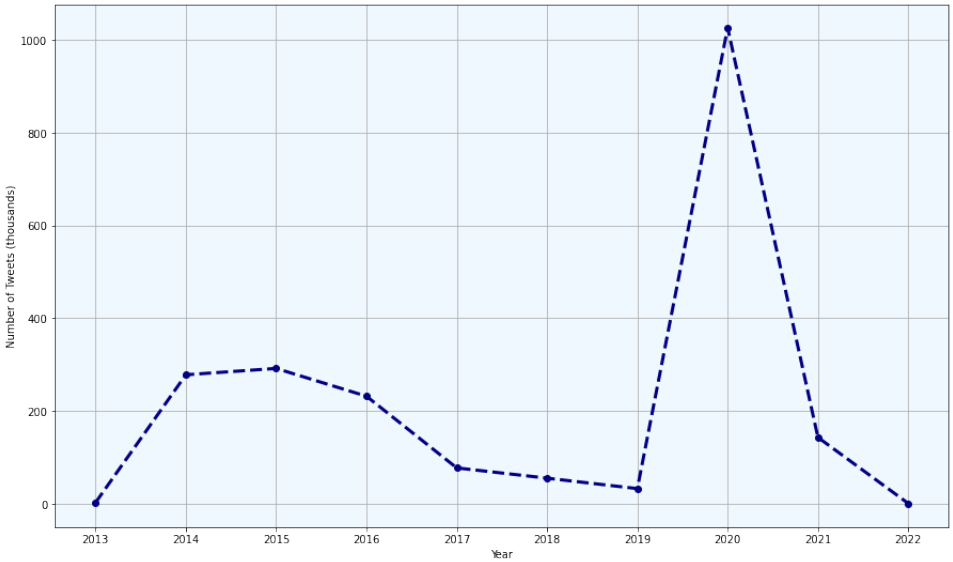


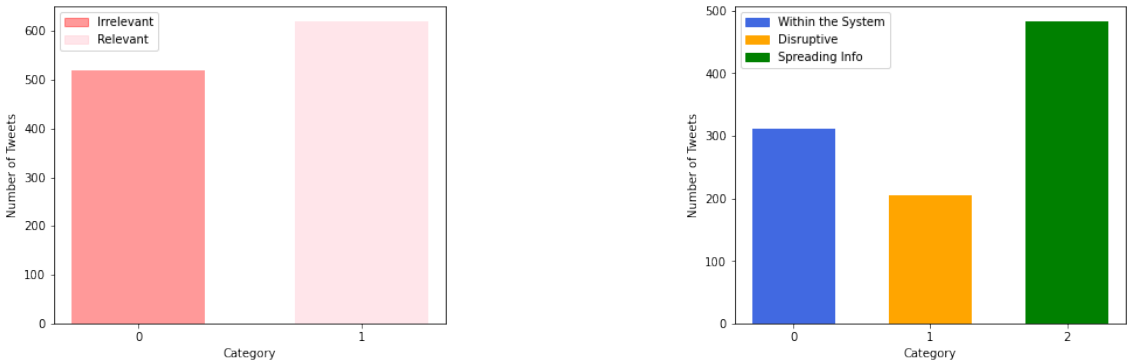
Figure 5.1: Number of unique tweets per year from 2013 through 2022. The sharp increase in number of tweets occurred in 2020 during which George Floyd was killed and a subsequent resurgence of the movement occurred.

5.2 Data Sampling

To train and test a supervised machine learning model, a sample of the data is needed. The sample is obtained using k-means clustering – a method that groups similar data points together into a given number of clusters. Because of limits in local storage, only 100,000 tweets were clustered and these were obtained from the dates 01-21-2020 through 05-13-2020. 3,000 clusters were used and one tweet was randomly selected from each of these clusters resulting in an initial sample of 3,000 tweets. Three Wellesley students were hired to label the tweets with two labeling independently and the third acting as the tie breaker. In the end, 1,140 tweets were labeled creating the final sample that is used in training and testing the machine learning models. The features included in this dataset are id, time tweeted, user, text, hashtags, number of retweets, number of favorites, number of replies, number of quotes, and advocacy label.

5.3 Data Labels

When the CAPS lab obtained the dataset of tweets, student researchers classified certain subsets of the tweets (such as those relating specifically to the events of Ferguson or Breonna Taylor) with one of nine possible calls to action contained in the body of the tweet: unrelated, within the system, disruptive, spreading information, encouraging, other, oppositional, and pressuring non-political elites. The tweets for this work were initially going to use these same labels used by the CAPS lab, but preliminary accuracy and F1-score results from classification models showed that there was not enough labeled data for each class to allow for successful classification using that number of labels. A new classification scheme was introduced where the tweets in the sample are first labeled as relevant or irrelevant to #BlackLivesMatter advocacy. This distinction is necessary because there are many tweets that utilize the hashtag #BlackLivesMatter or others related to the movement due to their popularity but do not actually contain content about the movement. When the tweets were labeled, the breakdown between the two groups was close to half and half: with 45% of the sample being irrelevant and 54% being relevant (figure 5.2a).



(a) Relevancy.

(b) Specific Calls to Action.

Figure 5.2: Numbers of tweets labeled per category by CAPS lab members. (a) displays irrelevant vs. relevant labels and (b) displays within the system vs. disruptive vs. encouraging of tweets initially labeled as relevant.

The tweets that are labeled as relevant are then further labeled by their forms of advocacy – within the system, disruptive, and spreading awareness. The labels categorizing

the tweets can be seen in table 5.1, and examples of these labels applied to tweets can be seen in table 5.2. There is a slight class imbalance present with these new labels – 31% of the tweets were labeled as 0 (within the system), 21% were labeled as 1 (disruptive), and 48% of the tweets were labeled as 2 (encouragement) as seen in figure 5.2b; however, the labels are far more evenly distributed compared to using all 9 labels of the original coding scheme.

Form of Advocacy	Label
Within the system forms of action (ex: voting, community gatherings, purchasing from Black businesses)	0
Disruptive (ex: protesting, boycotting, pressuring elites or companies)	1
Encouragement and spreading information	2

Table 5.1: Descriptions of three forms of calls to action with which tweets will be classified as. Examples of and the number associated with each are also provided.

Tables 5.3a and 5.3b show the percent of inter-rater agreement for both the binary and multi-class models. This inter-rater agreement was calculated by obtaining the percent of tweets where all labelers agreed on the final label. The binary labels were largely agreed upon, with both having an inter-rater agreement above 0.80 leading to an overall agreement on 85% of the tweets. The multi-class labels were more uncertain with both labels 0 and 1 having an inter-rater agreement less than 0.60 leading to an overall inter-rater agreement on only 74% of the tweets.

5.4 Conclusion

The creation of this sample is imperative to the remainder of the work because how accurate the final labels are depends on the information provided to the machine learning models. The process of using this sample in the models is detailed in the next chapter.

Example of Tweet	Call to Action
hi! i'd like to open up 6 bust commission slots for whoever donates $20+ to the BLM movement or similar supporting funds (NAACP legal defense fund, etc) i'll also do lineart busts for $10+ below are examples of what you'd be getting, DM me for details. #BlackLivesMatter https://t.co/kXvGdaLQiI	0 (Within the System)
we have the right to rebellion. we have the right to revolution. #BlackLivesMatter	1 (Disruptive)
We're tired. Tired of making hashtags. Tired of trying to convince you that our #BlackLivesMatter too. Tired of dying. Tired. Tired. Tired. So very tired #divinesenaya #Soûlkreôl #special #original #unique #music #love #humanity	2 (Encouragement)

Table 5.2: Examples of tweets that correspond to the three forms of advocacy. The first demonstrates donating to a fund benefiting BLM, the second demonstrates a more disruptive tweet, and the third expresses general sentiment regarding the movement.

Label	Inter-Rater Agreement
0	0.88
1	0.84
Overall	0.85

(a) Relevancy Agreement

Label	Inter-Rater Agreement
0	0.55
1	0.59
2	0.80
Overall	0.74

(b) Call to Action Agreement

Table 5.3: Inter-rater agreement scores from labeling the sample of tweets. (a) displays the proportion of tweets for which the labelers agreed that the tweet was relevant or irrelevant. (b) displays the proportion of tweets for which the labelers agreed that the tweet was talking about within the system, disruptive, or encouraging calls to action. The tables display this information for each individual class as well as for all the tweets combined.

Chapter 6

Classification Methods

This chapter will provide an overview of how the models outlined in the chapter 3 will be implemented – including both unsupervised and supervised models. It will also explain the process for determining which preprocessing techniques and vectorization methods to use and how the models will be evaluated to determine which will be applied to the entire #BlackLivesMatter dataset.

6.1 Preprocessing Techniques

Tweets contain many features that are not necessary when constructing a natural language model. To deal with these, the sample of tweets will be cleaned first by replacing urls with the word "URL" since the text of the link does not contain any context beneficial to a language model. The tweets will then be lowercased since there are many iterations of how common phrases used in the tweets (such as #BLM vs #blm) could be capitalized, and lowercasing the tweets allows for less unnecessary information. The remaining pre-processing techniques will only be used when testing supervised methods and involve removing stop words, stemming, and lemmatizing. Similar to the study by El-Zanaty, each of the models will be tested with all combinations of those three techniques as well as with none of the pre-processing techniques (El-Zanaty, 2019). These possible combinations are shown in table 6.1.

Combination	Stopword Removal	Stemming	Lemmatizing
1			
2	X		
3	X	X	
4	X	X	X
5		X	
6		X	X
7			X
8	X		X

Table 6.1: All combinations of pre-processing techniques that will be implemented, including stopword removal, stemming, and lemmatizing.

6.2 Vector Representation

To vectorize the tweets, two methods will be utilized for the supervised models – TF-IDF and Sentence-BERT. Since tweets are so limited in the amount of information they can contain due to their short length, these two embedding methods could possibly show different results based upon the amount of context that they can obtain. Both of the vectorizations will be tried with each combination of pre-processing techniques.

6.3 Unsupervised Learning Methods

Topic modeling methods will be used to determine if the forms of machine learning models can detect forms of advocacy without explicit instructions to do so. The two forms of topic modeling to be examined are LDA topic modeling and BERTopic. The number of topics used for LDA topic modeling will be determined based on the top words that are listed for each topic. The number of topics will be incremented by 5 each time, and once the topics

begin showing keywords indicating the three forms of advocacy the number of topics will stop being incremented. When the number of topics are decided upon, the LDA topic model will be implemented using TF-IDF vectorization. The model will output the top-30 most salient words for each topic. The value of using LDA topic modeling as a method of organizing tweets will be determined based upon how many of these advocacy specific words are present in each topic.

BERTweet has a hierarchical clustering variation that presents the relationships between the found topics. The number of topics will be automatically chosen by the model, and the model will use BERT sentence embeddings. The practicality of using this method will be determined by looking at how well the individual topics and relationships between the topics align with the predetermined classifications used for the supervised learning methods.

