Who Supports Black Lives Matter Protests?: A Machine Learning Approach

Eric Niblock (ejn9259), Liben Chen (lc4438), Sagar Patel (skp327), Burcu Kolcak (bk1959)

Abstract

Text classification regarding political and social sentiment on Twitter is valuable for research into collective-sense making processes (CSMP), as well as practical endeavors, such as the evaluation of community thought and well-being. Creating semi-automatic models capable of sentiment prediction, we focus on the Black Lives Matter (BLM) movement corresponding to the recent Floyd protests. Our models achieve reasonable levels of overall accuracy across three classes (Decision Tree (DT): 0.51 accuracy, Logistic Regression (LR): 0.51 accuracy, Deep Learning (DL) Model: 0.36 accuracy). Furthermore, we provide an example of the model's potential, comparing positive sentiment towards BLM protests with positive sentiment towards the now President-elect Biden. Using our models, as well as ordinary and weighted least squares regression (OLS and WLS, respectively), we found that these sentiments are correlated (DT, WLS: $p = 0.001$; LR, OLS: $p = 0.004$).

1 Introduction

In May 2020, George Floyd, an unarmed Black man, was killed while in police custody in Minneapolis. In the aftermath of Floyd's killing, protests against police brutality have spread in cities across the United States. The Floyd protests occur among a long history of protests for racial justice in the United States and the Black Lives Matter (BLM) movement, which has mobilized protests and other forms of political action since 2013, especially in local areas with more frequent police killings of Black civilians (Davenport, Soule and Armstrong, 2011; Reynolds-Stenson, 2018)

Unlike previous protests, the press explains the difference of current Floyd Protests by highlighting the convergence of the pandemic, economic devastation and police violence, the shift in public attitudes toward racial attitudes, and the role of social media as an underlying factor that triggers the magnitude and scope of the BLM movement in the aftermath of Floyd's killing (Badger, 2020).

Previous work on the potential role of social media in understanding the BLM movement

and protests against police brutality has studied social media platforms such as Facebook and Twitter. A few examples will suffice to illustrate this point. For example, Ince, Rojas and Davis (2017) examines how social media users contribute to the framing of BLM by using hashtags. Mundt, Ross and Burnett (2018) focus on the role of social media in strengthening social movements by analyzing public social media accounts and interviews.

Our study differ from previous work on multiple fronts. First, previous studies focus on the BLM movement prior to the killing of Floyd. Many scholars and journalists suggest that the role of social media in strengthening recent protests has been vital considering the magnitude and scope of the BLM movement in the aftermath of Floyd's killing. Second and most importantly, the goal of many previous studies has not been prediction.

In this paper, using Twitter data, we develop a model to identify support for the BLM movement in the aftermath of Floyd's killing. Social media platforms like Facebook and Twitter contain vast amount of data that is hard to moderate. To better understand the link between social media and the BLM movement, machine learning can be used to identify potential support for the BLM movement in larger collections of text. The construction of such a model is useful for a plethora of reasons. It can provide communities a form of social sense and awareness during times of instability, but furthermore, it can serve as a standard for developing issue-based classification methods for studying CSMP, and social, predicitive-analytics.

2 Data Understanding and Preparation

Our dataset is composed of tweets that we scraped from Twitter using the *Nasty* library.[1] We searched for tweets containing hashtags associated with BLM protests. Consistent with our theoretical considerations, we identified 28 hashtags listed in Table 5 of the Appendix.

We predicted a subset of these hashtags to be highly correlated with support (#BlackLivesMatter, #NoJusticeNoPeace, #GeorgeFloyd, etc.) while others would be highly correlated with dissent or critique (#AllLivesMatter, #BackTheBlue, #Riots, #ThinBlueLine, etc.).

[1] Nasty library is a tool for retrieving Tweets via the Twitter Web UI instead of using the Twitter Developer API

The sample period starts at the beginning of March, 2020, and ends at the end of October, 2020 to capture the dialogue surrounding the shifting sociopolitical climate encompassing race and racial policing in the United States in the aftermath of Floyd's killing.

The scraping technique provided the raw text associated with each tweet, which would eventually be processed and converted to a set of features in order to train our models. However, the text is not accompanied by a feature denoting explicit political sentiment (support, opposition, or neutral). For this reason, we derived a method by which to label the data with this target variable. The convention used for our analysis was that a label of 1 would denote support for the protests, a label of 0 would indicate neutrality or irrelevance with regard to the protests, and a label of -1 would indicate opposition to the protests.

2.1 An Alternative to Manual Labeling

Manual labeling is a widely used tool for studies that rely on machine learning classification using Twitter data. For example, to detect hate speech using Twitter data, Davidson et al. (2017) annotated a corpus of social media posts by matching twenty-five thousand samples with a hate speech lexicon. Yet, manual labeling comes with many difficulties concerning text classification. In our case of sentiment analysis concerning a specific sect of political speech, these difficulties arise in terms of (1) labor and expense, (2) semantic ambiguity/complexity, and (3) inconsistency within those responsible for labeling. The drudgery of (1) is relatively self evident - determining the sentiment of text is a complex and tedious process which takes more time than other labeling tasks (such as labeling photos or trends). Especially when faced with pressure to produce, the removal of manual labeling from model construction can save time for data scientists, or expense if the task is instead outsourced. Furthermore, as far as (2) goes, the text found within tweets and other social media posts is often out-of-context or lacking in self-evident meaning. This is in part due to the casual and informal nature of social media, with posts often containing various topics and meanings, along with mistakes and grammatical errors. This further complicates the task of manual labeling. Dovetailing from this, (3) the manual classification of tweets is often tackled by multiple labelers, who all must operate by use of a consistent set of rules, which is never fully comprehensive. The sample

space of speech and sentiment is so vast that no matter the contrived system of rules produced in order to sort text, there will always exist edge cases and complications that produce variability between labelers, and even variability within individual labelers themselves.

