A Comparative Analysis of Semi-Supervised and Self-Supervised Classification for Labeling Tweets about Police Brutality

Elaine Okanyene Nsoesie
Department of Global Health
School of Public Health
Boston University USA
onelaine@bu.edu

Wuraola Fisayo Oyewusi
Department of Research and Innovation
Data Science Nigeria
wuraola@datasciencenigeria.ai

Opeyemi Osakuade
Department of Research and Innovation
Data Science Nigeria
osakuade@datasciencenigeria.ai

Olubayo Adekanmbi
Department of Research and Innovation
Data Science Nigeria
olubayo@datasciencenigeria.ai

Abstract

Social media has proven to be influential in social justice advocacy. In 2020, thousands of Nigerians and allies used the EndSARS hashtag to protest police brutality in Nigeria. In this work, we aim to understand the conversation associated with the EndSARS hashtag by comparing the outcome of semi-supervised and self-supervised machine learning classification algorithms for the automatic labeling of tweets. The self-supervised, zero-shot learning algorithm had the best performance for automatic tweet labeling with average weighted recall of 0.73, compared to cosine similarity with TF-IDF(0.71), cosine similarity with universal sentence encoder (0.58) and Jaccard Similarity(0.222). The major topics of discussion included complaints about police brutality events, Lekki Massacre, activism, media coverage, lack of response from public figures and questions about moving forward. The pretrained predictive models for automatic tweet labelling will be made publicly available.[1].

1 Background

1.1 Police brutality and EndSARS

In recent years, there have been a number of protests on the Internet and in-person demonstrations against police brutality around the world. Organizations such as, Amnesty International have written about police brutality and human right violations in Hong Kong, Mexico, United States, Nigeria, and other regions.[2] In early October 2020, a series of reports about the shooting and killing of young men by the Special Anti-Robbery Squad (SARS) – a former branch of the Nigerian police force - led to public protests against police brutality in Nigeria. Nigerians and allies used the hashtag #EndSARS – a reference to a decentralized social movement against police brutality - on social media platforms to speak against these events and demand change. The SARS has a long record of abuses including allegations of profiling and attacking young Nigerians, mostly males, based on their fashion choices, tattoos, and hairstyles. They were also known to mount illegal roadblocks, conduct unwarranted

[1] https://git.io/J11wz

[2] https://www.amnesty.org/en/documents/afr44/4868/2016/en/

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

checks and searches, arrest and detain without warrant or trial, rape women, and extort young male Nigerians for driving exotic vehicles and using laptops and iPhones[3].

The EndSARS social movement against police brutality started in 2017 as a Twitter campaign using the hashtag #EndSARS to demand the disbanding of the unit by the Nigerian government. After experiencing a revitalization in October 2020 following more revelations of the abuses of the unit, mass demonstrations occurred throughout the major cities of Nigeria, accompanied by vociferous outrage on social media platforms. On the night of 20 October 2020, at about 6:50 p.m, members of the Nigerian Army were reported [4]to have opened fire on peaceful EndSARS protesters at the Lekki toll gate in Lagos State, Nigeria.Amnesty International stated that[5] at least 12 protesters were killed during the shooting, however it is said the number is definitely higher. There have been complaints about missing persons and some were confirmed to be hospitalized and in critical condition.

Twitter was widely used by citizens during this protest for sharing information covering personal opinions, events, and organization of support services, such as the provision of legal aid. Citizens also acted as observers and created useful situation reports. However, labeling these data to better understand the nuances of the conversation that emerged during this period is hard. This is because short-text classification can be a challenging task, due to the sparsity and high dimensionality of the feature space, and difficulty of separating valuable information from random tweets in the vast number of tweets created during this period. We apply techniques from data mining and machine learning to this task.

In this paper, we compare semi-supervised and self-supervised machine learning classification for automated labeling of the EndSARS tweets gathered during the period: 1st of October, 2020 to 31st December, 2020