If topic modeling is able to group tweets by advocacy automatically, that would be invaluable to this type of work since it would involve less time taken to label, train, and test models.

6.4 Supervised Learning Methods

As mentioned in chapter 5, two forms of supervised models will be tested – a binary classifier and a multi-class classifier. The binary classifier will be used to predict relevant and irrelevant tweets. The multi-class classifier will be used to predict tweets which are within the system, disruptive, or about spreading information concerning the #BlackLivesMatter movement. All of the following mentioned methods will be applied to each model.

Four Sci-kit Learn classification models will also be used, including k-nearest-neighbors, perceptrons, SVMs, and neural networks. With these, the data will be split into 70% training, 15% testing, and 15% validation. The hyperparameters of each model will be tuned using the validation data and the grid search method. These include the number of neighbors and metric in the kNN model and the number of hidden layers in the neural network. All of these models will be tested using each of the pre-processing and embedding

combinations, and the mean 5-fold cross validation score will be reported. The final supervised model to be tested is fine-tuning the Distilbert-Base-Uncased Hugging Face Transformer. This model uses the provided distilbert tokenizer, and the number of epochs will be chosen based upon the point at which overfitting is observed – when the training and validation accuracies appear to significantly diverge.

The models will be evaluated on their overall accuracy, precision, recall, and F1-scores as well as accuracies broken down by class. Emphasis will be placed on F1-score especially for the multi-class classifier due to its better representation of precision and recall per class when there is class imbalance present. Consideration will also be given to the overall inter-rater agreement scores as well as the per-class agreement scores. The chosen model will ideally perform at least as well as human labelers. The best performing binary model will first be chosen and applied to the entire dataset of 21 million tweets. The multi-class classifier will then be utilized on the tweets tagged as relevant by the binary model.

6.5 Conclusion

The choice of how to implement and evaluate these models is pivotal to the analysis contained in the later portions of this research. The necessary pre-processing techniques, embedding methods, and hyperparameters need to be tried and tested to ensure the best possible results can be obtained. The accuracy of the labels and subsequent analysis and conclusions about the movement that are derived from the chosen models is dependent on the effort put into ensuring that sound methods are used.

Chapter 7

Classification Results

The following chapter aims to answer the first research question about what models best classify activism by describing the results of unsupervised and supervised machine learning models. The goal of this section is to determine which methods and models will be applied to the entire dataset of 21 million tweets.

7.1 Unsupervised Learning Results

LDA topic modeling results in a list of the top 30 most relevant terms for each topic. The data used was the entire dataset of tweets, and 20 topics were chosen based on exploratory analysis that showed this number allowed for the most distinct groupings relating to the desired labels. The top five words for each topic are shown in Appendix table A.1. Based upon these and other words provided, some general themes could be concluded for each topic. The topics include words relevant especially to the within the system and disruptive forms of action. For within the system forms of advocacy, associated topics include topics 2, 5, and 14 which contain the words vigil, vote, and voting. For disruptive forms of protest, associated topics include topics 4, 5, 9, 12, 15, and 16 which contain the words petebuttigieg, boycott, protestors, protests, and protest. The remaining topics included words that could be interpreted as encouraging or spreading awareness about the movement.

BERTopic provides the top three or four words associated with each topic. 16 topics were used based on the recommended number provided by BERTopic. Figure 7.1 shows the hierarchical clustering of these topics. Overall, many of the individual topics correspond to the three forms of activism. Topic 9 includes protest and protests and topic 11 includes boycott and money which are both relevant to disruptive activism. Topic 7 includes petition and sign, topic 6 includes votes, and topic 0 includes vigil which all point towards within the system calls to action. Encouraging tweets are harder to pinpoint because they do not call out specific actions, so any of the topics could theoretically be interpreted as encouragement of the movement. Additionally, the hierarchical clustering does not add much context – the topics are not grouped in any clear manner.

Although the results of unsupervised learning do show interesting trends in the tweets, it does not quite attack the task at hand of clearly grouping tweets by the calls to action contained. Because of this, supervised learning is most likely required for this kind of task.

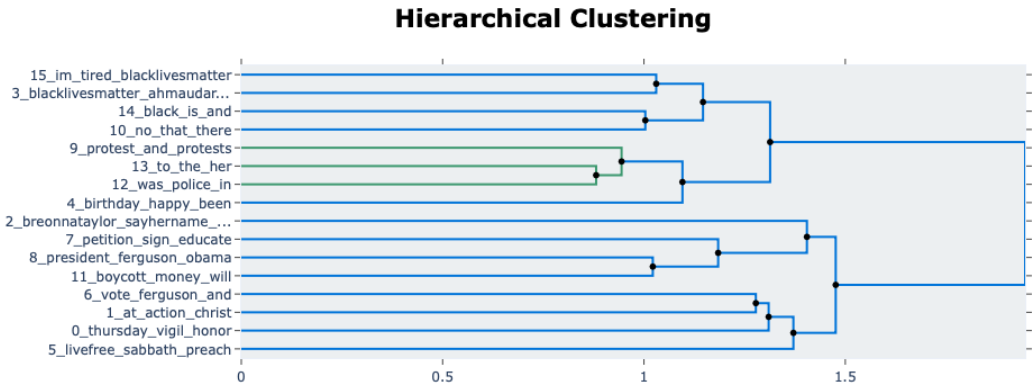


Figure 7.1: Hierarchical clustering of topics provided by BERTopic.

7.2 Supervised Learning Methods

The primary metrics used to evaluate the success of each of the models are weighted F1-scores, precision, and recall because of the moderate class imbalance; however, accuracy

was also used as a general baseline. Additionally, the percent of inter-rater agreement was calculated for each label, and this was used to determine the minimum accuracy per class that would ideally be reached. For the binary model, the overall inter-rater agreement is 0.85 with agreement for class 0 being 0.88 and agreement for class 1 being 0.84. For the multi-class model, the overall inter-rater agreement is 0.74 with agreement for class 0 being 0.55, agreement for class 1 being 0.59, and agreement for class 2 being 0.80.

When all possible combinations of pre-processing steps with each of the models are run for the multi-class task with TF-IDF vectorization (table 7.1), the highest macro-F1 score is associated with the use of none of the pre-processing techniques. The average of these scores was 0.69. All of the variations provide similar results, but the use of stopword removal, stemming, and lemmatization performed the worst with an average F1-score of 0.59. None of these pre-processing steps were therefore used in any of the models.

	kNN	Perceptron	Neural Network	SVM
No Cleaning	0.65	0.70	0.70	0.70
Stopword Removal	0.59	0.55	0.59	0.62
Stemming	0.57	0.61	0.58	0.61
Lemmatization	0.59	0.54	0.59	0.61
Stopword Removal and Lemmatization	0.59	0.55	0.59	0.62
Stopword Removal and Stemming	0.59	0.57	0.58	0.61
Lemmatization and Stemming	0.58	0.55	0.60	0.61
Stopword Removal, Lemmatization, and Stemming	0.59	0.57	0.58	0.61

Table 7.1: Accuracy scores for each of the combinations of pre-processing techniques when applied to multi-class classification using TF-IDF vectorization.

7.2.1 Binary Classifier

Traditional Machine Learning Models

The weighted accuracy, precision, recall, and F1-scores for the binary classification when using no pre-processing techniques and either sBert embeddings or TF-IDF vectors can be seen in tables 7.2 and 7.3. The best model with the use of sBert embeddings is the SVM, with an accuracy, precision, and recall of 0.80 and an F1-score of 0.79. Yet, all of the models produced similar results with scores between 0.73 - 0.80. The best model with the use of TF-IDF vectorization is the neural network with accuracy, recall, and F1-scores of 0.77 and a precision of 0.78. The results were more varied compared to those from the sBert embeddings with the kNN producing lower scores ranging from 0.63 - 0.67 and the SVM having an F1-score of 0.62. Additionally, the accuracies by class for the SVM model using sBert embeddings can be determined using a confusion matrix (figure 7.2). Class 0 (irrelevant) has an accuracy of 0.61, and class 1 (relevant) has an accuracy of 0.91 – both below the inter-rater agreement scores.

sBert	Accuracy	Precision	Recall	F1-Score
kNN	0.74	0.73	0.74	0.73
Perceptron	0.78	0.79	0.78	0.75
SVM	0.80	0.80	0.80	0.79
Neural Network	0.77	0.78	0.77	0.77

Table 7.2: Metric scores for binary classification when using sklearn methods with sBERT sentence embeddings.

Large Language Model

The results of fine-tuning the distilbert model can be found in table 7.4. The overall accuracy of the Distilbert model for the multi-class labels is 0.87, which is 0.02 percentage points higher than the inter-rater label agreement. The model also produced a precision,

TF-IDF	Accuracy	Precision	Recall	F1-Score
kNN	0.67	0.66	0.67	0.63
Perceptron	0.72	0.77	0.72	0.70
SVM	0.75	0.77	0.75	0.62
Neural Network	0.77	0.78	0.77	0.76

Table 7.3: Metric scores for binary classification when using sklearn methods with TF-IDF vectorization.

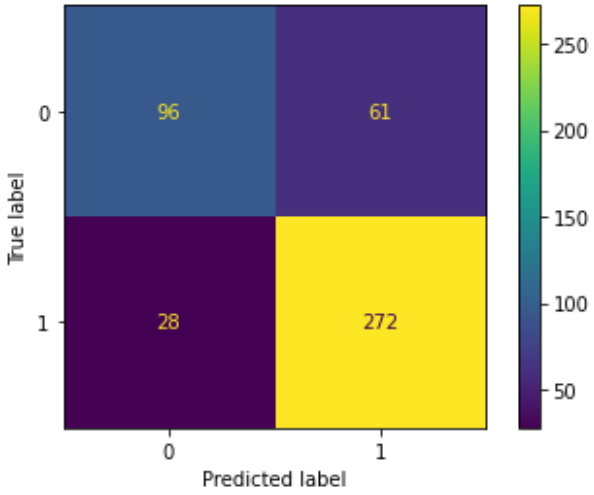


Figure 7.2: Confusion matrix displaying the true and predicted values from the binary SVM.

and F1-score of 0.87 and a recall of 0.86. By class, the model results in a 0.08 percentage point lower accuracy than the inter-rater agreement score for category 0, but it results in a 0.08 percentage point higher accuracy for category 1.