Various approaches have been used to avoid the problems associated with manual labeling. Go, Bhayani and Huang (2009) employed the use of emojis to classify social media posts, a proxy for direct access to the emotional information embedded within the text. In our case, we presume that the use of hashtags provides a similar insight into the political attitude and emotion regarding BLM. A similar method has been employed in previous work, though without the narrow focus of a social/political topic (Hasan, Agu and Rundensteiner, 2014). However, unlike the purely automated approach of using hashtags as indicators, our approach is somewhat of a hybrid model, in that we first determined which hashtags corresponded to tweets best expressing either support or dissent of the BLM protests.

We randomly sampled 100 tweets from each of the 28 scraped hashtags and manually labeled that subset. We determined that #BackTheBlue and #NoJusticeNoPeace were the most pure in terms of correlating with our suspected categorizations, that is, they possessed the lowest rates of negated expression (using #NoJusticeNoPeace in a manner that is unsupportive of protests, or using #BackTheBlue in a manner that is supportive of protests). In Table 1, we present the descriptive statistics of three hashtags we use for our models.

Table 1

Hashtag	Sampled Positive	Sampled Neutral	Sampled Negative	Total Tweets	Total Cleaned
#NoJusticeNoPeace	83	17	0	5297	5291
#BackTheBlue	0	27	73	6280	5675
Neutral/Irrelevant	0	100	0	7364	5701

Results of our random sample from three sources. #NoJusticeNoPeace and #BackTheBlue proved most successful in representing support and opposition, respectfully. Neutral/Irrelevant tweets were collected by scraping Twitter with no reference to hashtag, and clearly have little to no relevant sentiment.

2.2 Data Preparation and Cleaning

In order to ensure the purity of each class, further cleaning was employed in order to automatically remove as many neutral or irrelevant tweets from our support and oppositional sets. Again, we took a hybrid approach, where the data was visually inspected for the frequent use of hashtags which signaled irrelevancy (for example, #BlueTigers was often used in conjunction with #BackTheBlue, an expression of support for a sports team, and irrelevant in the context of our study). Any tweets which possessed these signaling hashtags were subsequently removed. Further preprocessing, such as the removal of punctuation, emojis, spacing, and irrelevant characters was run on all of the data. Most crucially, all of the hashtags were removed from the data, as to avoid the model simply learning the automated process we employed to create the data. Table 2 demonstrates the cleaned text, and what constitutes each classification (1 being supportive, 0 being neutral, and -1 being oppositional).

Table 2

Label	Unprocessed Tweet	Preprocessed Tweet
1	And you really think things have "progressed" and changed ???? #NoJusticeNoPeace #NateWoods	And you really think things have progressed and changed
1	Some local ladies and I are going to #RunWithMaud tomorrow morning. Because we are mothers and can't imagine the grief Ahmaud's mother will feel this Mother's Day. #NoJusticeNoPeace #BlackLivesMatter	Some local ladies and I are going to tomorrow morning Because we are mothers and cant imagine the grief Ahmauds mother will feel this Mothers Day
0	Imma need the Rona to get it together before July.... it's my golden birthday this year \U0001f97a	Imma need the Rona to get it together before July its my golden birthday this year
0	Okay guys for Wednesdays' YouTube upload would you like a star wars jedi fallen order video or a complication of funny, epic, & fail moments(clips) of my gaming moments on stream?	Okay guys for Wednesdays YouTube upload would you like a star wars jedi fallen order video or a complication of funny epic fail moments clips of my gaming moments on stream
-1	#RacistDemocrats hate that Betty and Jorge Rivas, owners of Sammy's Mexican Grill, support President Trump. If you live in the Tucson area, please visit them! They're pro-law enforcement, too!\n\n#HispanicsForTrump #LatinosForTrump #BackTheBlue #MAGA\n\nhttps://t.co/3kZPMqwSMZ https://t.co/zHcGMayYqi"	hate that Betty and Jorge Rivas owners of Sammys Mexican Grill support President Trump If you live in the Tucson area please visit them Theyre prolaw enforcement too
-1	Democrats war on law enforcement is dangerous for LEOs & our communities.\n \nWe must reject their scare tactics designed to teach our kids to fear law enforcement.	Democrats war on law enforcement is dangerous for LEOs our communities We must reject their scare tactics designed to teach our kids to fear law enforcement

2.3 Minimizing the Role of Selection Bias

There are numerous sources of selection bias within our model. The decision to use only specific hashtags in an attempt to model an incredibly extensive and variable sample space (that is, event-based political expression) may exclude certain actors and their methods of expression regarding the current social movements. However, this would represent an issue in most methods — it does not seem that there exists a feasible method by which to determine if the subset of chosen tweets is fully representative of the population of individuals expressing opinions on this issue.

There also is a potential for the control tweets to be biased. With approximately 8,000 samples, it is highly unlikely that our control data is representative of the neutral class (any speech unrelated to our sample space of concern). Furthermore, it is possible that some of the control data does in fact contain tweets relevant to our sample space. However, this possibility is mitigated considering that our sample space is relatively small compared to the entirety of tweets produced on any given time period.