1.2 Supervised, Semi-Supervised and Supervised Machine Learning

Classification is one of the most useful methods of deriving insights from text data. Supervised machine learning methods have realized remarkable performance for classification tasks utilizing labeled data. However, labeled tweets are usually expensive to obtain, especially when they are large, thus, not always achievable[8].When annotated data are unavailable, unsupervised machine learning techniques are used to observe the relationship between features by relying on the similarities among the data or on probabilistic approaches[5] When supervised and unsupervised methods are combined, a technique called semi-supervised learning is created, which can be applied to datasets with small amounts of labeled data and large amounts of unlabelled data.However, semi-supervised learning techniques all rely on active learning and pseudo labels, which are inadequate or computationally expensive[8]. Self-supervised learning has emerged as an important training paradigm for learning model parameters that are more generalizable and yield better representations for many downstream tasks. It typically involves learning through labels that come naturally with data, for example, words in natural language. Self-supervised tasks typically pose a supervised learning problem that can benefit from lots of naturally available data and enable pre-training of model parameters that act as useful prior to supervised fine-tuning[4]

In this paper, we use unsupervised learning to cluster and perform topic modeling on the tweets to select label topics and present a comparison between semi-supervised and self-supervised classification for EndSARS tweets using semantic text similarity and Zero-shot classifier respectively.

1.3 Clustering and Text Similarity

1.3.1 Topic modeling

Topic modeling is an essential algorithm used in extracting high-quality information from a large amount of unstructured text such as tweets, using computational methods and techniques.The basic idea behind topic models is to treat the documents as mixtures of topics and each topic is viewed as a probability distribution of the word.Latent Dirichlet Allocation (LDA) is a topic modeling algorithm widely used by researchers to extract topics from large text data. Blei et al. first proposed LDA as a

[3] https://en.wikipedia.org/wiki/End_SAR

[4] Paquette, Danielle. "Why are people talking about Nigeria and #EndSARS?". Washington Post. ISSN 0190-8286. Archived from the original on 10 February 2021. Retrieved 24 October 2020

[5] https://en.wikipedia.org/wiki/2020_Lekki_shooting

topic discovery graphical model in 2003 with the basic idea that documents exhibit multiple topics[2]. Examples of LDA usage in analyzing tweets includes applications in health[6] [3] and other domains [7]

1.3.2 Semantic text similarity

Generally, Semantic similarity is a metric of the conceptual distance between two terms, based on the closeness of their meanings[1]. Cosine similarity measures the similarity between two vectors by comparing the cosine angle between the two vectors, this determines whether two vectors are pointing in roughly the same direction. Jaccard similarity or intersection over union measures the similarity by dividing the size of intersection by the size of the union of two sets.

1.3.3 Zero-shot Classification

In the Zero-shot learning (ZSL) setting, we train a classifier from labeled training examples from seen classes and learn a mapping from input feature space to semantic embedding space. ZSL aims to classify class labels, which were never exposed during the training pipeline.

2 Methodology

2.1 Data Collection

The analysis included data collection, data cleaning and preprocessing, exploratory analysis, and labelling

Figure 1 shows the end to end methodology for automatic labelling of tweets related to police brutality

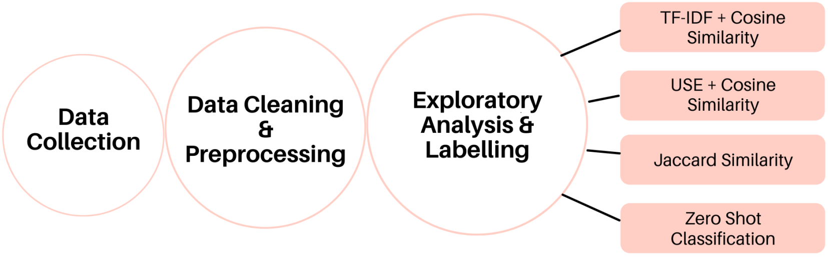
Data Collection
Data Cleaning & Preprocessing
Exploratory Analysis & Labelling
TF-IDF + Cosine Similarity
USE + Cosine Similarity
Jaccard Similarity
Zero Shot Classification

Figure 1: Automatic EndSARS tweet labeling methodology

332,371 tweets with EndSARS hashtags were collected between 1st of October, 2020 to 31st December, 2020 using the Twitter streaming API [6].

2.2 Data Cleaning & Preprocessing

Irrelevant hashtag-related tweets removal: The tweets consist of 23,166 unique hashtags which are quite ambiguous. Figure 2 shows the top sixteen hashtags used in the tweets, excluding the #endsars hashtag.

Convert to lowercase letters. Conversion into lower case letters is necessary because text analysis is case-sensitive. In the probabilistic model, the frequency of each letter is counted so that “Text” and “text” are treated as different words if the case conversion is not applied. All the letters are converted into lower cases.