7.2.2 Multi-Class Classifier

Traditional Machine Learning Models

The weighted accuracy, precision, recall, and F1-scores for the multi-class classification when using no pre-processing techniques and either sBert embeddings or TF-IDF vectors

Metric	Score
Accuracy	0.87
Precision	0.87
Recall	0.86
F1-Score	0.87

(a) Metrics

Label	Inter-Rater Agreement	Testing Accuracy
0 (Irrelevant)	0.88	0.80
1 (Relevant)	0.84	092
Overall	0.85	0.87

(b) Inter-rater Agreement

Table 7.4: Performance of fine-tuned Distilbert Model for binary classification. (a) displays the accuracy, precision, recall, and F1-scores. (b) displays the inter-rate agreement scores and the by-class accuracy scores.

can be seen in tables 7.5 and 7.6, respectively. The best model with the use of sBert embeddings is the neural network, with an accuracy, precision, and recall of 0.76 and an F1-score of 0.75. All except the kNN model produced similar results with scores around 0.75. The best model with the use of TF-IDF vectorization was the perceptron, with an accuracy of 0.75, precision of 0.77, recall of 0.75, and F1-Score of 0.72. The neural network again produced very similar results with an F1-score 0.73 but a lower precision of 0.74. The kNN performed much worse than the others with an accuracy of 0.53 and a very low F1-score of 0.40. The accuracies by class for the neural network model using sBert embeddings are again shown using a confusion matrix (figure 7.3). Class 0 (within the system) has an accuracy of 0.69, class 1 (disruptive) has an accuracy of 0.64, and class 2 (encouragement) has an accuracy of 0.85 – this time all above the inter-rater agreement scores. Overall, these results and those of the binary models clearly indicate a better performance with the use of sBert embeddings compared to TF-IDF vectors.

Large Language Model

The results of fine-tuning the distilbert model can be found in table 7.7. The overall accuracy of the multi-class Distilbert model is 0.86, which is 0.12 percentage points higher than the inter-rater label agreement. The model also produced a precision of 0.84, a recall

sBert	Accuracy	Precision	Recall	F1-Score
kNN	0.71	0.70	0.71	0.69
Perceptron	0.75	0.75	0.75	0.74
SVM	0.76	0.76	0.76	0.74
Neural Network	0.76	0.76	0.76	0.75

Table 7.5: Metric scores for multi-class classification when using sklearn methods with sBERT sentence embeddings.

TF-IDF	Accuracy	Precision	Recall	F1-Score
kNN	0.53	0.49	0.53	0.40
Perceptron	0.75	0.77	0.75	0.72
SVM	0.73	0.74	0.73	0.67
Neural Network	0.75	0.74	0.75	0.73

Table 7.6: Metric scores for multi-class classification when using sklearn methods with TF-IDF vectorization.

of 0.85, and an F1-score of 0.83. By class, the model results in higher accuracies than all of the inter-rater agreement scores with class 0 having a testing accuracy of 0.85, class 1 having an accuracy of 0.81, and class 2 having an accuracy of 0.86.

7.3 Conclusion

These results determine that understanding relevancy of specific forms of activism using natural language processing and machine learning is possible. Furthermore, especially advanced techniques, such as large language models like Distilbert and embedding methods like sBert, perform especially well whereas unsupervised and traditional machine learning models are not as helpful for this specific task. Not only did Distilbert achieve high metric

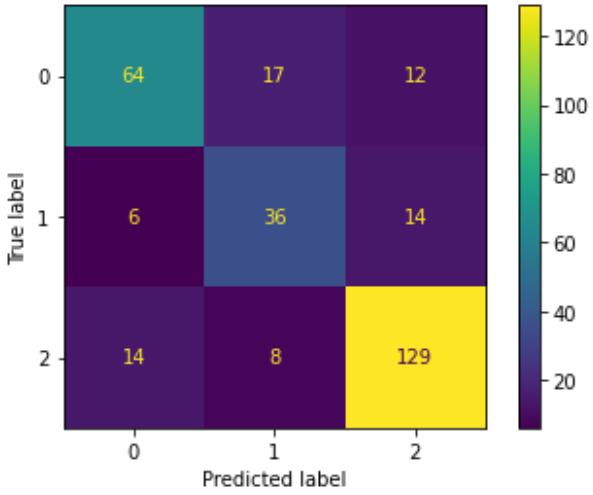


Figure 7.3: Confusion matrix displaying the true and predicted values from the multi-class neural network.

scores overall, it also far outperformed human labeler accuracies. This success confirms the supports the decision to use the fine-tuned Distilbert model to obtain binary and multi-class labels for the entire #BlackLivesMatter dataset.

Metric	Score
Accuracy	0.86
Precision	0.84
Recall	0.85
F1-Score	0.83

(a) Metrics

Label	Inter-Rater Agreement	Testing Accuracy
0 (Within the System)	0.55	0.85
1 (Disruptive)	0.59	0.81
2 (Encouraging)	0.80	0.89
Overall	0.74	0.86

(b) Inter-rater Agreement

Table 7.7: Performance of fine-tuned Distilbert Model for multi-class classification. (a) displays the accuracy, precision, recall, and F1-scores. (b) displays the inter-rate agreement scores and the by-class accuracy scores.

Chapter 8

Data Analysis Methods

This chapter will explain the methods used to analyze the dataset of 21 million tweets once the labels have been applied from the fine-tuned Distilbert machine learning models. There will be general data analysis of all of the tweets as well as more specific analysis of popular accounts and intersectionality with other social movements with the goal of answering research questions 2, 3, and 4.

8.1 General Data Analysis

To determine how calls to action have changed throughout the #BlackLivesMatter movement, time series analysis will be conducted. The proportion of irrelevant tweets vs. relevant tweets over the length of the movement will be examined followed by an analysis of the proportions of the three forms of activism. Since the movement is almost 10 years old, this time series analysis will be on the month scale rather than day or week which would cloud the analysis with too much unnecessary information. These figures will also provide context regarding how these proportions relate to police brutality events.

Because the broader analysis of the movement must be conducted on the month scale, additional time series analysis will be conducted to determine how these proportions change day-to-day in the month surrounding police brutality events. Nine different events will be examined, and these were chosen based on which sparked the most Twitter

activity. Looking more closely during these time periods could present subtle trends that would not be clear by looking at the overall movement.

Although topic modeling was not especially helpful in grouping tweets by form of advocacy, it can be used to provide insight into how the tweets were ultimately labeled by the models. The list of salient words provided by LDA topic modeling can show what words are most pertinent to a group of tweets. This method will be used individually with the groups of tweets labeled as disruptive and within the system to understand potential key words in tweets that signal a certain call to action. This will be done for the tweets surrounding different police brutality events to also demonstrate how these key words might change over time with the movement. Topic modeling will be conducted specifically for disruptive and within the system tweets because these are more concrete forms of activism.

8.2 Analysis by Type of User

As part of the CAPS Lab's research, a list of 1,000 Twitter accounts was created based upon which users present in the entire set of #BlackLivesMatter tweets had the most followers. Members of the CAPS Lab then labeled these popular accounts as being individuals or organizations and further labeled the accounts by the type of industry they belong to. The lists of these industries are presented in table 8.1. The proportions of calls to action for these accounts will be analyzed to determine if they differ widely from "everyday" users that are represented more so by the entirety of the dataset as analyzed in section 8.1. Additionally, whether these proportions vary widely between industries and how similar individuals vs. organizations are within the same industry will be examined. Of the popular accounts' tweets, the tweets with more than 150,000 total engagements (likes, retweets, comments, and quotes) will be examined specifically. These can provide information on which types of activism receive the most attention and support.

The CAPS Lab also compiled a list of 1,000 accounts that had the most tweets in the entire dataset – labeled as "prolific users". These should be the users who tweet the

most relevant information since they are so active; however, there are many spam accounts and accounts that use the hashtag solely to reach a greater audience rather than actually speaking on the movement. Using the labeled tweets can provide an understanding of what kinds of tweets these accounts are actually publishing in terms of whether they are actually aimed at the movement.

Industry	Number of Tweets
Business	38
Comedy	35
Fashion	15
Film	339
Music	551
News	288
Politics	186
Publishing	47
Religion	7
Social Media	389
Sports	97
TV	52

(a) Individuals' Industries

Industry	Number of Tweets
Activism	547
Business	71
Music, Art, and Film	127
News	5341
Politics	59
Publications	651
Social Media	759
Sports	261
TV and Radio	92

(b) Organizations' Industries

Table 8.1: Industries identified by CAPS Lab members and number of labeled tweets for popular accounts of individuals and organizations. (a) displays those of individuals. (b) displays those of organizations.

8.3 Intersectionality Analysis

The final area of analysis will be how the movement relates to intersectionality. Similar frames to those in Bonilla and Tillery's work will be utilized – their work examined LGBTQ+ issues, feminism, and Black Nationalism (Bonilla and Tillery, 2020). Yet, this work will look more broadly at issues relating to minority racial and ethnic groups instead of specifically just Black Nationalism because of the recent movements like #StopAsianHate and activism for Indigenous Peoples. To determine which tweets relate to these frames, the key words located in table 8.1 and variations of them will be searched for in tweets. Once the tweets have been obtained, time series analysis will again be used to determine how the proportion of relevant tweets by each topic has changed over time.

Topic	Key Words
LGBTQ+ Issues	LGBTQ+, gay, lesbian, bisexual, transgender, queer, pride month
Feminism	Feminism, women, #metoo, #sayhername
Marginalized Racial and Ethnic Groups	Asian, Indigenous, Hispanic, LatinX, colonialism, Black Nationalism, #StopAsianHate

Table 8.2: The frames and examples of the key words used to identify tweets talking about intersectional issues.

Chapter 9

Data Analysis Results

This chapter will present the results of analyzing the labels assigned by the chosen fine-tuned Distilbert model with the goal of answering the remaining research questions. Section 9.1 will answer question 2 about how the calls to action have changed over time. Section 9.2 will provide results specific question 3 about how trends differ with popular and prolific users, and section 9.3 will then answer question 4 about the intersectionality of #BlackLivesMatter. Although the proportion of irrelevant tweets will be considered in some of the analysis, the research questions will primarily be answered by looking at the tweets pertaining to the three forms of activism.