3 Modeling & Evaluation

3.1 Decision Tree

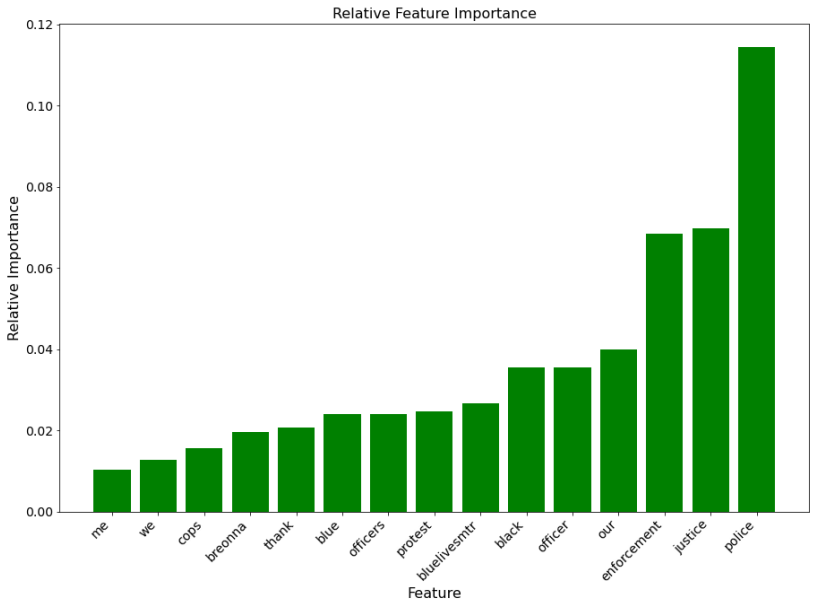
All of our baseline models made use of a vectorizer, which mapped all of the training data (composed of the above sets) into an $n \times d$ matrix, of n tweets and d features. Tokenization, the process of creating these d features, occurred at a very base level — every unique word that appeared more than five times throughout the total body of words used was considered a feature, along with every n-gram of length two. The cut off of five was employed to ensure that processing time did not become unmanageable, and the use of n-grams has widely been seen as a method to preserve sentence structure, meaning and coherence, which is especially important in attempting to derive sentiment from text (Violos et al., 2018).

Furthermore, each model was tested against a variety of different datasets, in an effort to produce a multitude of metrics by which to evaluate its success. Our primary method of evaluation stemmed from employing the model on our subset of manually labeled tweets, to

help show the veracity of our automatic labeling process.

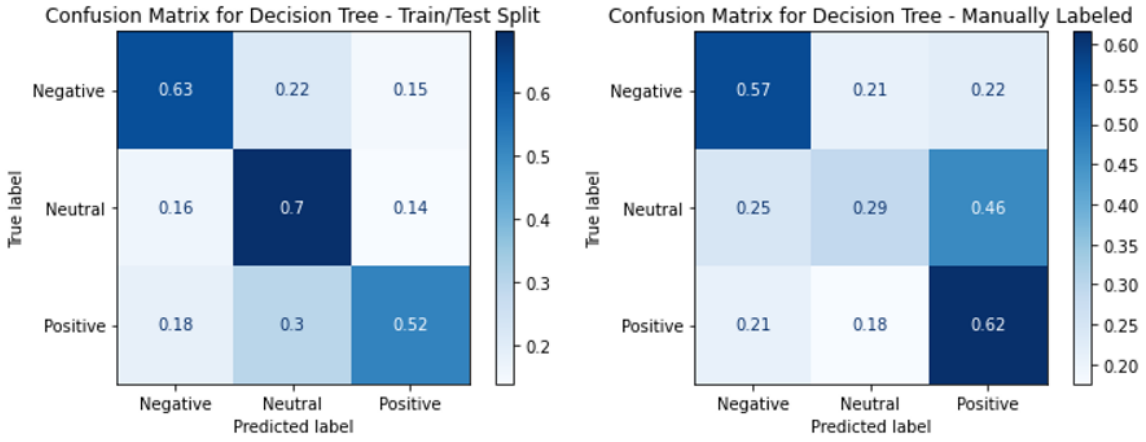
One of our approaches employed a simple decision tree which provided an insight into feature significance (Figure 1), and a rough estimate of accuracy. There was no hyperparameter configuration (leaf size, split size, max depth) that surpassed our manually-labeled test set accuracy of 51.6%, which we obtained by use of the out-of-box configuration. Our choice of using accuracy (in reference to error analysis) rather than precision/recall is primarily due to the fact that accuracy is well-defined in multiclass problems, whereas precision/recall is difficult to generalize. Though the accuracy is not particularly reassuring in this case, even this simple model did provide us with various different significant features that seemed to suggest our model is on the right track. As expected, the model has picked up on words related to social movements and unrest such as "police," "justice," and "enforcement."

Figure 1



Most important features as determined by the decision tree — as expected, most are politically charged with respect to BLM.

Figure 2



Confusion matrices for the decision tree; resulting class-accuracy for automatically labeled test set (left) and manually labeled test set (right)

Figure 2 displays two confusion matrices, each displaying a different measure of success in the model. The train/test split refers to the evaluation of the model on test data that comes directly from the automatic labeling procedure (it is important to remember that the labels generated from the automatic procedure will not be completely accurate). The second confusion matrix refers to the results obtained by use of the manually-labeled testing data, which in general will have accurate labels. Though both matrices seem to suggest that the model is functional regarding the classification of supportive and oppositional behaviors, the manually-labeled test data revealed the model's ineffectiveness at classifying neutral/irrelevant tweets.

We conjecture that the discrepancy between the accuracy regarding neutral/irrelevant statements is due primarily to another form of selection bias. There is likely an inherent difference between the automatically gathered neutral/irrelevant data (which was scraped randomly from Twitter), and the neutral/irrelevant data found within the manually labeled set. The manually labeled data contains control tweets that also originate from the #NoJusticeNoPeace and #BackTheBlue datasets (this was done in an attempt to include neutral, yet

political statements, within the neutral/irrelevant camp, i.e. "there is a protest happening in downtown Boston"). These neutral/irrelevant tweets are likely more political in nature, whereas the automatically collected neutral/irrelevant data likely constitutes something closer to random noise. The manually-labeled testing results falter because the model is unfamiliar with the more political, and yet neutral forms of expression. This problem may be corrected by the use of more data, or the incorporation of some manually labeled data within the training phase.