Remove @user, symbols, and links. Twitter has special rules regarding reserved symbols and links that could cause confusion in later analysis. The @ symbol is used when the user mentions some other user to read this tweet. Remove punctuation and digits. This is a general step used in many text

[6] https://developer.twitter.com/en/docs/tutorials/consumingstreamingdata

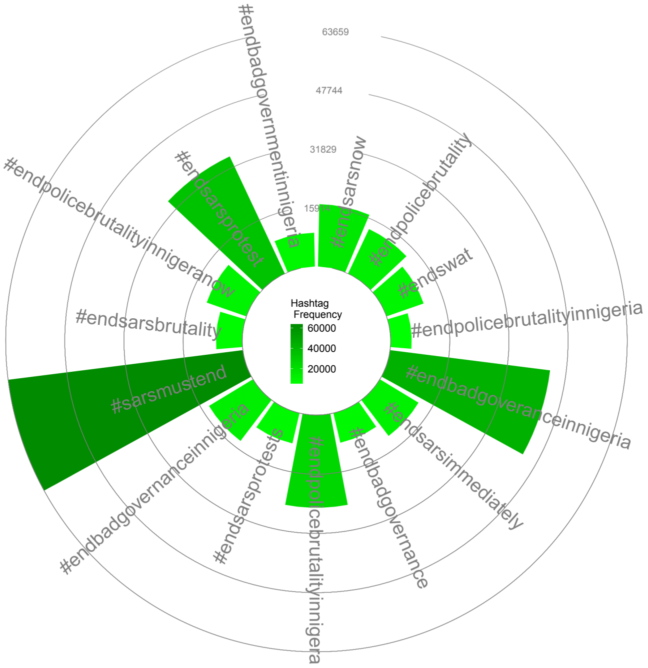

Figure 2: Top sixteen hashtags used in the tweets collected

mining techniques. Punctuation in sentences makes the text more readable for humans, but a machine does not distinguish punctuation and digits from other characters. Punctuation is removed because text analysis is not concerned with the digits. Numeric digits usually do not influence the meaning of the text.

Remove stopwords. Stopwords refer to the words which usually have no analytic value, words such as ‘a’, ‘and’, ‘the’ etc. These words make the sentences more readable to humans but confound the analysis. Words can be added to the list of stopwords depending on the specific requirements.

Stemming. Stemming is the process of eliminating affixes from words to convert the words into their base form; for example, stemming “run”, “runs” and “running” into “run”.

3 Exploratory Data Analysis and Label selection

As referenced in [7]the Lekki massacre event started at about 6:50 pm on the night of 20 October 2020. We explored a portion of the dataset that focused on this event where Figure 3 shows the most frequent words before 6:50 pm on that day, which reveals tweets about curfew, government, peaceful, protesters etc. Figure 4, on the other hand, shows the word cloud for tweets from 6:50 pm till the end of the day, which reveals tweets about killings, lekki, and blood.

3.1 Topic modeling

The 2020 EndSARS protest consisted of several activities as it took place in many cities across the country and also outside the country for a couple of days. Twitter became a forum for discussion and display of events during this period. To extract the different topics associated with the EndSARS protest tweets, we explored topic modeling. Clustering of tweet data that is processed by the LDA model method produced 6 topic clusters, where each topic has different words that are interconnected. The six topics includes complaints about police brutality events; the Lekki Massacre; activism; media coverage of the ongoing protests; lack of response from public figures including the president of