9.1 Calls to Action Over Time

The overall proportions for all tweets by label can be seen in figure 9.1. 48% of the tweets were labeled as irrelevant, 8% were labeled as within the system, 10% were labeled as disruptive, and 34% were labeled as encouragement. This emphasizes how much people use the hashtag unrelated to promoting the movement. The proportion over time of irrelevant and relevant tweets for the entire length of the movement as well as the total number of tweets by month is displayed by figure 9.2. The number of tweets sharply increased to over 1,218,000 during the month after Eric Garner's death. The number of tweets gradually declines to just over 130,000 the month right before Alton Sterling and

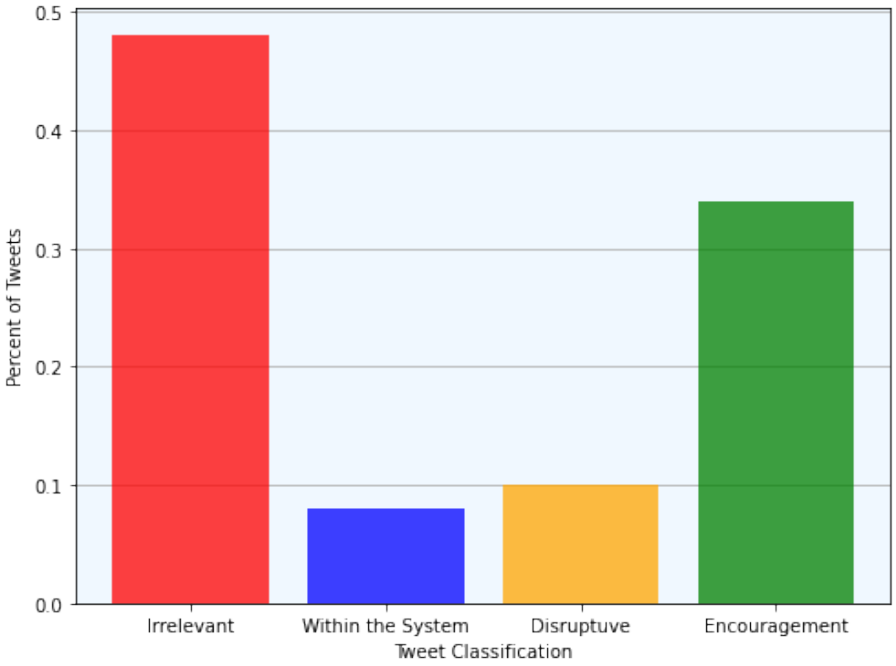


Figure 9.1: Percentage of tweets classified as each form of activism for all tweets.

Philandro Castile's deaths, but following these events, the number increases again to over 1,857,000. There is a long decline until around Breonna Taylor and George Floyd's killings in March and May of 2020, respectively. In June of 2020, the number of tweets reached its maximum with a total of 9,551,072 tweets. There is then a sharp drop until Daunte Wright's death in April of 2021. Following this event, the number of tweets sees another decline and reached the lowest point since the end of 2014 with a value of 21,116 tweets.

These patterns in the numbers of tweets confirm the findings of a some of the studies. The greatest numbers are typically around the months when police brutality events occur. There was more activity in the earlier part of the movement from 2014 through 2016, and then a resurgence in 2020. Yet, the data during the last part of 2021 demonstrates that perhaps this momentum has worn off. The trends do not align with the idea from other studies that George Floyd's killing was a turning point for the movement – there is no longer consistently more tweets than there were prior.

Prior to Eric Garner's death in July of 2014, the proportion of irrelevant tweets remained above 77% with an average of 89% irrelevancy. Following Tamir Rice's death in

November of 2014, that proportion remained around or below 75% for the rest of the time period with an average of 51% irrelevancy. During the spikes in number of tweets, such as during the months of Alton Sterling, Stephon Clark's, Daunte Wright's deaths, the proportion of irrelevant tweets sharply declines – the value of these decreases to near 40%. Compared to the number of tweets, the proportions of relevant vs. irrelevant tweets have remained much more consistent over the entire length of the movement following Eric Garner's death.

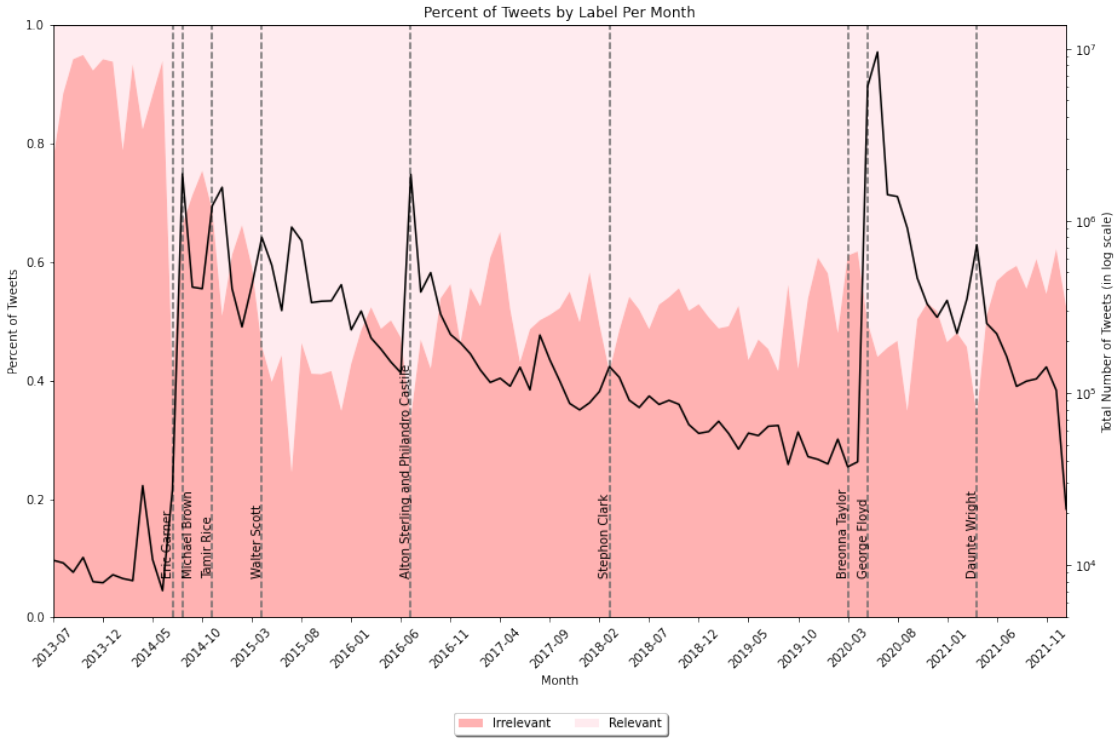


Figure 9.2: Percentages of all tweets classified as irrelevant or relevant over the entire length of the movement. The black line the total number of tweets. The figure also plots lines that signify when certain police brutality events occurred.

Instead of proportions of irrelevant vs. relevant tweets, figure 9.3 displays only the proportions of relevant tweets broken down by encouraging, disruptive, and within the system activism. The figure additionally shows the number of relevant tweets by month. The number of tweets reflects the same patterns as in figure 9.2 – there are sharp increases following events of police brutality with a gradual decline prior to Breonna Taylor's death

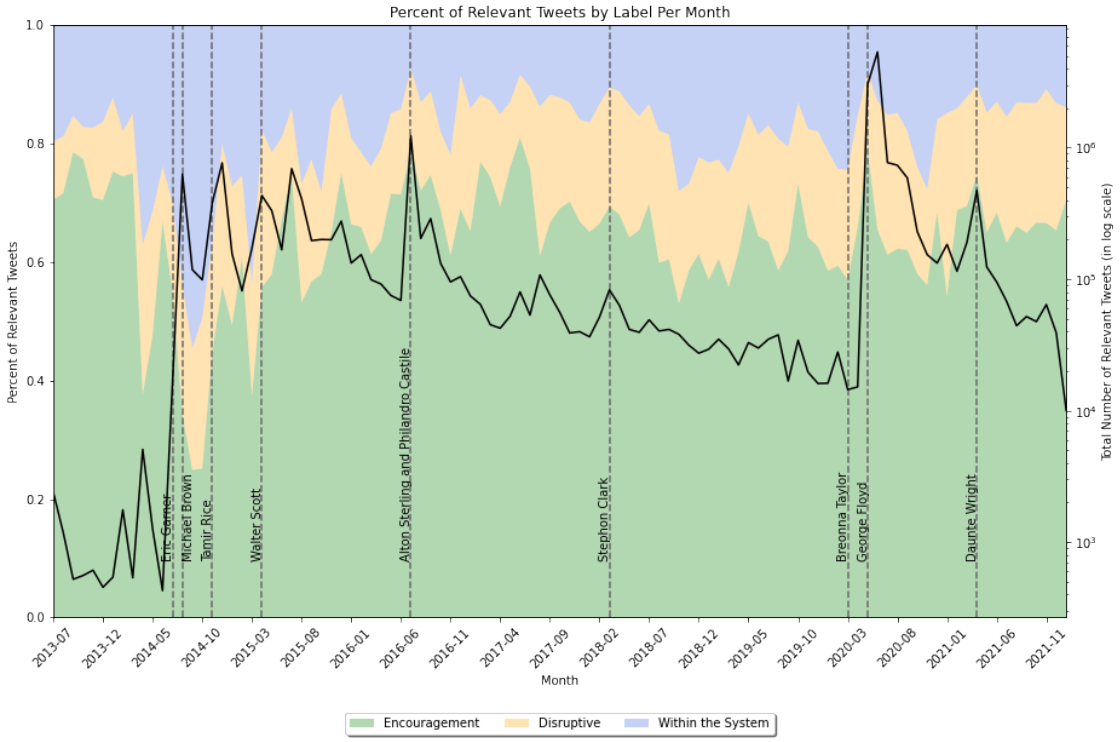


Figure 9.3: Percentages of all tweets classified as within the system, disruptive, or encouraging over the entire length of the movement. The black line the total number of tweets. The figure also plots lines that signify when certain police brutality events occurred.

and another decline since Daunte Wright's death.

The overwhelming proportion of relevant tweets are those that spread encouragement of the movement with the average percentage of encouraging tweets being 64%. The average percentage of disruptive and within the system tweets are 17% and 19%, respectively. Throughout the entire length of the movement, these proportions remain relatively consistent. The only period where this dynamic shifts is around the later portion of 2014 – between Eric Garner and Tamir Rice's deaths. During this period between April and November of 2014, the average proportion of within the system tweets increased to 39% and the proportion of encouraging tweets decreased to 42%. The proportion of disruptive tweets was relatively unaffected with an average of 20%.