3.2 Logistic Regression

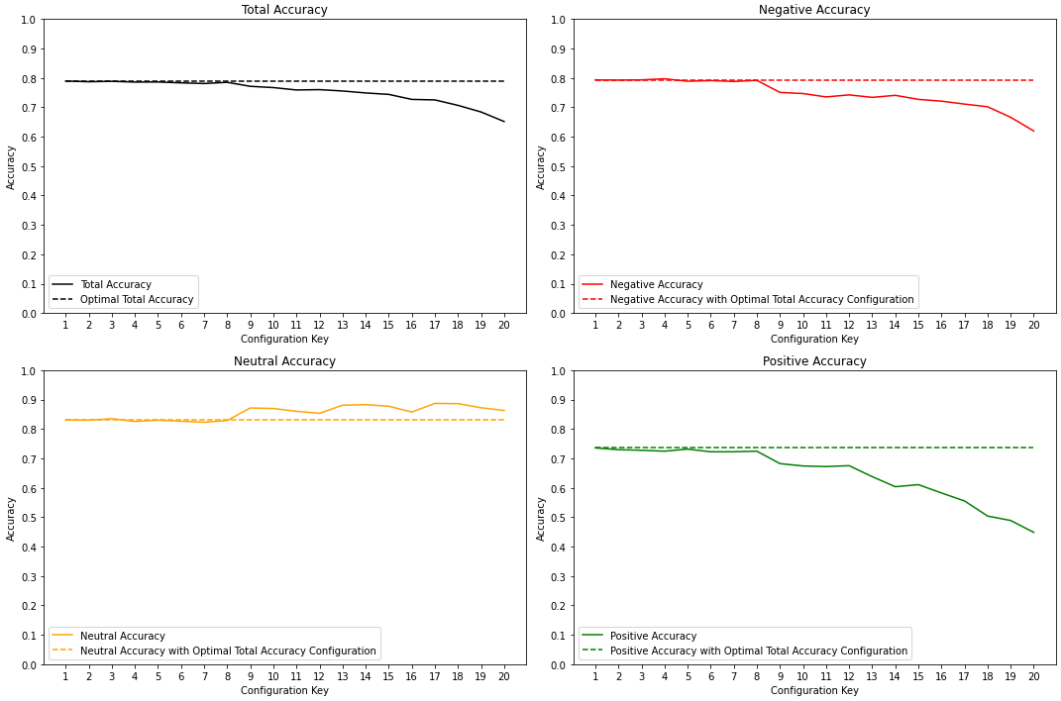
The next approach in model selection leveraged the fact that tweets have an inherent constraint of 280 characters. We hypothesized that this scarcity in text would lead to a greater discrepancy in how single words and n-grams would contribute to the model. This led to the decision to engineer features with a hybrid approach, where single words and n-grams would be modeled with distinct minimum frequency thresholds. Furthermore, the n-gram range hyperparameter was also tuned but little change was detected due to the minimum frequency threshold. As a result this parameter was set to 5 throughout all configurations. Our results indicated there was little variation in performance by differentiating the minimum frequency of single words and multiple words, and instead the most important factor consistently was a lower term frequency threshold.

Once the feature set was determined using the optimized vectorizer, the next step was to build out the logistic regression model. At this point, the out-of-the-box configuration actually performed better than the various hyperparameters we tested without any performance concerns, so hyperparameter tuning was not employed in the optimized model.

Again, accuracy was the performance metric measured here. Once the model was tuned in the training data split, the accuracy in the testing data split was 77.9%. Interestingly, both the negative and positive accuracy trended strongly with the total accuracy across all model configurations. However, the neutral accuracy was inversely correlated, so as the neutral accuracy increased both polarized sentiments decreased. These results are included in Figure 3, and the configurations are defined in Table 6 of the Appendix.

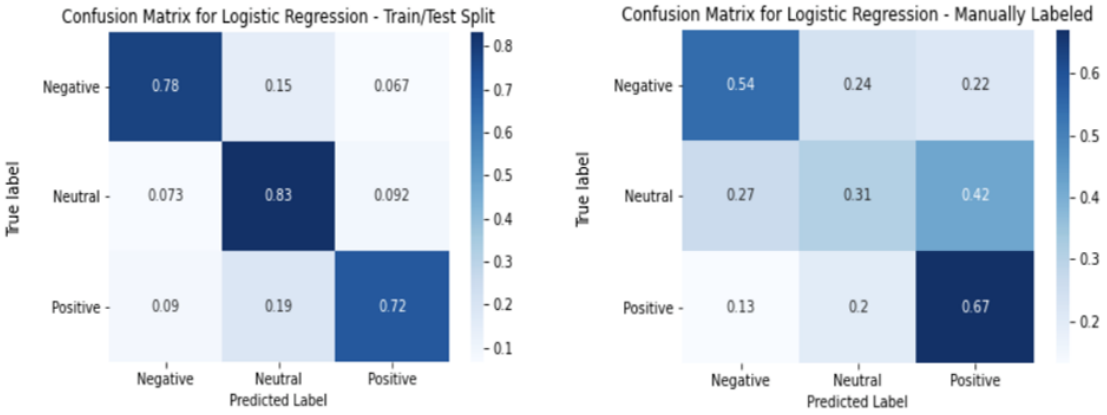
In addition to testing against the testing data split, we further tested our model against the manually labeled data using the same process as described in the previous section. The tuned model performed with a 50.1% overall accuracy. When broken down into the sentiment components, the positive accuracy was 67.1%, the negative accuracy was 54.5%, and the neutral accuracy was 31.0%. The confusion matrix for both the testing data in the automatic labeling approach and manually labeled approach are provided in Figure 4. The discrepancy between the neutral classes of each confusion matrix is again explained by the form of selection bias presented at the end of the previous section.