[7] https://en.wikipedia.org/wiki/2020$_{L}ekki_{s}hooting$

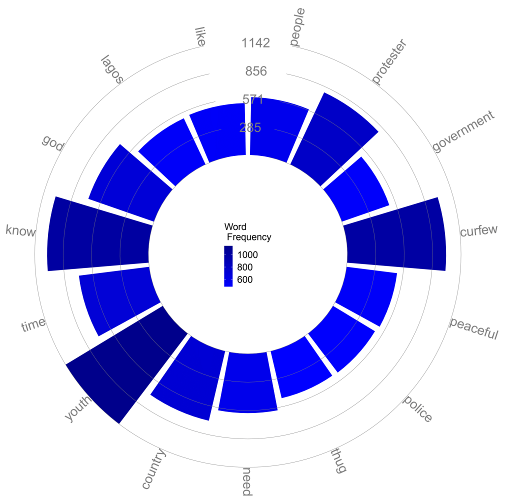

Figure 3: Most frequent words before 6:50 pm

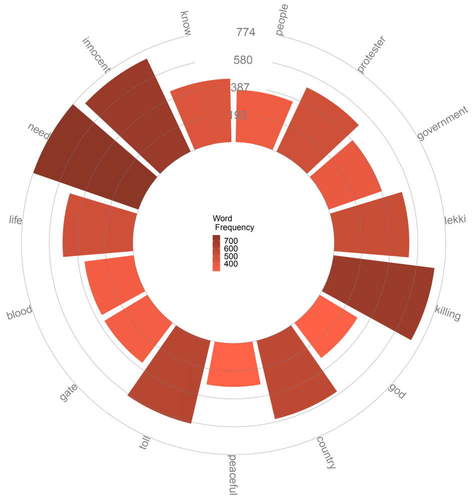

Figure 4: Most frequent words from 6:50 pm till the end of the day

Nigeria; and questions about moving forward and ending police brutality in Nigeria. From henceforth, these topics are referred to as follows: Activism, Police brutality, Media coverage, Lekki massacre, Questions about moving forward/back, No response from public figures. Cluster results for each topic was visualized with pyLDAvis as shown in Figure 5.

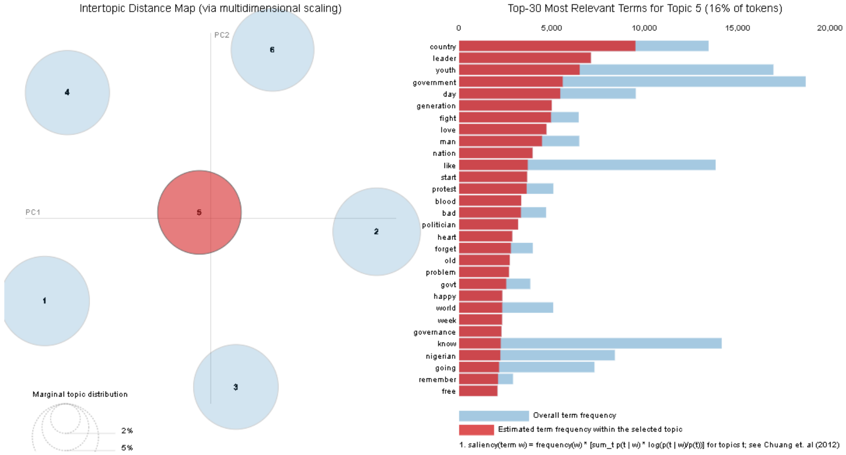

Figure 5: Sample Topic Visualization with pyLDAvis

3.2 Semi-Supervised labeling

0.5% of the dataset, which is 1085 samples, were randomly selected and labeled by volunteers[8] based on the topics selected from the topic modeling. Each tweet was labeled into a maximum of three classes, making it a multilabel dataset. Tweets that did not belong to any of the 6 topics selected were tagged irrelevant. The tweets (labeled and unlabelled) were vectorized with tf-idf, Universal Sentence Encoder (USE) and we explored the Cosine and Jaccard similarity to measure the distance between vectors of the unlabelled tweets and the sample labeled tweets in each label class. From the predictions made, we select the top-3 most confident predictions made by the model.

[8] Thanks to Mary Salami,Osuolale Emmanuel,Samson Kosemani,Samuel Akinseinde for their meticulous work on the data labelling and Adyasha Maharana for sharing her thoughts on USE

3.3 Self Supervised labeling

We leveraged the 6 topics extracted from the topic modeling and implemented the pre-trained zero-shot topic classification model by hugging face [9] to predict labels for each tweet. From the predictions made, we selected the top-3 most confident predictions.

3.4 Performance evaluation

The semi-supervised and self-supervised models were evaluated on another 0.1% of unlabelled data to compare the results of both methods. Our results are based on weighted average F1-score, precision, recall, and hamming loss. This metric is chosen because the tweets were multi-labeled. Each model gives a probability for each label between 0 and 1, but they are independent and do not sum to 1.