Figures 9.4 and 9.5 display the number of tweets broken down by call to action for each day in the month surrounding the corresponding police brutality event. In almost all

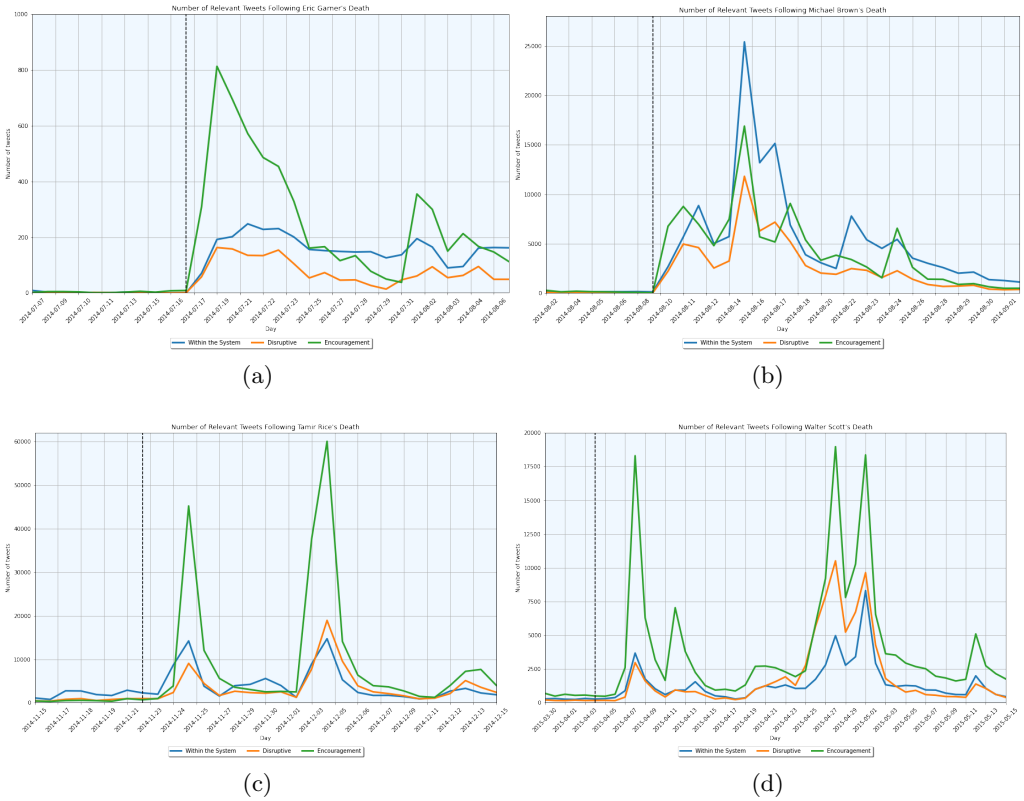
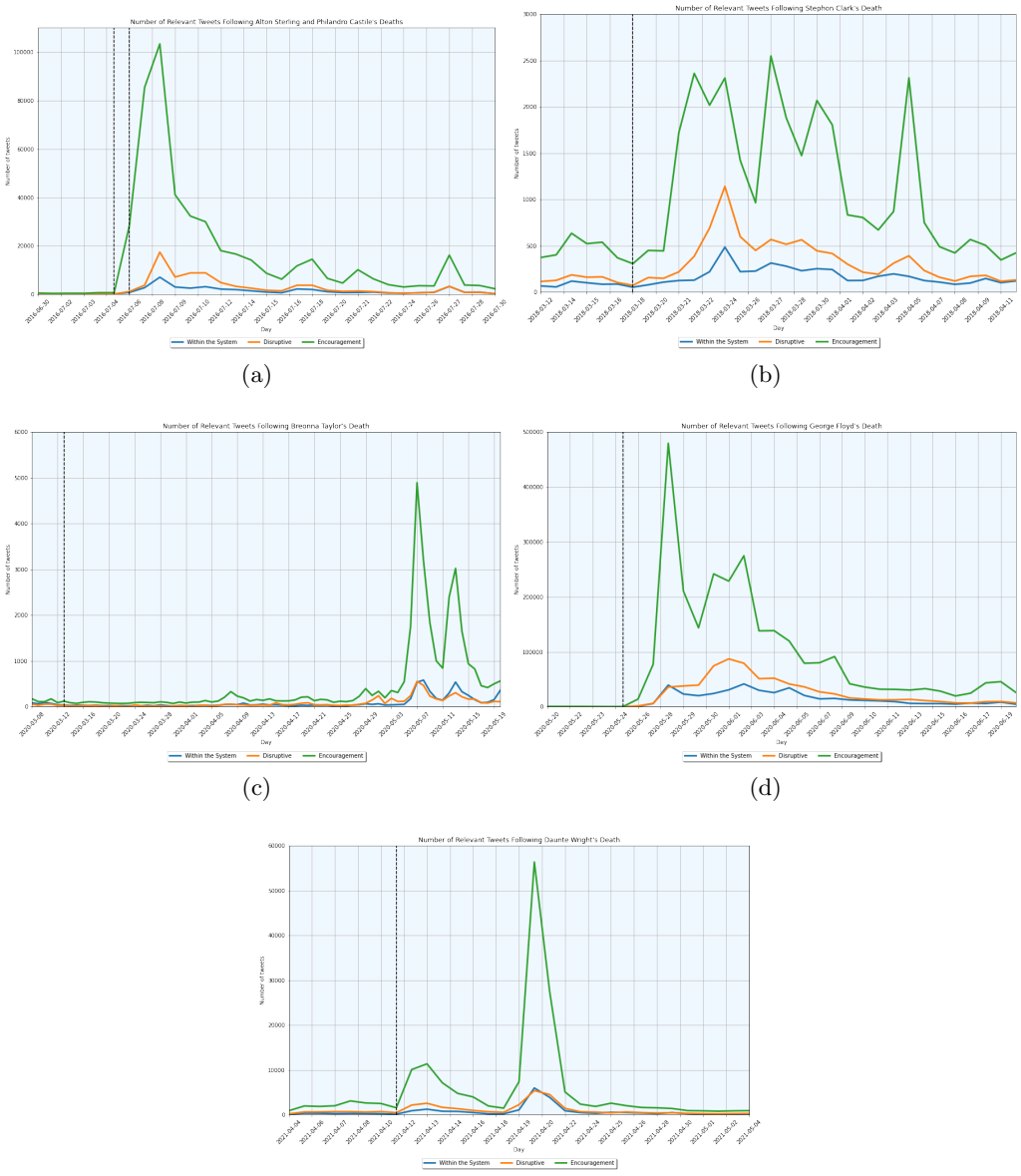


Figure 9.4: Number of tweets per call to action in the month following the killings of the following individuals (a) Eric Garner (b) Micahel Brown (c) Tamir Rice (d) Walter Scott

of the figures, the dominant call to action is encouragement of the movement; however, in figure 9.3b, which displays the time period surrounding Michael Brown's death, the dominant call to action is instead within the system activism.

Between just within the system and disruptive tweets, within the system is only more prevalent in figures 9.4a (Eric Garner) and 9.4b (Michael Brown). These also happen to be the first two police brutality events. In figures 9.5a (Alton Sterling and Philandro Castile), 9.5b (Stephon Clark), and 9.5d (George Floyd), disruptive tweets are instead more prevalent. In all of the remaining figures, the number of within the system and disruptive tweets are almost identical or there is no consistently dominant call to action between the two. With the first events leaning more towards within the system calls to action and later events advocating for disruptive forms of action, this could potentially demonstrate a move towards more disruptive actions to support the #BlackLivesMatter



(a) (b) (c) (d) (e)

Figure 9.5: Number of tweets per call to action in the month following the killings of the following individuals (a) Alton Sterling and Philandro Castile (b) Stephon Clark (c) Breonna Taylor (d) George Floyd (e) Daunte Wright

movement.

In both the graphs showing the time period surrounding Tamir Rice's death and Walter Scott's death, there are two distinct peaks in the number of tweets. Within these, during the first peak the number of within the system tweets outweighs the number of disruptive tweets. Yet, during the second peak, the number of disruptive tweets outweighs the number of within the system tweets. This is possibly a sign of growing frustration over the days and weeks following an individual's killing that results in greater emphasis on protests or boycotts.

Tables A.2 and A.3 show the top-15 most salient words provided by LDA topic modeling associated with within the system and disruptive tweets in the weeks after Michael Brown's and Daunte Wright's deaths, respectively. These two police brutality events were chosen because they come from two different time periods of the movement and differ in their percentage of tweets per call to action. As noted earlier, the tweets after Michael Brown's killing saw a much greater presence of disruptive activism compared to within the system activism whereas tweets surrounding Daunte Wright's killing were more evenly distributed between the two forms. The words regarding Michael Brown's death seem much more precise. For within the system tweets, listed words include gofundme, fundraisers, colorofchange, and profits – all words associated with fundraising and petitions. For disruptive tweets, the words include protests, gas/tear (listed separately but clearly referencing tear gas used in protests), march, and riot. All of these words clearly talk about the physical demonstrations especially prevalent following Michael Brown's death.

The within the system words for tweets surrounding Daunte Wright's death include accountability, firefischer, and usmayors, which could all be referencing the idea calling out political and other figures for lack of action regarding the movement. The disruptive words include protestors and protests but do not include any other words that seem to specifically evoke disruptive activism. Additionally, there are many words that overlap between the two groups of tweets – justice, guilty, and verdict are present in both lists. This perhaps shows that how disruptive and within the system activism are written about

in tweets is dependent on the specifics of a police brutality event rather than the general ideology of the movement.

9.2 User-Specific Results

Figure 9.6 displays the proportion of tweets per activism category by industry of popular accounts of individuals. Additionally, there is a column that displays the proportions of all tweets, regardless of popularity of the user. Overall, the percentage of within the system tweets is higher for almost all of these popular accounts – the only industry where this proportion is lower is music. For all tweets, within the system accounts for 15.9% of tweets, but this percentage is only 15% for the music industry. The second lowest category is sports with a percentage of 16.5%. Although music and all tweets have a similar proportion of within the system tweets, the music industry has far less disruptive tweets (10.1% vs. 18.7%) and far more encouraging tweets (74.8% vs. 65.5%). Religion notably has no disruptive tweets and an overwhelming majority of within the system tweets. News's proportions are almost evenly distributed between the three categories, which makes sense given that their reporting is supposed to be relatively unbiased. Tv, comedy, film, and politics all have relatively similar proportions of within the system tweets, but differ widely in their proportion of disruptive tweets – comedy and film have far fewer disruptive tweets.

Figure 9.7 displays a similar graph except it pertains to popular accounts of organizations rather than individuals. Social media has the greatest proportion of within the system tweets and has a relatively even split between the three categories. Almost all of the industries have a greater proportion of tweets supporting disruptive actions compared to that of all tweets. Activist groups and music related organizations have the highest proportion of disruptive tweets, while businesses have the lowest. This could be due to businesses' fear of potential financial impact that the other types of accounts do not have as much concern about.