Figure 3



Performance as a function of data configuration: interestingly, neutral accuracy improves at the expense of the positive and negative classes

Figure 4



Confusion matrices for the logistic regression; resulting class-accuracy for automatically labeled test set (left) and manually labeled test set (right)

3.3 Deep Learning Model

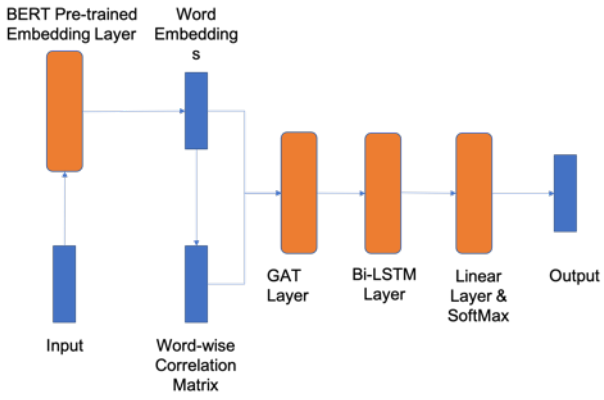
Deep Learning (DL) models have been widely successful in text classification problems, with architecture that is suited for preserving structure and meaning. We take advantage of the recent developments in transfer learning techniques within natural language processing – we use BERT (Devlin et al., 2019) to generate the pretrained embedding by which to encode the contextual information into each local position of a sentence. Because the pretrained embedding already contains features from other large-sized corpuses, we are able to leverage these features which may not be present within our data set.

Using this word embedding, we then generate a word-wise correlation matrix, the elements of which indicate the level of relevancy between any of two positions in a given sentence. This correlation matrix is treated as an adjacency matrix to be inputted into the graph attention layer (GAL) (Velicković et al., 2017) together with our pretrained word embeddings. In the GAL, the model learns an attention matrix that identifies the appropriate weight which should be assigned to each word. To give a concrete example, suppose we have an attention matrix of size $n \times n$, where n is the sentence length. The element i, j of the matrix will be large if the model thinks the relation between the context around position i of the sentence and the

context around position j of the sentence is of great importance in determining the label of the sentence. Visualizations of this process can be found in the attention maps section 6.1 of the Appendix . By incorporating the GAL, we are able to provide the model with flexibility in prioritizing which knowledge is best to learn. The output of the GAL will have the same dimension as the input of that layer. The difference is that, now each position of the word embedding is properly weighted so that the model knows which aspect of the sentence is most important in delivering signals with regard to the classification task.

We then pipe the attended word embedding into the Bidirectional LSTM layer (Hochreiter and Schmidhuber, 1997). This layer helps the model to understand the flow of information from the beginning of the sentence to the end of the sentence, as well as from the end to the beginning, due to its bidirectional nature. This layer offers more of a contextually-rich learning ability such that the model maintains an effective method of summarizing these hidden representations into a single lower-dimensional vector (i.e., the cell state vector) that compresses all information learned. Finally, we perform the standard scoring procedure in the last layer of our model, which takes in the cell state vector and outputs the probability regarding different labels. Figure 5 displays the process flow.

Figure 5



Process flow

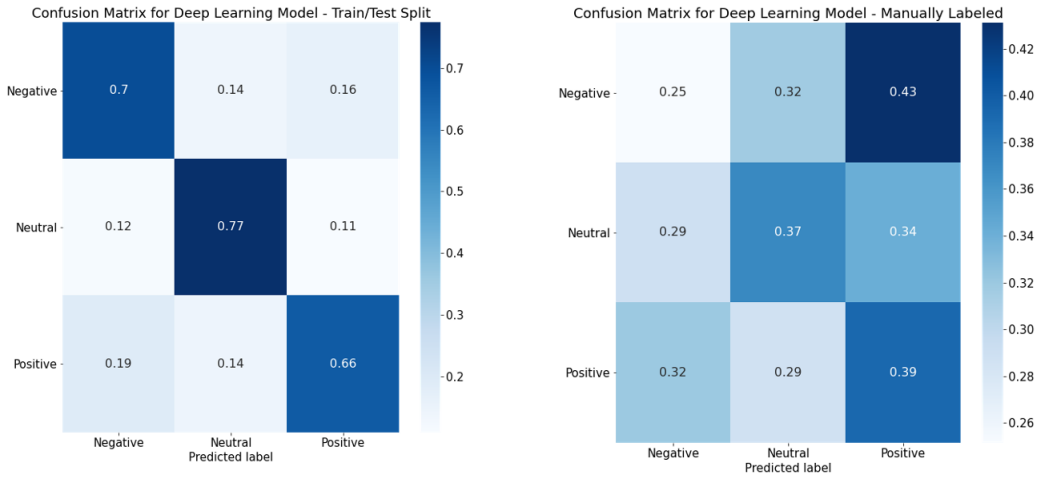
Overall, the design of the model allows us to sensitively capture the local signals by use

of the GAL, as well as efficiently incorporate local signals into global representations of the sentence via use of the Bi-LSTM.

We evaluate the result of the model with confusion matrix and accuracy score. The deep learning model achieved 72.6% overall accuracy on the automatically labeled dataset and 35.6% overall accuracy on the manually labeled dataset.

Figure 6 shows the confusion matrices on two datasets respectively. In terms of the automatically labeled data, we notice that the model is strongest in learning the neutral class and has a balanced learning ability on negative and positive classes. As for the manually labeled data, the model performs marginally better than a random guess. This contrasts that of the simple model, which still maintains adequate performance on the manually labeled dataset. One potential reason might be that the deep learning model is more sensitive on the shift in the underlying distribution of the data than the simple model.

Figure 6



Confusion matrices for the deep learning model; resulting class-accuracy for automatically labeled test set (left) and manually labeled test set (right)

It is a bit surprising that the deep learning model performs worse than the simple models that we proposed in the previous sections. However, after second thought, we consider it reasonable because the deep learning model puts strong emphasis on understanding the context,

which might be overkill given the short text data on Twitter.

