Analyzing Public's Reaction towards Black Lives Matter Campaign using Machine learning-based Approach through Spark

Rio Rizki Aryanto¹, Dhomas Hatta Fudholi²
¹Master Program in Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia
²Department of Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia

E-mail: ¹rio.aryanto@students.uii.ac.id, ²hatta.fudholi@uii.ac.id

Abstract. Since the end of May 2020, there is a massive wave of protest in the United States addressed to the government regarding the case of police violence towards black people. On May 25, 2020, George Floyd, a 46-years-old black American man was killed during an arrest for allegedly using a counterfeit bill in Minneapolis. This study analyzes the public’s reactions towards the Black Lives Matter campaign using a supervised machine learning-based approach. The proposed model uses logistic regression with word vectors as its feature. The model classifies the public’s reaction represented by tweets and crawls using Spark Streaming into sentiment class, i.e. positive, negative and neutral. In addition, named entity recognition analysis was also conducted in this study. The aim is to find who else besides George Floyd whose rights have been fought for by the public. SparkNLP is used to build the logistic regressions model, sentiment analysis and named entity recognition. This study finds that most of the public tweets had a negative tone addressed to the Floyd incident specifically and to the violence towards black people in general. Another finding is that the campaign not only fought for George Floyd, but also fought for the other victims like Rayshard Brooks, Dominique Fells and Eric Garner.

Keywords: social media analytics, Spark, machine learning, sentiment analysis, black live matter

1. Introduction
Incident of George Floyd's death, a 46-years-old black American man on May 25, 2020 sparked a huge wave of protests in the United States. Even today, protest is still often found and reported in the media as the Black Lives Matter campaign. On May 25, 2020 George Floyds was killed during an arrest for allegedly using a counterfeit bill in Minneapolis. A white police officer named Derek Chauvin was knelt on Floyd’s neck for almost 8 minutes. In the end, Floyd’s life could not be saved anymore. The day after the incident, the videos made by witnesses and footage from security cameras became public. As a result, anger emerged from the public regarding police violence towards black people and triggered a move known as Black Lives Matter campaign.

The incident is one of many cases that show the inequality received by black people, especially African-American in the United States. This inequality has indeed become an old case and has even become public information for the people of the United States. In the past, there have been countless
cases related to police violence towards African-American men. As today, the term white privilege was born, the privileges received by white people in the United States.

According to FBI and US Census Bureau data for 2018 about drug abuse arrest, black people had a higher arrest rate compared to its population rate. The opposite happened to white people and other ethnic where these two had a lower arrest rate compared to its population rate. The data said about 750 for every 100,000 African-Americans were detained for substance abuse reasons. While for the white people, it is just about 350 for every 100,000 people. The United States’ national survey of drug use also stated that there are equal numbers of use among white people and African-Americans, however, blacks are more likely to be arrested. This is also the evidence of the inequality received by black people.

African-American are also detained 5 times more than white people and 2 times more than Hispanic-American. In 2018, 13% of the United States’ population were African-American, but one-third of prison inmates in the United States’ prison were also African-Americans. The majority of the United States’ population were white people with about 60%. These people only contribute 30% of prison inmates. That means, for every 100,000 African-American citizens, 1,000 people were imprisoned. By the same token, only 200 white people were imprisoned. All of these statistics were reported by the US Census Bureau in 2018.

The other famous black people inequality was about shootings by police officers. Still according to the US Census Bureau and supported by Statista.com data in 2019, of the 1,004 fatal shooting cases that occurred in the United States, 36.8% happened to the white people, 23.4% African-Americans, 15.7% Hispanic and the rest for the other ethnic. African-American’s fatal shooting percentage was higher than its population percentage in the United States. This only happened for black people whereas in others it happened the opposite. Fatal shooting itself is a shooting by a police officer that leads to death of the victim or suspect.

Above data just shows us that there is inequality received by black people in the United States which this inequality usually came from the police department. Recent cases that happened to George Floyd arouse the enthusiasm of African-American in the United States in particular and black people around the world in general together advocate for equality of black people in society. This study analyzed the public reaction towards Black Lives Matter campaign in the United States that was conveyed through social media Twitter. As today, the protest did not only happen on the street, but also through social media. The public reactions would be classified into sentiment classes while for deeper analysis, named entity recognition would be used as an additional.