Figure 9.8 displays these same proportions but for industries that appear in both of the

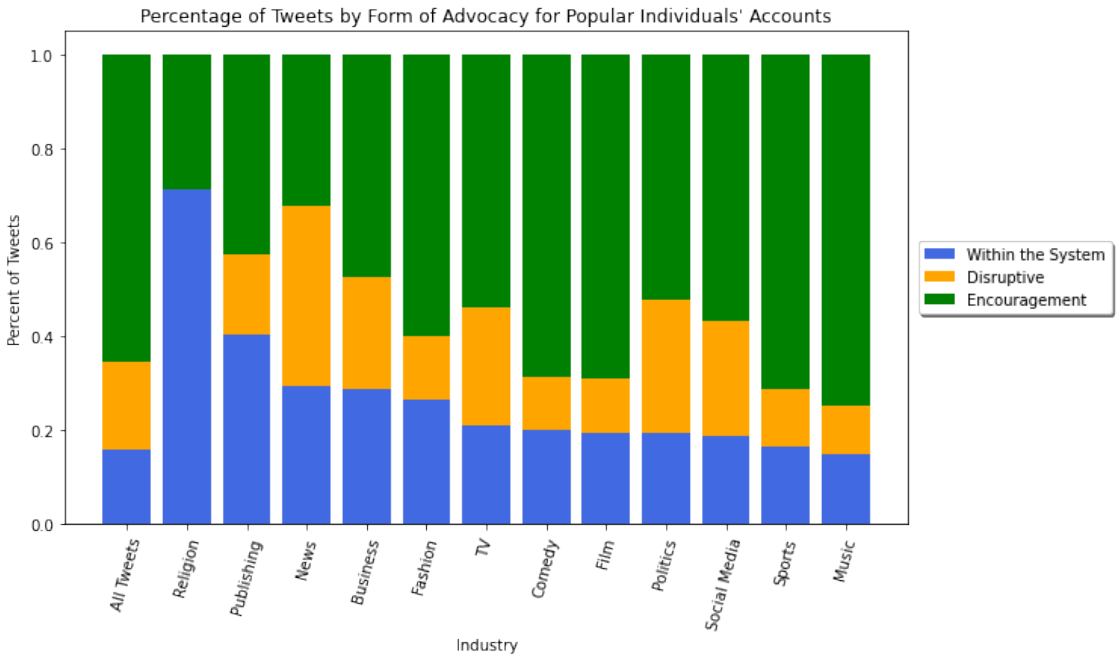


Figure 9.6: Breakdown of tweets by form of activism for accounts of popular individuals. The chart also includes the breakdown of these categories for all tweets for purpose of comparison.

previous figures. For tv/radio, news, and politics, the proportions are relatively similar; however, for the other categories, there is relatively substantial variation. This is especially apparent for individuals versus organizations in music/art/film, publications, and sports – individuals tweet far less about disruptive actions. Yet, individuals from publications and sports favor within the system activism more so than organizations. Additionally, organizations in music/art/film and social media tweet less about general encouragement of the movement. Overall, these proportions seem dependent on the industry in question with no clear trends in organizations' vs. individuals' preferences of calls to action, but popular accounts generally tweet more specific calls to action compared to all users.

Figure 9.9 displays the tweets by user and call to action that received the highest amount of engagement – retweets, quotes, and likes. Notably, 12 of the 40 tweets are by YourAnonCentral (also known as Anonymous), a hacker group that opposes and leaks data from government and corporate groups that promote inequality. Eight of their tweets promote disruptive calls to action, which makes sense given the context of their platform.

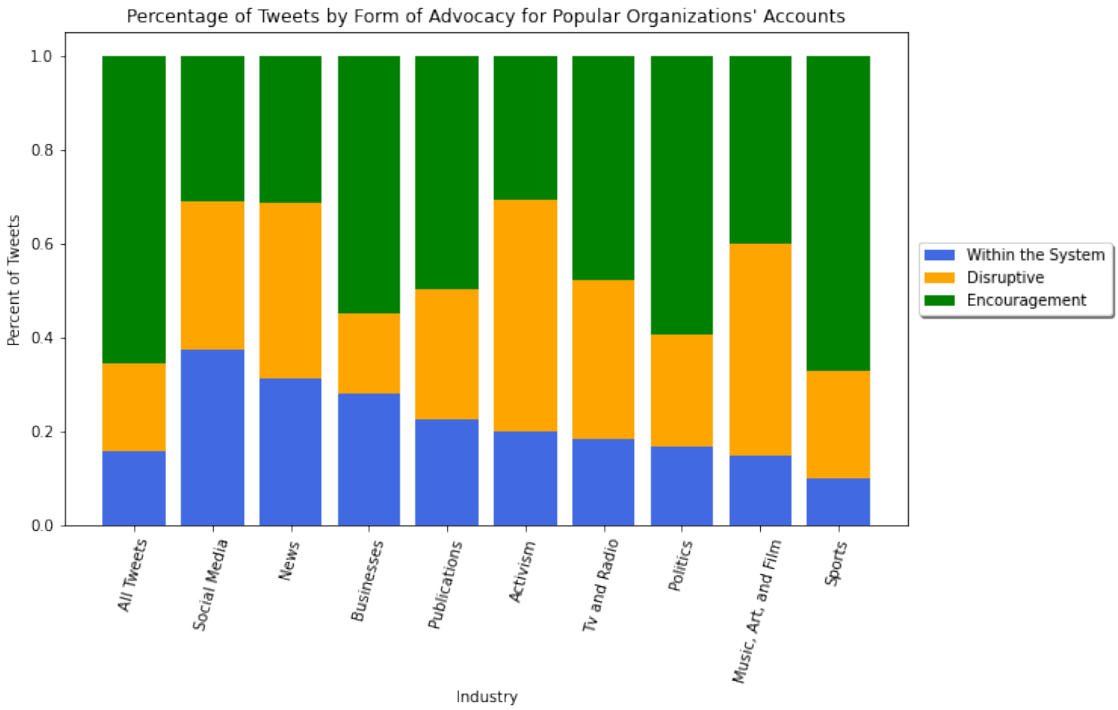


Figure 9.7: Breakdown of tweets by form of activism for accounts of popular organizations. The chart also includes the breakdown of these categories for all tweets for purpose of comparison.

Additionally, only nine of these popular tweets are about disruption at all. This is interesting given that YourAnonCentral is the only anonymous account whereas the other's are tied to extremely famous individuals or organizations.

This, along with there only being two within the system tweets, potentially show how extremely famous accounts lean more towards encouraging tweets when talking about #BlackLivesMatter because that form of activism is considered less divisive with members of their audience who may not be supporters of the movement. Another possible explanation is that these famous individuals are engaging in "slacktivism" – forms of activism that are "easily performed, and are considered more effective in making the participants feel good about themselves, than they are at achieving the stated political goals" (Morozov, 2009). Therefore, celebrities could be primarily tweeting this form of activism because it is commonly considered the easiest to do.

Figure 9.10 displays the proportion of tweets made by prolific accounts for each label.

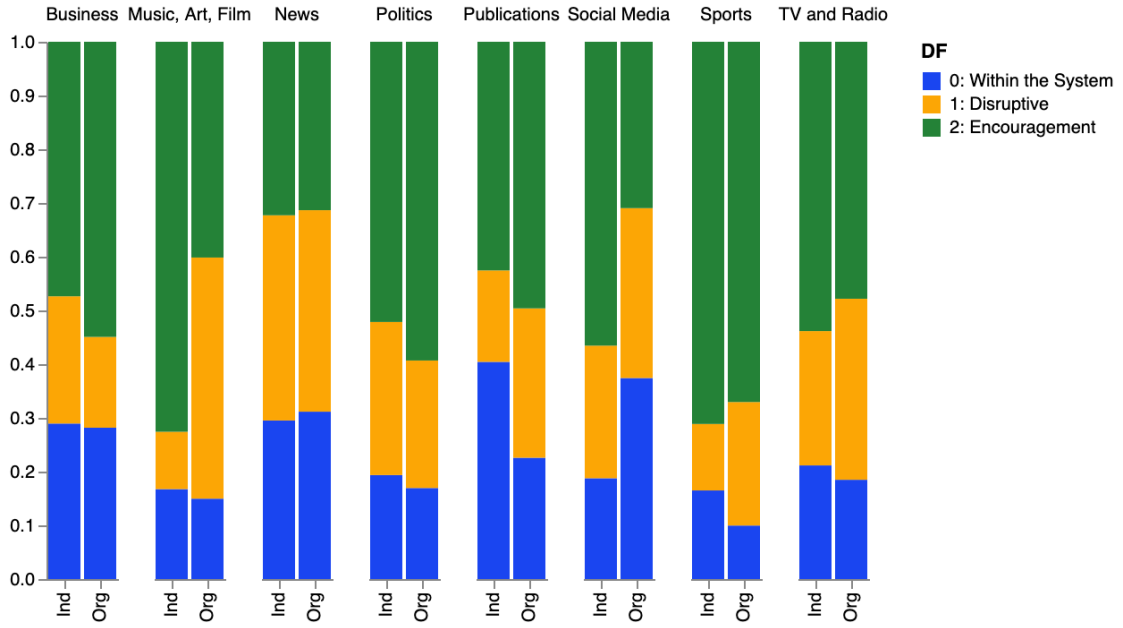


Figure 9.8: Breakdown of tweets by form of activism for organizations and individuals whose industries are present in both groups.

The trends are very similar to those shown in 9.1; however, there is a dramatic increase in irrelevant tweets and a slight decrease in encouraging tweets. This perhaps emphasizes that the accounts tweeting the most are not actually activists or supporters of the movement but spam or bot accounts using popular hashtags to increase engagement.

9.3 Analysis of Intersectionality

Figure 9.11 displays the proportion of relevant tweets that correspond to other social movements outside of #BlackLivesMatter. The three types of movements shown related to LGBT+ issues, feminism, and other minority races (Asian, Hispanic/LatinX, and Indigenous peoples). The most prominent of these is feminism. This intuitively makes sense because half of the Black population falls into this category, whereas the other two issues do not have as large of an overlapping population. There is a slight increase in intersectionality with issues regarding the LGBTQ+ community and other marginalized races; however, the predominant form of intersectionality regards feminist issues. All three is-

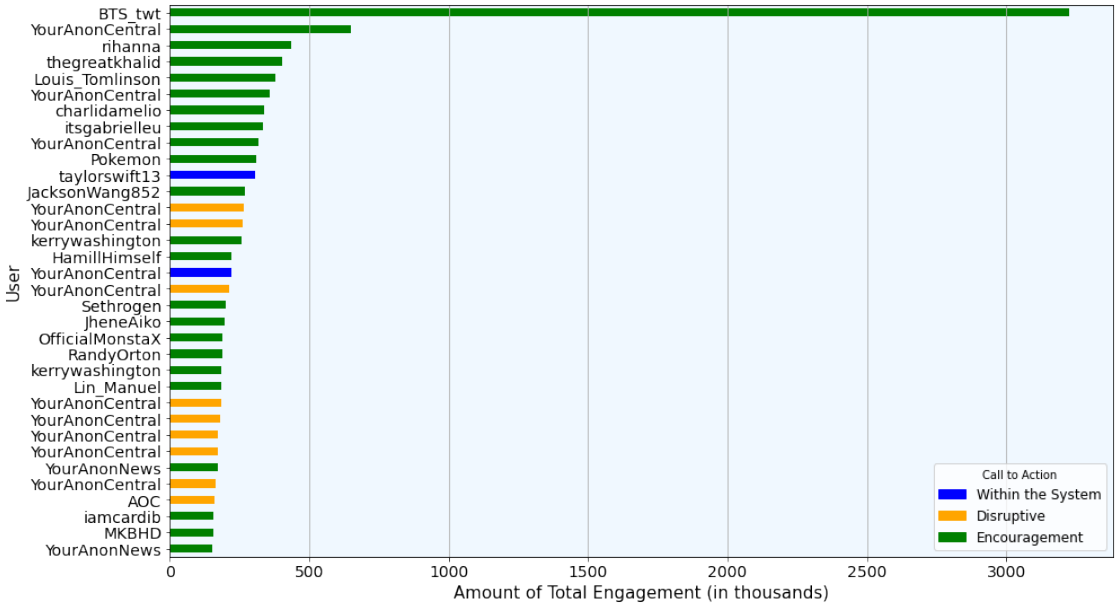


Figure 9.9: Bar chart for tweets with over 150,000 total engagements – likes, retweets, and quotes. The chart is sorted by total engagements descending and displays the classification of the tweet.

sues stay at similar proportions from July of 2013 through March of 2015, but then there is a drastic surge in the proportion of tweets related to feminism. After that we see the proportion related to feminism hover around 10%. This most likely correlates with the #SayHerName campaign which began in December of 2014 [1]. Additionally, there is a further increase in this proportion starting in late 2017. This can most likely be attributed to the beginning of the #MeToo movement around that time. Surprisingly, the proportion of feminism related tweets drastically drops after Breonna Taylor's killing in March of 2020 even though she is the only female victim of police brutality shown in the graph. This contradiction could be the result of the delay in attention to the Breonna Taylor case. The case did not receive much media attention until around two months after her actual killing [2]. This follows with what is seen in figure 9.11 because there is a resurgance in the proportion of feminism related tweets around May of 2020.