The way that the deep learning model operates — blending local information with information from the contextual words into a hidden representation — might obscure the meaning of the local information by introducing too much contextual noise. Especially, with social media data, the writing is not always logical and organized, often containing large amount of contextual noise. On the other hand, our simple model, the logistic regression, used features such as n-grams which tend to deliver clear and strong local signal, hence yielding better performance overall. That said, we still consider the deep learning model an option for the political attitude classification, because of the rapid development in the field of natural language processing. More powerful techniques worth trying is coming into the toolkit in the foreseeable future.

4 Deployment

Having produced a model capable of classification concerning sentiment surrounding the BLM protests, such a model could be employed in various different circumstances, and for various different ends. One feasible application concerns the verification and tracking of collective sense-making processes (CSMP). CSMP are at the heart of public discourse, and represent the evolution and interaction of voices within a community, as they grapple with problems, solutions, and perceptions. Most importantly, CSMP often act as a precursor to social change, or more dramatic forms of social unrest. The advent of social media has allowed for greater investigation into CSMP, being both highly visible as well as accessible. Social media provides the data necessary to track social change in real time, thereby maintaining situational awareness in times of unease (Oh, 2015).

CSMP are at the intersection of multiple disciplines, including political and social science, as well as information science. Their implications stem further than the tracking of social unrest. CSMP provide us with a standard for reality, a place of shared meaning and existence. They are present whenever individuals collectively process and act upon information.

As one of many possible applications, we decided to employ our model in the study of

the CSMP surrounding perceptions of BLM concerning the related protests, and the 2020 Presidential Election. We employed the decision tree classifier and logistic regression model developed in the previous section and constructed linear regressions relating the perceived support of the protests to the total voting percentage for the now President-elect Biden.

We began by collecting around 300,000 tweets in total from three different hashtags during the same time period as before. We chose to use #NoJusticeNoPeace, #BlueLivesMatter, and #Protest for no other reason other than to gather a good spread of sentiment from our data set. The cleaning and preprocessing method is the same as was employed in the previous section, with the addendum that the tweets had to contain geographical data which could identify their state of residence. Users located outside of the United States were excluded.

Table 3

Hashtag	Total Tweets	Total Cleaned
#NoJusticeNoPeace	81664	15050
#BlueLivesMatter	99488	12982
#Protest	110522	19703

Total tweets collected before and after cleaning and removing duplicates, for the purpose of correlation analysis between support for BLM and support for Biden

From here, the set of usable tweets was classified by use of the decision tree and logistic regression models previously trained. The supportive and oppositional tweets were then grouped by state, and the ratio of positive sentiment tweets to the sum of positive and negative sentiment tweets was calculated. We preformed both ordinary least squares (OLS) as well as weighted least squares (WLS) analysis in order to evaluate the correlation between the data. The weighted least squares analysis was done by weighting each point (state) by the number of tweets associated with it.

We created a statistical test where our null hypothesis is that there is no correlation between support for Biden and support for the protests ($\beta_1 = 0$), and an alternative hypothesis where the relationship is positive ($\beta_1 > 0$). The null hypothesis was able to rejected at the 1% level for both the OLS LR model and the WLS DT model. The performance of both optimal

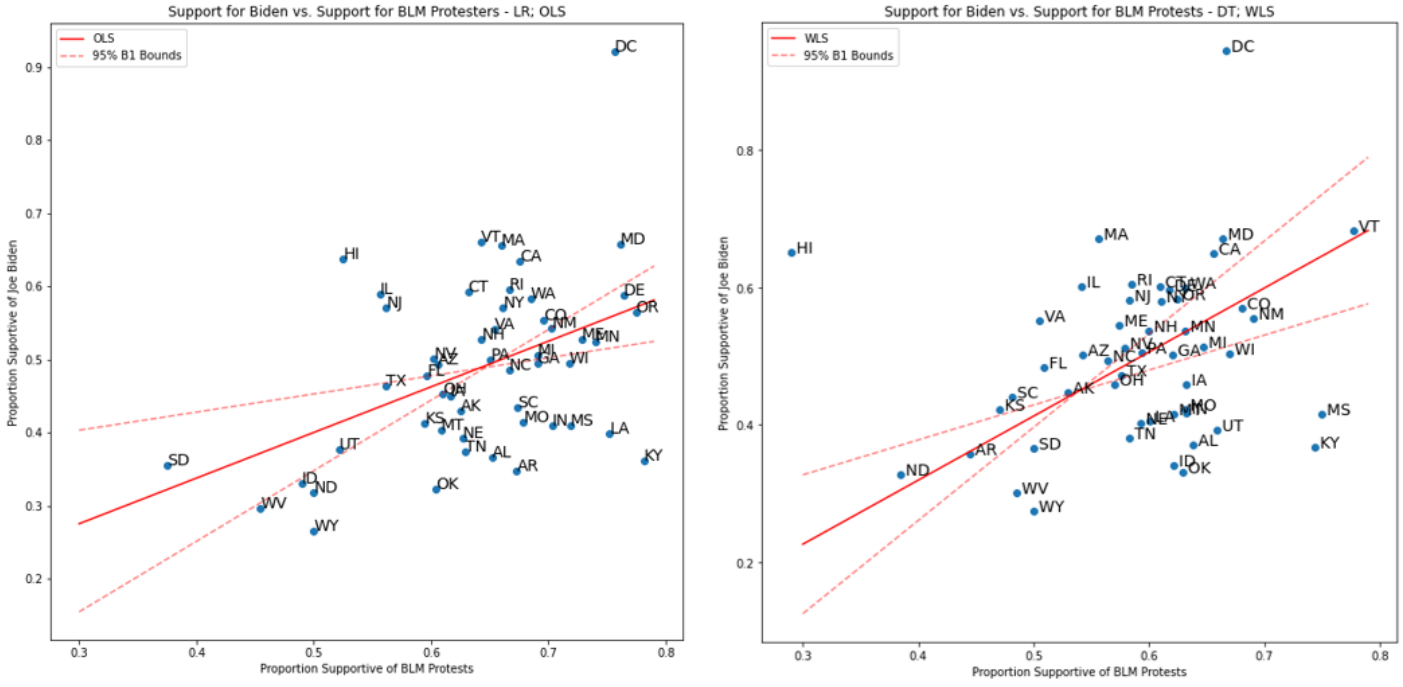
models are included in Figure 7 and the results of all four model performances are included in Table 4.