2. Related Works

Machine learning-based approaches for analysis sentiment have been done many times in the past. Before machine learning-based approaches, the lexicon-based approach was known and widely used in many researches. Lexicon-based approaches make use of dictionaries or corpus to detect the sentiment. The quality of sentiment resulting from lexicon-based approach is very dependent on its dictionary or corpus. If the dictionary or corpus used is complete enough and is able to cover the scope of the problem raised, it is likely that the results of the sentiment classification will also be good.

In the previous work, one example of lexicon-based approach was done using an emoticon corpus [2]. Besides emoticon corpus, the researchers who used this approach also usually used N-Gram or hashtags as done on [3]. This study uses them to label the training data. Lexicon-based approach might feel old, but it could be used for problems or topics that are more general. Something that could be a challenge if using a machine learning-based approach. The other benefit using lexicon-based approach is it is easier to implement compared to machine learning-based approaches. On the other side, lexicon-based approaches also came with some drawbacks. This approach cannot provide the correct sentiment label for different subjects and usually is unable to capture the rich relational structure of the text. Therefore, this approach is mostly used as a complement or preprocessing stage before being complimented by a machine learning-based-based approach.

Lexicon-based approach’s challenges could be handled using a machine learning-based approach. Basically, a machine learning-based approach would learn the data and its relationship so it is able to
address data that may not have been studied before. Something that could not be done using lexicon based approach. To capture the rich relational structure of the lexicon, researchers used the word vector in their sentiment classifier [4]. Some studies also used supervised machine learning algorithms for its sentiment classifier like logistic regressions [5]. In their study, [5] use logistic regressions as base learner and majority vote as ensemble method. They also integrated their sentiment classifier with Hadoop.

A machine learning-based approach is indeed very closely related to very large amounts of data, where the application requires a lot of effort. The greater the data used, the heavier the workload of the machine which can result in the longer the processing time. Implementation of big data platforms could be the solution to this problem. One of the most popular big data analysis platforms is Spark. Spark is widely used by many researchers due to its ability in fault tolerance and real-time data processing. [6] use Spark for their real-time sentiment analysis for specific advertising targets. They used social media Twitter data for the study. [7] also conducted a study about sentiment analysis using social media Twitter data. The attempt of the study was to make financial decisions such as stock market predictions.

Spark implementations in big data analysis especially social media data were quite popular. [10] uses Spark streaming to stream Twitter data about job vacancies and then classified it into job categories. Spark was used to facilitate the effective job search in their system. Their work succeeded in implementing real-time analysis of filtering job vacancies among million streaming tweets which led to the improvement of the job search’s effectiveness. Spark is also capable of conducting sentiment analysis like the one done in [11]. Their study implements Spark in a comparison analysis of sentiment analysis classifiers such as naive bayes, logistic regressions and decision tree. Another Spark implementation is real-time sentiment analysis using Spark streaming by [12], who in the study also conducted fake-account detection analysis. His work found that the classification performances of the system for real-time modes is about 80.93%. The proposed model in this study uses Spark streaming to retrieve the Twitter streaming data and analyze it using SparkNLP. Instead of fake-account detection analysis similar to [12], this study adds named recognition analysis to extract deeper insight from the tweets.

![Figure 1. Workflow of the Proposed System](image)

3. Methods
Focus of this study is to analyze the public reactions towards Black Live Matter campaign that happened in the United States. Similar to [6] and [7], this study would make use of real-time features in Spark as its main architecture. Real-time process happened in a crawling process which uses Spark streaming and Twitter streaming API to collect tweet data related to predefined keywords. While for sentiment and named entity recognition analysis, it was done every 15 minutes which the schedule was run using the help of crontab.

This study aims to analyze the public reaction that was conveyed through social media Twitter. Public reactions would be classified into sentiment classes using logistic regression method as a base learner for the classifier. The implementation of several analyses in this study were also done using Spark
models such as Spark MLlib for sentiment analysis classification and SparkNLP for its named entity recognition. Figure 1 shows how the workflow of the proposed system which contains data collections, preprocessing, sentiment analysis and named entity recognition. The detail of each component is explained below.