[1] 'About #SayHerName', *The African American Policy Forum*, https://www.aapf.org/sayhername, (accessed 5 April 2023)

[2] Jewél Jackson, 'Why Black Women Like Breonna Taylor Still Need 'Say Her Name' Movement', *Lousville Public Media*, 2020, https://www.lpm.org/2020-07-06/why-black-women-like-breonna-taylor-still-need-say-her-name-movement, (accessed 4 April 2023)

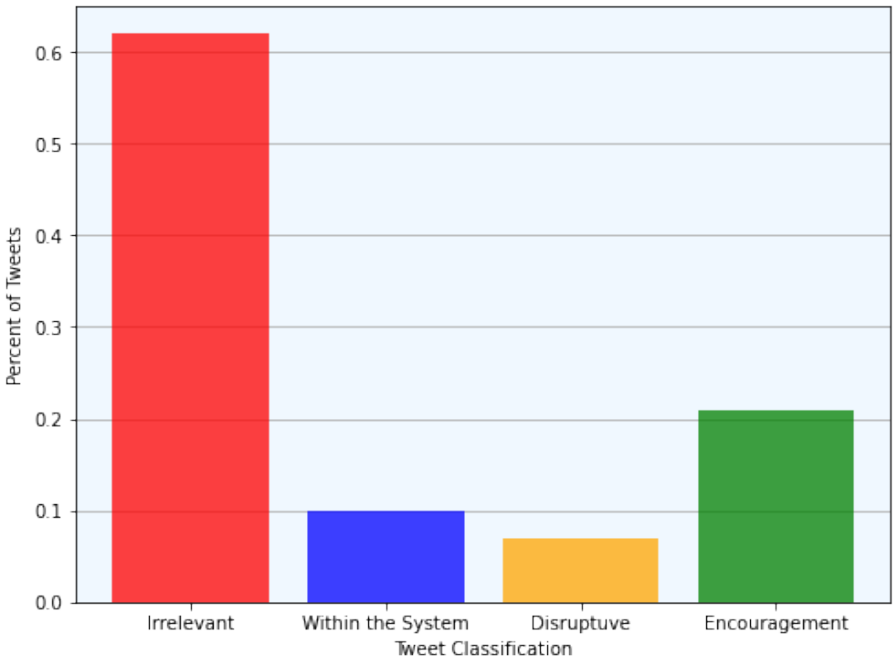


Figure 9.10: Percentage of tweets by form of activism for prolific accounts. These are the accounts that were identified for having tweeted the most about the movement.

Although supporters of the movement might not be tweeting about issues relating to the LGBTQ+ community and other marginalized races, it could be because of other explanations. One possibility is that most of these tweets were labeled as irrelevant because in the sample used to train the Distilbert model, there were no examples labeled regarding these intersectional issues. If the model did not see any of these tweets in training, it might not have the understanding that these types of tweets are still considered relevant to the movement.

9.4 Conclusion

The results indicate a general preference towards encouragement of and spreading information about #BlackLivesMatter as the predominant form of activism. Yet, following the initial police brutality events in 2013 and 2014, there has been a shift away from within the system and towards advocating for disruptive calls to action. When looking at popular accounts, encouragement still remains the majority form of activism, but the preference

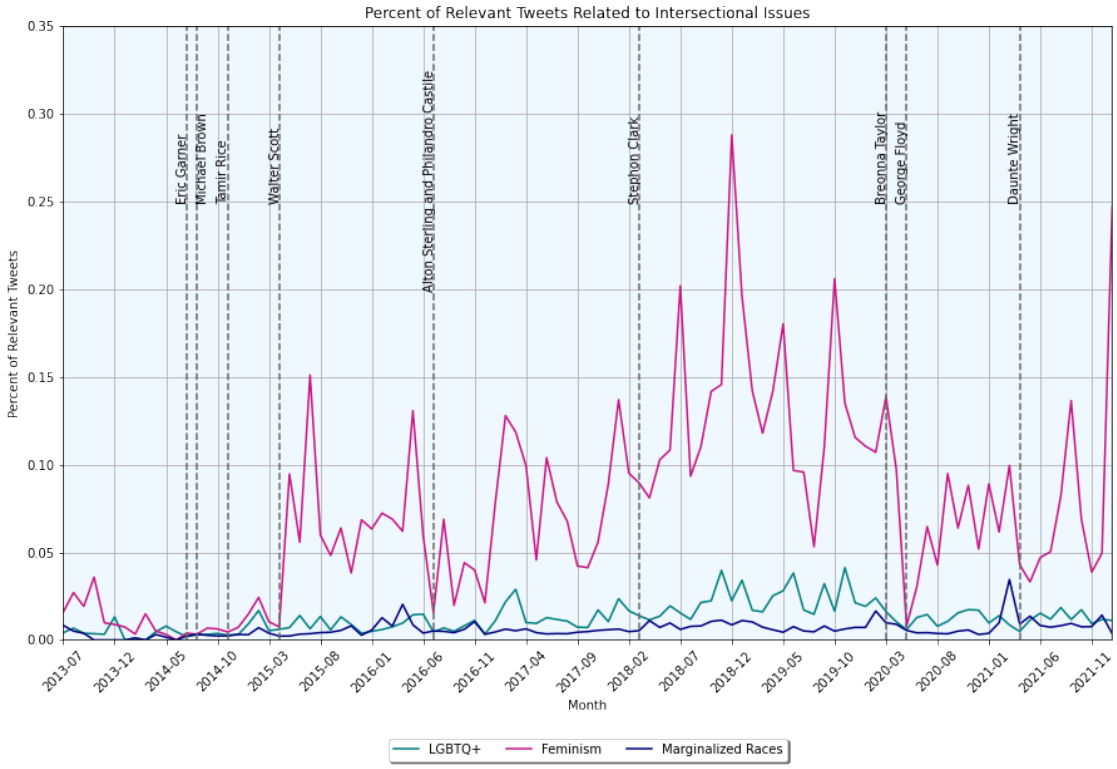


Figure 9.11: Proportion of tweets per month relating to three different intersectional social mvoements – LGBTQ+ issues, feminism, and social movements related to other marginalized races.

towards within the system or disruptive forms of activism varies between industries and whether the account represents an organization or an individual. Lastly, BLM is an intersectional movement, but this is primarily focused on feminist issues. Overall, these results can provide insights about the movement's classification as a New Social Movement which will be discussed further.

Chapter 10

Conclusion

10.1 Discussion

The primary motivation of this work was to understand how advanced data science tools can be utilized to answer social science questions. The specific lens used in studying this was the classification of activism in #BlackLivesMatter tweets and the opportunity for subsequent analysis of the movement. Through trials of natural language processing techniques and machine learning models, it was concluded that fine-tuning Distilbert for both binary and multi-class classification provided the best results in determining the call to action contained in a tweet. Not only did these models generate objectively high metrics, they also outperformed human accuracy – the overall goal of many machine learning algorithms. These results provide clear evidence that natural language processing tools can successfully be applied to social science contexts. Yet, simpler models, such as support vector machines and k-nearest-neighbors, performed worse, demonstrating that large language models are perhaps necessary in certain domains, such as #BlackLivesMatter tweets, where the amount of text is so extremely limited.

When analyzing the movement using the results of the classification models, there were clear relationships between the calls to action themselves as well as with outside factors – such as the type of user and intersectionality. Overall, encouragement has been the predominant form but when looking at disruptive and within the system calls to

action, there has been a shift towards disruptive calls to action since Tamir Rice's death. Compared to all tweets, those by popular accounts tweet less purely about encouragement of the movement and more about specific within the system or disruptive calls to action. There is no clear relationship between how organizations vs. individuals in a given industry tweets. Although these popular accounts publish fewer tweets strictly encouraging the movements, these are the tweets that gain the most support. Lastly, the amount of intersectionality, when looking at feminist, LGBTQ+, and other minority race issues, has diminished since 2020.

In terms of #BlackLivesMatter's classification as a New Social Movement, the results provide much supporting evidence. In support of its classification as this type of movement, BLM unequivocally galvanizes public support shown by the significant proportion of the tweets that outright promote and encourage the movement. Additionally, there has been a shift towards disruptive calls to action roughly starting after Tamir Rice's killing in 2014. This supports the idea that New Social Movements work through demonstrations rather than political or institutional channels. Finally, BLM is greatly focused on identities outside of the direct focus of the movement. The large proportion of tweets especially related to feminist issues emphasize the intersectional nature of #BlackLivesMatter that falls in line with the New Social Movement paradigm.

10.2 Limitations

10.2.1 Data Sample

The primary limitation of this work was the amount of labeled data that could be collected. Initially, this work was supposed to look at nine categories of activism, which would have provided more nuance, but it was not possible to get enough tweets labeled for each category within the scope of this project. Even when the number of categories was limited to three, there was still only a labeled sample of just over 1,100 tweets. For an entire dataset of 21 million data points and considering the possible variations of tweets, the

sample could have been improved by having more labeled tweets. Additionally, the sample was obtained from January through mid-May of 2020, which only represents a very specific time period of the movement. This does not capture tweets from the early days of the movement or even those from during and after the killing of George Floyd. It is necessary that this sample is as representative as possible since this directly impacts the success of the classification models to accurately interpret the tweets they are given.

10.2.2 Dataset

When looking at the entire dataset of 21 million tweets, these tweets were collected using hashtags related to #BlackLivesMatter; however, this might not be the best method to collect related tweets. Although during the early days of the movement and Twitter, hashtags were the primary avenue of participating in conversations surrounding BLM, many tweets today do not utilize hashtags because they do not offer the same success as they once did (Wagner, 2015). To better ensure all related tweets are being collected, this could instead require searching for tweets that contain these keywords anywhere in the body of the tweet rather than specifically the hashtag.