Table 4

Model	β_1	95% Confidence Interval	p-value
Decision Tree - OLS	0.3256	[-0.082, 0.733]	0.115
Decision Tree - WLS	0.9316	[0.508, 1.355]	0.001
Logistic Regression - OLS	0.6247	[0.248 , 0.966]	0.004
Logistic Regression - WLS	0.2791	[-0.131, 0.856]	0.146

Results from various least squares regressions. DT WLS and LR OLS both provided sufficient evidence to reject the null-hypothesis — there appears to be a correlation between support for BLM and support for President-elect Biden

Figure 7



The results of the best performing least squares regression analyses, with 95% confidence bounds regarding the fit. An ordinary least squares regression is displayed through use of the LR model (left), and similarly a weighted least squares regression with regards to the DT model is shown (right)

It is important to take these results with a grain-of-salt, and perhaps think of them as more of a proof-of-concept demonstration. We acknowledge that the weighting approach of the WLS model is not entirely characteristic of the problem at hand, as participation on Twitter certainly varies by geographic location. As such, stronger domain knowledge should be applied to this problem. Furthermore, the ideal approach for deployment of this correlation analysis would involve models of a higher caliber.[2]

4.1 Ethical Implications and Risks

Classification of political speech, especially when done in tandem with geographical information and personal identification information must be handled with care. Models such as these, regardless of application, should only be used at a global level in order to understand a community at large, to promote safety and welfare, and to foster a dialogue between opposing parties.

The overall goal of this model is to serve as a tool for future research. As noted, it's primary purpose involves the analysis of CSMP which occur whenever groups face collective struggle. As we have shown, the model may be useful in issue-based predictive analytics, whether that concerns presidential elections, public safety, or targeted community-based initiatives. The results of analyses that are based on this model are dependent on the model's classification accuracy, however, in using the model on a global/community level, there is no significant risk associated with false positives or negatives. Therefore the attainment of the model accuracy within this work suffices in most applications, though further model development may be necessary, depending on the application.

5 Conclusion

The development of political/social issue-based classification is useful for evaluating the fabric of a society's current structure. Our model seeks to discriminate between views regarding the BLM protests, though by use of the automatic labeling process, the outline of this work can be adapted to similar issue-based events. Our automatic labeling procedure was

[2]We decided against creating least squares regressions concerning the DL model due to the superior performance of the other models.

validated by use of a small subset of manually labeled data, which should always be done in the future. This helped to show the model's veracity.

A classification model such as this is useful for many reasons, though it primarily provides insight into social turbulence, and can serve as a method for predictive analytics. In the case of the most recent Presidential election, our model was able to show significant correlation between support for BLM protests and support for President-elect Biden.

As was mentioned, the automatic labeling process serves as a proxy for manual labeling. The benefits of such a process greatly out-weight the costs, provided that there is not significant risks associated with false positives and false negatives. These risks can be avoided by tackling problems at a global, rather than user-level, scale. This research was conducted using a fairly small set of data, with limited computational resources, and a limited time frame. The development of an industry-capable model of this sort (by using magnitudes more data) would likely provide even stronger results.

References

Badger, Emily. 2020. "How Trump's use of federal forces in cities differs from past presidents." *NY Times* . July 23 https://nyti.ms/2X6tbWL.

Davenport, Christian, Sarah A. Soule and David A. Armstrong. 2011. "Protesting while black? The differential policing of American activism, 1960 to 1990." *American Sociological Review* 76(1):152–178.

Davidson, Thomas, Dana Warmsley, Michael Macy and Ingmar Weber. 2017. "Automated hate speech detection and the problem of offensive language." *arXiv preprint arXiv:1703.04009* .

Devlin, Jacob, Ming Wei Chang, Kenton Lee and Kristina Toutanova. 2019. "BERT: Pre-training of deep bidirectional transformers for language understanding." *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1(Mlm):4171–4186.

Go, Alec, Richa Bhayani and Lei Huang. 2009. "Twitter sentiment classification using distant supervision." *CS224N project report, Stanford* 1(12):2009.

Hasan, Maryam, Emmanuel Agu and Elke Rundensteiner. 2014. Using hashtags as labels for supervised learning of emotions in twitter messages. In *ACM SIGKDD workshop on health informatics, New York, USA*.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9(8):1735–1780.

Ince, Jelani, Fabio Rojas and Clayton A Davis. 2017. "The social media response to Black Lives Matter: how Twitter users interact with Black Lives Matter through hashtag use." *Ethnic and racial studies* 40(11):1814–1830.

Mundt, Marcia, Karen Ross and Charla M Burnett. 2018. "Scaling social movements through social media: The case of black lives matter." *Social Media+ Society* 4(4):2056305118807911.

Oh, O., Eom C. Rao H. R. 2015. "Role of social media in social change: An analysis of collective sense making during the 2011 Egypt Revolution." *Information Systems Research* 26(1).

Reynolds-Stenson, Heidi. 2018. "Protesting the police: anti-police brutality claims as a predictor of police repression of protest." *Social Movement Studies* 17(1):48–63.

Velicković, Petar, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò and Yoshua Bengio. 2017. "Graph attention networks." *arXiv* pp. 1–12.

Violos, John, Konstantinos Tserpes, Iraklis Varlamis and Theodora Varvarigou. 2018. "Text Classification using the N-gram graph Representation model over high frequency data streams." *Frontiers in Applied Mathematics and Statistics* 4:41.