3.1. Data collection
Data were obtained from Twitter that have been crawled in the period between June 4 to June 26, 2020. To note that, in this study there is a small missing time series data gap on June 7 due to technical problems. The system started with a crawling process using Twitter streaming API accessed from Python Tweepy and Spark streaming features. In order to obtain tweets that could represent public reaction toward Black Lives Matter campaign, the keywords filter features in Twitter streaming API is used. Tweets that would be taken are tweets that contain one of these keywords i.e George Floyd, Black Lives Matter or Racism. Using Spark streaming features, real-time tweets that have been collected will be saved in CSV files every 5 seconds. This CSV file will later be used in the sentiment analysis process and named entity recognition.

Instead of using historical data, this study uses stream data that is collected from 4th of June 2020. The data collecting starting date is just a few days after George Floyd’s murder happened (on May 25). The decision to use stream data has been triggered by events that happened in early June such as Trump’s threat to deploy military on protestant on June 1 and the charge of three former Minneapolis police officers on June 3 following Chauvin’s charge on May 29. This study was interested in analyzing public responses toward Black Live Matter campaigns after some movements made by the United States’ government. Was the public anger subsided or it did not affect anything at all? Using streaming data, we are able to capture every public response through Twitter after that time and also cover a longer period of time monitoring.

Spark streaming itself gives some benefits on its implementation. First, it recovers fast from failures as Spark streaming is popular as a fault-tolerance streaming system. Spark streaming also could handle heavy computation with its load balancing schema, which it would balance all the jobs across its workers. Interestingly, Spark streaming is also capable of doing interactive queries on streaming data on its worker memory. Moreover, the streaming data are able to be built into a machine learning model using rich libraries such as Spark MLlib.

As mentioned before, the tweets are expected to be able to represent the current condition in the United States. To accommodate this purpose, additional filters will be made before the tweets stored in CSV files. Tweets to be used must be in English and contain one of the following 3 keywords, namely “George Floyds”, “Black Lives Matter” or “blacklivematter”. In the end, a total of 1,299,181 tweets have been collected during the period of monitoring.

3.2. Preprocessing
Before analysis, the tweets must be preprocessed first. In this step, there are 3 processes done i.e tokenization, removing stop words and word vector formations. Spark was also used in the preprocessing stages. Tokenization is the process of separating single unit text into words. After that, the words which are categorized as stop words such as “this”, “that”, “was”, “were”, “during”, etc would be removed. Lastly, word vector formation was done. Word vectorization itself is a process to convert filtered words into numerical data. Spark’s Word2Vec model was used to create the word vector which in its implementation the vectors were built using the skip-gram model.

It is generally known that logistic regression only accepts numeric input as its feature, while our input text has been converted into a matrix of word vectors. However, we could still use logistic regression as our model since in Spark MLlib, the logistic regression model would automatically convert the matrix of word vectors into a vector by averaging all the values in the sequence. So, instead of using a matrix of word vectors, Spark’s logistic regression used a vector representation of it as its feature.
3.3. Sentiment analysis
Our logistic regressions classifier was built and trained using Sentiment140’s public dataset obtained from [8], which contains two different data i.e data-train and data-test. Our classifier was trained using the data-test, because it has three different sentiment classes and its suit best for this study’s case. Actually, Sentiment140’s data-train had more records than the data-test one, however, it only has two different sentiment classes.

Basically, logistic regression was used for binary classification problems. In Spark MLlib, binary logistic regression can be generalized into multinomial logistic regression and predict multiclass classification. In binary classification, to predict whether an input is positive or negative, the default threshold of probability i.e 0.5 was used. Probability above 0.5 would be classified as positive or negative otherwise. On the other side, there are two options that could be used for multiclass classification problems in Spark, whether using thresholds or depending only on each class probability returned from the multinomial logistic regression model. Using thresholds, Spark would return a class which has both of the highest probability and its value is above the threshold. Otherwise, Spark would return a class with the highest probability as a predicted class and not consider any thresholds. In this study, we did not define any thresholds for the model.