10.2.3 Classification Models

Tweets are not limited to expressing one form of activism, which could potentially explain why the inter-rater agreement scores were so low. Tweets could both support protests as well as voting for a certain politician, which would make it difficult to discern what the true value of the tweet's label should be. One solution would be hiring more labelers to ensure the accuracy of the chosen label; however, a possibly more robust solution would be labeling the sample in a way that supports the use of a multi-label classification model which can assign a datapoint to multiple groups.

10.3 Future Work

The primary avenue for future work would be obtaining a more comprehensive sample and again looking at all nine possible labels of activism. This could provide extremely interesting insights on how specific forms of disruptive, within the system, and encouraging calls to action differ in their popularity. A direction that would altogether avoid the limitations provided by the sample used to train the classification model would be utilizing zero-shot classification. This machine learning approach uses a pre-trained large language model, such as BERT, and a set of provided labels to classify data (Lampert et al., 2009). It does not require any training on a sample of the data but instead uses natural language processing methods to automatically understand the relationship between the labels and the documents looking to be classified. Additionally, tweets of more types of users, such as politicians and grassroots activist organizations, could be analyzed in addition to purely popular accounts. Although some of these were captured in the dataset of popular accounts, they might not be representative of all accounts from these areas of the movement.

Another avenue includes looking at other social media platforms. Although Twitter was the initial home base for the #BlackLivesMatter, the movement has since become extremely popular on other platforms as well. This was especially apparent during the George Floyd protests – there was a high volume of activity among BLM related hashtags on both Instagram and Tik Tok [1]. Additionally, besides other platforms becoming relevant, in the past year Twitter has seen many changes due to the change in ownership. Many users have decided to leave the platform altogether. Together, these suggest that looking to other platforms besides Twitter might provide a more holistic and accurate analysis of the movement.

[1]Rachel Janfaza, 'TikTok serves as hub for #blacklivesmatter activism', *CNN*, 2020, (accessed 6 April 2023)

Appendix A

LDA Topic Modeling Results

Table A.1: Top 5-salient words outputted by LDA topic model for each of 20 topics.

Topic	Word 1	Word 2	Word 3	Word 4	Word 5
1	citizens	org	act	misconduct	enact
2	blacktwitter	thursday	facebook	pls	vigil
3	blm	racist	white	racism	city
4	going	make	brown	ppl	petebuttigieg
5	vote	need	address	word	right
6	&	birthday	unarmed	today	way

7	sayhername	breonna	america	taylor	years
8	irunwithmaud	fuck	let	speak	sandrabland
9	lives	matter	berniesanders	boycott	support
10	think	stop	cops	ericgarner	community
11	breonnataylor	killed	statement	murder	needs
12	barackobama	protestors	said	saying	better
13	ahmaudarbery	say	murdered	free	men
14	policebrutality	life	day	movement	voting
15	blackhistory month	justiceforall	history	times	protests
16	irunwithmaud	peaceful	protest	time	solidarity
17	stand	tired	run	trying	fight
18	justicefor ahmaudarbery	power	women	men	happening
19	peace	tell	american	really	justiceforahmaud

20	blackhistory	want	home	video	new

Table A.2: Top-15 most salient words from topic modeling performed on the tweets in the weeks after Michael Brown's death. (a) shows those from the tweets labeled as within the system, and (b) shows those from the tweets labeled as disruptive.

(a) Within the System Tweets

Number	Word
1	youtube
2	gofundme
3	wilson
4	watch
5	officer
6	colorofchange
7	johnson
8	return
9	fundraisers
10	profits
11	twitter
12	calling
13	michael
14	state
15	gas

(b) Disruptive Tweets

Number	Word
1	twitter
2	pic
3	instagram
4	change
5	wilson
6	darren
7	robbery
8	protests
9	gas
10	tear
11	missouri
12	ericgarner
13	murder
14	march
15	riot

Table A.3: Top-15 most salient words from topic modeling performed on the tweets in the weeks after Daunte Wright's death. (a) shows those from the tweets labeled as within the system, and (b) shows those from the tweets labeled as disruptive.

(a) Within the System Tweets

Number	Word
1	served
2	justice
3	guilty
4	georgefloyd
5	chauvin
6	thing
7	derek
8	accountability
9	verdict
10	usmayors
11	firefischer
12	care
13	doesn
14	sayhername
15	jury

(b) Disruptive Tweets

Number	Word
1	guilty
2	chauvin
3	dauntewright
4	defense
5	brooklyn
6	justice
7	daunte
8	wright
9	derek
10	verdict
11	minneapolis
12	protestors
13	prosecution
14	protests
15	nelson

Bibliography

Anderson, M. (2018). An analysis of #BlackLivesMatter and other Twitter hashtags related to political or social issues. Retrieved March 8, 2023, from https://www.pewresearch.org/internet/2018/07/11/an-analysis-of-blacklivesmatter-and-other-twitter-hashtags-related-to-political-or-social-issues/

Ankita, Rani, S., Bashir, A. K., Alhudhaif, A., Koundal, D., & Gunduz, E. S. (2022). An efficient CNN-LSTM model for sentiment detection in #BlackLivesMatter. *Expert Systems with Applications*, *193*, 116256. https://doi.org/10.1016/j.eswa.2021.116256

Badaoui, S. (2020). *Black Lives Matter: A New Perspective from Twitter Data Mining* (tech. rep.). Policy Center for the New South. https://blog.sodipress.com/wp-content/uploads/2020/11/black-lives-matter-A-new-perspective-from-twitter-data-mining.pdf

BERTopic: Neural topic modeling with a class-based TF-IDF procedure [arXiv:2203.05794 [cs]]. (2022). https://doi.org/10.48550/arXiv.2203.05794

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bonilla, T., & Tillery, A. B. (2020). Which Identity Frames Boost Support for and Mobilization in the #BlackLivesMatter Movement? An Experimental Test. *American Political Science Review*, *114*(4), 947–962. https://doi.org/10.1017/S0003055420000544

Church, K. W. (2017). Word2vec. *Natural Language Engineering*, *23*(1), 155–162.

Della Porta, D., & Diani, M. (2006). *Social movements: An introduction* (Second). Blackwell Publishing.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [arXiv:1810.04805 [cs]]. Retrieved February 27, 2023, from http://arxiv.org/abs/1810.04805

Edrington, C. L. (2022). Social Movements and Identification: An Examination of How Black Lives Matter and March for Our Lives Use Identification Strategies on Twitter to Build Relationships. *Journalism & Mass Communication Quarterly*, *99*(3), 643–659. https://doi.org/10.1177/10776990221106994

Edrington, C. L., & Lee, N. (2018). Tweeting a social movement: Black lives matter and its use of twitter to share information, build community, and promote action. *The Journal of Public Interest Communications*, *2*(2), 289–289. https://doi.org/10.32473/jpic.v2.i2.p289

El-Zanaty, Z. (2019). Zeyad at SemEval-2019 Task 6: That's Offensive! An All-Out Search For An Ensemble To Identify And Categorize Offense in Tweets. *Proceedings of the 13th International Workshop on Semantic Evaluation*, 823–828. https://doi.org/10.18653/v1/S19-2144

Freelon, D., Mcilwain, C., & Clark, M. (2016). Beyond the Hashtags: #Ferguson, #BlackLivesMatter, and the online struggle for justice.

Garza, A. (2016). A Herstory of the #BlackLivesMatter Movement [Google-Books-ID: Y7ErDAAAQBAJ]. In *Are All the Women Still White? Rethinking Race, Expanding Feminisms*. SUNY Press.

Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S., & Gurusamy, V. (2014). Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.

Lampert, C., Nickisch, H., & Harmeling, S. (2009). Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer.

Lebron, C. J. (2017). *The making of black lives matter: A brief history of an idea*. Oxford University Press.

Lee, C. S., & Jang, A. (2021). Questing for Justice on Twitter: Topic Modeling of #StopAsianHate Discourses in the Wake of Atlanta Shooting. *Crime & Delinquency*. https://doi.org/10.1177/00111287211057855

MacDonald, F., & Dobrowolsky, A. (2020). *Turbulent times, transformational possibilities?: Gender and politics today and tomorrow* [Google-Books-ID: P8bhDwAAQBAJ]. University of Toronto Press.

Mishra, S., Agarwal, S., Guo, J., Phelps, K., Picco, J., & Diesner, J. (2014). Enthusiasm and support: Alternative sentiment classification for social movements on social media. *Proceedings of the 2014 ACM conference on Web science - WebSci '14*, 261–262. https://doi.org/10.1145/2615569.2615667

Mitchell, T. M., et al. (2007). *Machine learning* (Vol. 1). McGraw-hill New York.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Pichardo, N. A. (1997). New social movements: A critical review. *Annual Review of Sociology*, *23*(1), 411–430. https://doi.org/10.1146/annurev.soc.23.1.411

Qi, L., Li, R., Wong, J., Tavanapong, W., & Peterson, D. A. M. (2017). Social Media in State Politics: Mining Policy Agendas Topics. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 274–277. https://doi.org/10.1145/3110025.3110097

Ramos, J., et al. (2003). Using tf-idf to determine word relevance in document queries. *Proceedings of the first instructional conference on machine learning*, *242*(1), 29–48.

Raschka, S., Liu, Y., Mirjalili, V., & Dzhulgakov, D. (2022). *Machine learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python*. Packt Publishing.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter [arXiv:1910.01108 [cs]]. Retrieved February 22, 2023, from http://arxiv.org/abs/1910.01108

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [arXiv:1908.10084 [cs]]. (2019). Retrieved February 22, 2023, from http://arxiv.org/abs/1908.10084

Srivatsa, S., Mohan, T., Neha, K., Malakar, N., Kumaraguru, P., & Srinivasa, S. (2022). Zero-shot Entity and Tweet Characterization with Designed Conditional Prompts and Contexts. https://doi.org/10.48550/ARXIV.2204.08405

Tillery, A. B. (2019). What kind of movement is black lives matter? the view from twitter. *The Journal of Race, Ethnicity, and Politics*, *4*(2), 297–323. https://doi.org/10.1017/rep.2019.17

Tong, X., Li, Y., Li, J., Bei, R., & Zhang, L. (2022). What are People Talking about in #BackLivesMatter and #StopAsianHate? Exploring and Categorizing Twitter Topics Emerging in Online Social Movements through the Latent Dirichlet Allocation Model. https://doi.org/10.48550/ARXIV.2205.14725

Wirtschafter, V. (2021, June 17). *How george floyd changed the online conversation around BLM* [Brookings]. Retrieved March 10, 2023, from https://www.brookings.edu/techstream/how-george-floyd-changed-the-online-conversation-around-black-lives-matter/

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, *10*(5), 593. https://doi.org/10.3390/electronics10050593