6 Appendix

Table 5: Hashtags

All Statistically Tested Hashtags		
#BlackLivesMatter	#BLM	#Protest
#Protesters	#ICantBreathe	#Justice
#HumanRights	#Activism	#Equality
#AntiRacism	#CriminalJusticeReform	#Racism
#SocialJustice	#GeorgeFloyd	#CivilRights
#SocialChange	#Riot	#Riots
#Antifa	#PoliceBrutality	#NoJusticeNoPeace
#Looting	#AllLivesMatter	#WhiteLivesMatter
#iMatter	#BrownLivesMatter	#BackTheBlue
#BlueLivesMatter		

Table 6: Configuration Definitions

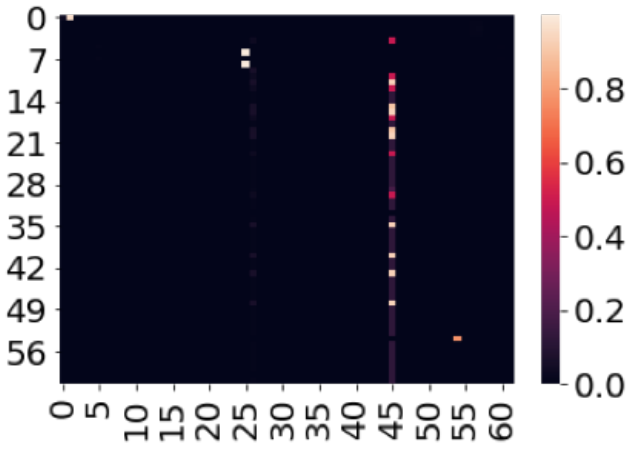
Key	Single Word min df	n-gram min df	Accuracy
1	5	5	0.789
2	5	10	0.787
3	5	20	0.786
4	5	50	0.786
5	10	5	0.786
6	10	10	0.783
7	10	20	0.781
8	10	50	0.785
9	50	5	0.77
10	50	10	0.767
11	50	20	0.759
12	50	50	0.76
13	100	5	0.755
14	100	10	0.749
15	100	20	0.744
16	100	50	0.727
17	200	5	0.725
18	200	10	0.707
19	200	20	0.684
20	200	50	0.652

6.1 Attention Maps

In the following, we randomly select for display the attention map of three sentences that the model correctly labeled. Each map is of size $n \times n$, where n is the sentence length of the corresponding post. We will provide interpretations on these maps.

In Figure 8, for the tweet "Next Saturday July 25th at 12pm we will have another protest", we observe that the model pays particular attention on the correlation between the location information around 45th position and many other positions in the sentence. The 45th position is the start of the phrase "another protest". This might suggest that the model captures particularly strong signals on this phrase and relies on it to classify this sentence.

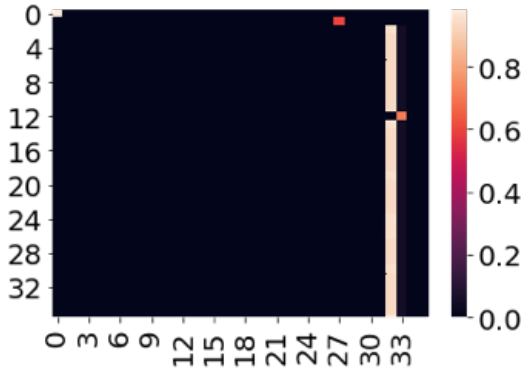
Figure 8



Tweet Contents: ["Next Saturday July 25th at 12pm we will have another protest"] Predicted Label: Supportive; True Label: Supportive

The attention map in Figure 9 is for the tweet "Baltimore County Maryland Patriots". The model pays particularly strong attention to the local information around 32nd position, which is the start of the word "Patriots". In many tweets, this word is frequently used to refer to users who are not supporting the protest. The model captures this signal well. Notice that, in this tweet, only the word "Patriots" potentially conveys political ideology, and the rest of the words are all location words that are irrelevant to any political attitude. The model successfully discerns the useful signal (i.e., "Patriots") from the noise (i.e., Baltimore County Maryland). This serves as proof that the attention layer helps the model focus on the portion of the sentence which is most relevant.

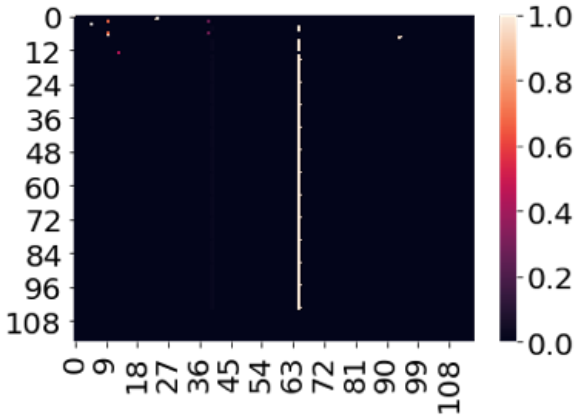
Figure 9



Tweet Contents: ["Baltimore County Maryland Patriots"] Predicted Label: Unsupportive; True Label: Unsupportive

In Figure 10, the model pays extra attention to the correlation between the local information around 65th position, which is the start of the phrase "arent protect". This tells us the model learns that the direction of the hidden representation at the 65th position is highly similar to the hidden representation of the whole sentence. To put it in a more mathematically rigorous way, the cosine similarity between these two vectors is high. This interpretation gives us an intuitive understanding of how the model learns to assess which part of the sentence is important.

Figure 10



Tweet Contents: ["An ANIMAL IS AN ANIMAL can't change it Police officer arent protect where be NYCMayor Hiding like a coward as usual"] Predicted Label: Supportive; True Label: Supportive

To summarize, these three case studies on sample attention maps were generated by use of the deep learning model, which serves to improve the interpretability of the model by giving an intuitive understanding as to how the black-box deep learning model attends to different contextual information.