One-vs-one approach was used in multiclass classification implementation on Spark. Technically, Spark would regress each class against the pivot class using binary logistic regression. By default, Spark would choose the first class with label 0 as a pivot class. For example, for K possible outcomes or class, Spark would have a multinomial logistic regression model, which contains K-1 binary logistic regression. The class with the highest probability from the multinomial logistic regression model would be chosen as the predicted class.

In Figure 1, we can also see that beside sentiment analysis, this study also adds additional analysis using named entity recognition. Only one entity type was considered in this study i.e person entity. This additional analysis was done to obtain deeper insight about Black Lives Matter campaign.

4. Result and discussion
Most of the public tweets found between June 4 to June 26, 2020 had a negative tone, as seen in Figure 2. The gap with other sentiments is also quite high with almost 5 times higher. A total of 73% of the tweets had negative sentiment while for both of positive and neutral sentiment, it has similar percentage with each 14% and 12.8%.

![Figure 2. Proportion Sentiment of Public Tweets](image)

During the period of monitoring, the number of negative tweets always surpasses the number of both positive and neutral tweets (see Figure 3). It seems there is a negative conversation happening on Twitter related to Black Lives Matter campaign. This might indicate that the public's rage had happened for a long period of time. They massively expressed their opinion through the Twitter platform with a negative tone.
Peak of negative tweets happened between June 13 to June 16, 2020. The blue line seen in Figure 3 had a slope up around that time. The peak might have been triggered by the incident that happened to Rayshard Brook, a 27-year-old African American man who was fatally shot by Garret Rolfe, an Atlanta Police Department (APD) officer on the night of June 12, 2020. This incident further exacerbated the situation and increased the public’s anger level. After June 16, 2020 the number of negative tweets started to decrease back to before Brook’s incident. On the other side, both positive and neutral sentiment had quite a stable number of tweets almost all the time.

Since most of the tweets had a negative sentiment, it could be misleading about the carried message, whether their anger addressed either to Floyd’s incident or Black Lives Matter campaign. There is a possibility that several segments of society do not support the Black Lives Matter campaign. This crucial question could be answered using wordcloud analysis. The wordcloud in Figure 4 was formed using a collection of words in negative tweets.

Words like killed, hung, died, cop, police and etc, that are highly related to Floyd's incident have been found. It seems that the public's anger has been addressed to Floyd’s incident rather than the message of Black Lives Matter campaign. The anger kept rising as the similar incident happened in close proximity to different people. Since the message of Black Lives Matter campaign was not clearly found in negative tweets, it might be carried on the positive tweets. On the positive tweets’ wordcloud as seen in Figure 5, the words like police reform and justice act were found. Here, the message of Black Lives Matter started to be seen. There is an expectation from the public that there would be changes especially in the United States’ police department. They demand reform within the police department and expect there would be equality in the society for black people.
This study also conducted named entity recognition analysis which aims to find deeper insight related to Black Lives Matter campaign. We only consider one entity in this analysis i.e. person entity. Black Lives Matter campaign might have been triggered by Floyd’s incident, however, our previous analysis found that the rage of the public just rose in the middle of June due to the Rayshard Brook’s incident. There is possibility of other similar cases which are also being fought for through this campaign. Please bear in mind that the word coloring both in negative (Figure 4) and positive wordcloud (Figure 5) did not represent classes or anything, it was just for beautification.

Besides George Floyd and Rayshard Brooks, the result of named entity recognition in Figure 6 also shows the other names like Dominique Fells and Eric Garner. These two were also the victims of the violence towards black people in the United States. Dominique Fells, a Philadelphia transgender woman, have been murdered and the body was found on the banks of the Schuykull River in the second week of June 2020. According to CBS Philadelphia, the authorities believe that 36-years-old Akhenaton Jones is responsible for Fells’ murder. This case might not relate specifically to police violence however, it is still related to the violence towards black people. It also happened around the same time to Brooks’ incident. Fair to say, that the peak of negative sentiment found during the period was also triggered by Fells’ murder.

The other name found was Eric Garner, who on July 17, 2014 died due to prohibited chokehold by a New York Police Department (NYPD) Daniel Pantaleo while arresting him. The incident took place in
the New York City borough of Staten Island. Although it happened about 6 years ago, the public still remembered about the incident since there were no significant changes happening in the police department afterwards. Even on Jan 15, 2015, it concluded that Pantaleo’s chokehold was administered and recommended charges, according to court documents. The public is back to fight for what they have fought before 6 years ago.