4 Result and Discussions

Table 1 shows sample tweets and the predicted labels for each method, comparing each of the predicted labels to the human labels, Zero-shot learning performed better especially with predicting accurate labels for the top 1 prediction. Although for tweets 2 and 3, USE labels were accurate for the top 1 prediction, the send labels are not related to the tweet, e.g., Lekki Massacre label should not be part of the top 3 labels for tweet 3 Jaccard similarity did not perform well on these tweets, since its predictions were almost the same for all tweets. Zero-shot learning proves to be a better approach to providing a more generalized label for multi-label tweets.

s/n	**Sample tweet**	**Human label**	**USE label**	**Jaccard label**	**TF-IDF label**	**Zero-shot label**
1	is there a missing person around you particularly during the movement friends and family should put out their information through the link below	Activism, Questions about moving forward/past, Media coverage	Media coverage, Lekki Massacre, Police brutality events	Lekki Massacre, Media coverage, Questions about moving forward/past	Police brutality events, Activism, Media coverage	Activism, Questions about moving forward/past, Media coverage
2	the change begins with you yes you	Activism	Activism, No Response from public figures, Questions about moving forward/past	Lekki Massacre, Media coverage, Questions about moving forward/past	Questions about moving forward/past, Activism, Irrelevant	Activism, Questions about moving forward/past, Media coverage
3	the people we elected to represent us in the national assembly are not say anything about the law in the north against those that are not from the north or not Muslim no motions raised they will soon get what they want	No Response from public figures	No Response from public figures, Lekki Massacre, Questions about moving forward/past	Lekki Massacre, Media coverage, Questions about moving forward/past	Activism, Questions about moving forward/past, Irrelevant	No response from public figures, Questions about moving forward/past, Activism
4	our point just started wed continue to protest as we see things unfold positively to our favor	Activism	Questions about moving forward/past, Media coverage, Activism	Lekki Massacre, Media coverage, Questions about moving forward/past	Questions about moving forward/past, Irrelevant,Activism	Activism,Questions about moving forward/past,Police brutality events

Table 1: A comparison of sample tweets and predicted labels

Table 2 shows the comparison between the Similarity methods and Zero-shot learning in terms of evaluation metrics. Hamming loss shows the fraction of the wrong labels to the total number of labels, it penalizes only individual labels i.e. at least one of the labels is correct while Weighted average precision, recall, and F1 score consider how many of each label there were in its calculation. Zero-shot learning gives a higher score for Precision, Recall, F1 score, and the lowest score for hamming loss.

Method	**Hamming Loss**	**Weighted Average Precision**	**Weighted Average Recall**	**Weighted Average F1 score**
Cosine similarity with TF-IDF	0.349	0.42	**0.71**	**0.50**
Cosine similarity with USE	0.397	0.44	0.58	0.44
Jaccard Similarity	0.535	0.02	0.22	0.04
Zero-Shot Learning	0.343	0.53	**0.73**	**0.51**

Table 2: Comparison of evaluation metrics for each method of tweet labelling

[9] https://joeddav.github.io/blog/2020/05/29/ZSL.html

5 Conclusion

We present a comparative analysis of semi-supervised and self-supervised machine learning classification for automatic labeling of tweets gathered during the EndSARS protest. In our findings self-supervised learning with zero shot had the best performance for multilabelling . While the training dataset will not be publicly available, the predictive models for automatic tweet labelling will be publicly released. Furthermore, our analysis highlights how machine learning approaches can be used to study social problems.

References

[1] Oscar Araque, Ganggao Zhu, and Carlos A Iglesias. A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems*, 165:346–359, 2019.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] Nina Cesare, Olubusola Oladeji, Kadija Ferryman, Derry Wijaya, Karen D Hendricks-Muñoz, Alyssa Ward, and Elaine O Nsoesie. Discussions of miscarriage and preterm births on twitter. *Paediatric and perinatal epidemiology*, 34(5):544–552, 2020.

[4] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, 2010.

[5] Herman Kamper, Aren Jansen, and Sharon Goldwater. A segmental framework for fully-unsupervised large-vocabulary speech recognition. *Computer Speech & Language*, 46:154–174, 2017.

[6] Adyasha Maharana, Kunlin Cai, Joseph Hellerstein, Yulin Hswen, Michael Munsell, Valentina Staneva, Miki Verma, Cynthia Vint, Derry Wijaya, and Elaine O Nsoesie. Detecting reports of unsafe foods in consumer product reviews. *JAMIA open*, 2(3):330–338, 2019.

[7] Edi Surya Negara, Dendi Triadi, and Ria Andryani. Topic modelling twitter data with latent dirichlet allocation method. In *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, pages 386–390. IEEE, 2019.

[8] Oluwafemi Oriola and Eduan Kotzé. Improved semi-supervised learning technique for automatic detection of south african abusive language on twitter. *South African Computer Journal*, 32(2):56–79, 2020.