In the end, Black Live Matter campaign was not only about George Floyd but also the other victims such as Rayshard Brook, Dominique Fells and even Eric Garner. It is not only limited to cases about police violence, but about justice and strives for degree equality of black people among the society. The public tries to address this issue to the government of the United Stated and demand a reformation within the police department as their first step.

5. Conclusions

Based on the results of the study, we found that there is a public’s rage emerging due to Floyd’s incident which triggered a huge movement called Black Lives Matter campaign. About 73.1% of public tweets contain a negative tone, with most of it being addressed to Floyd’s death incident. Another finding is about Black Lives Matter campaign’s message. It could be found in the collection of positive tweets which some words such as police reform and justice act exist. Public expects a reformation within the United States’ police department and hopes there would be equality for black people among the society.

The proposed model used Spark as its main architecture. Spark streaming was used together with Twitter Streaming API to crawl the data while Spark MLlib and SparkNLP were used for sentiment classification and named entity recognition analysis. Even though the system uses a simple algorithm such as logistic regression, the result of the analysis gives us quite an understanding of the public's reaction towards Black Lives Matter campaign. Named entity recognition also gives a deeper insight about what happened and what is being fought for through the campaign.

References

[1] Prabhat A and Khullar V 2017 Sentiment classification on big data using Naïve bayes and logistic regression Int. Conf. on Computer Communication and Informatics (ICCCI) (India: Coimbatore) pp 1–5
[2] Read J 2005 Using emoticons to reduce dependency in machine learning techniques for sentiment classification Proc. of the ACL Student Research Workshop (ACLstudent ’05) (USA: Association for Computational Linguistics) pp 43–8
[3] Zhang L, Ghosh R, Dekhil M, Husu M and Liu B 2011 Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis (California: HP Laboratories)
[4] Mass A L, Daly R E, Pham P T, Huang D, Ng A Y and Potts C 2011 Learning Word Vectors for Sentiment Analysis Proc. of the 49th Annual Meeting of the Association for Computational Linguistics (Portland: Oregon) pp 142–50
[5] Lin J and Kolcz K 2012 Large-scale machine learning at twitter Proc. of the 2012 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD ’12) (New York: Association for Computing Machinery) pp 793–804
[6] Nair L R, Shetty S D and Shetty S D 2017 Streaming Big Data Analysis for Real-Time Sentiment based Targeted Advertising Int. J. of Electrical and Computer Engineering 7 pp 402–7
[7] Das S, Behera R K, Kumar M and Rath S K 2018 Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction Procedia Computer Science 132 pp 956–64
[8] Go A, Bhayani R and Huang L 2012 Twitter Sentiment Classification using Distant Supervision Retrieved from http://help.sentiment140.com/for-students
[9] F.F da silva N, Hruschka E R and Hruschka Jr E R Tweet sentiment analysis with classifier ensembles Decision Support System 66 pp 170–9
[10] Nair L R and Shetty DR S D 2015 Streaming Twitter Data Analysis Using Spark for Effective Job Search J. of Theoretical and Applied Information Technology 80 pp 349–53
[11] Baltas A, Kanavos A and Tsakalidis A K 2017 An Apache Spark Implementation for Sentiment
Analysis on Twitter Data Algorithmic Aspects of Cloud Computing ALGO CLOUD 2016
Lecture Notes in Computer Sci. 10230 ed Sellis T and Oikonomou K pp 15–25
[12] Kılınc D 2019 A Spark-based Big Data Analysis Framework for Real-time Sentiment Prediction
on Streaming Data J. of Software: Practice and Experience 49 pp 1352–64
[13] BBC News 2020 George Floyd: Tiga 'perlakuan tidak adil' yang dialami warga kulit hitam di
Amerika Serikat BBC News Retrieved from https://www.bbc.com/indonesia/dunia-52902071
[14] BBC News 2020 Kematian George Floyd: Bagaimana rasanya menjadi warga kulit hitam di AS
BBC News Retrieved from https://www.bbc.com/indonesia/dunia-52917132