ONLINE THREATS DETECTION IN HAUSA LANGUAGE

Abubakar Yakubu Zandam
Department of Computer Science
Federal University Dutse

Isa Inuwa-Dutse
University of Huddersfield
Federal University Dutse

Fatima Muhammad Adam
Department of Computer Science
Federal University Dutse

ABSTRACT

One of the widely used technological inventions is the Internet which gives rise to online social media platforms such as Twitter and Facebook to proliferate. These platforms are quite instrumental as a means for socialisation and information exchange among diverse users. The use of online social media to spread information can be both beneficial and harmful. From the positive side, the information can be useful in the areas of security, economy and climate change. Motivated by the growing number of online users and widespread availability of contents with the potential of causing harm, this study examines how online contents with threatening themes are being expressed in Hausa language. We collected the first collection of Hausa datasets with threatening contents from Twitter and develop a classification system to help in curtailing security risks by informing decisions on tackling insecurity and related challenges. We employ and train four machine learning algorithms: Random Forest (RF), XGBoost, Decision Tree (DT) and Naive Bayes, to classify the annotated dataset. The result of the classifications shows an accuracy score of 72% for XGBoost, 71% for RF, 67% for DT and Naive Bayes having the lowest of 57%.

1 INTRODUCTION

Online social media platforms have become quite instrumental for socialisation and information exchange. These platforms enable users to express their opinions and share messages with one another Nemes and Kiss (2021). Twitter[1] and Facebook[2] are widely used for effective dissemination of information Chen et al. (2020). Data from online social media have been widely used for various purposes such as understanding people's opinions Zishumba (2019), topic identification Kusumawardani and Basri (2017), Antypas et al. (2022), text classification Chaure et al. (2019) and entities recognition Nie et al. (2020). Text classification and Named Entity Recognition (NER) are one of the important tasks that is applied on social media data to extract useful information. Although text classification and NER are distinct tasks, the two are language-specific and the algorithms use for their implementation are influenced by the language type. While there exists some multilingual models, many languages do not have the linguistic resources sufficient for NLP-related tasks; hence, considered to be low resource languages (LRL) Tsvetkov (2017). The term LRL refers to languages with less resources and less statistical methods Magueresse et al. (2020). Hausa language is considered low resource due to limited resources to effectively build many downstream tasks in NLP. Hausa is one of the most widely spoken language with an estimate of over 100 million speakers majority residing in parts of Southern Niger and Northern Nigeria Inuwa-Dutse (2021). Hausa is one of the common languages that millions of users leverage as medium to write and posts information on social media. The use of online social media to spread information can be both beneficial and harmful. From the positive side, the information can be useful in the areas of security, economy and climate change. As one of the most crucial aspect of human life, information spread on social media can inform sound decisions on tackling insecurity and related challenges. Although information extraction and

[1] https://www.twitter.com
[2] https://www.facebook.com

analysis have been widely studied, especially English, information extraction in Hausa language based on data from Twitter will enrich the downstream tasks in LRL. In this study, we propose to leverage text classification to detect online posts with threatening themes or security risks in Hausa language. Thus, the aim is to develop a machine learning model capable of detecting Hausa posts with threatening content on Twitter. We create relevant Hausa datasets and build a model trained on the collected dataset to classify threatening online content. Essentially, the study contributes the following:

- Annotated corpus of threatening tweets and terms in Hausa language. The collection will offer a rich set of Hausa lexicons that can be use for various downstream tasks.
- A classification model capable of detecting threatening tweet in Hausa language. The model will classify post based on the identified entities as a threat-containing tweet or not.

The remaining part of the paper is structured as follows. Section 2 offers relevant studies and Section 3 provides the approach we followed in the study. Section 4 presents the relevant results and Section 5 concludes the study and offer pointers to future studies.

2 Related Work

As the name suggest, named entity recognition (NER) is concerned with identifying terms such as a name, place organization in a given text and classifying them into some predefined categories Roy (2021). NER has many applications in NLP such as information extraction, question-answering, machine translation, automatic summarisation and semantic annotation. Approaches in NER can be either ruled-based or machine learning approach Yi et al. (2020). Ruled-based approach focuses on searching for matched entities from a predefined category of entities. This approach relies on grammatical rules crafted by the language experts. This often results in precision but low recall. Another limitation of this approach is that the entities are hard-coded and has to be updated when there is new discovery of entity. On the other extreme, the machine learning approach involves searching a pattern and relationship in a text to create and train a model on a large collection of annotated dataset. A well-trained model should be able to classify previously unseen datapoints. Applying NER on social media textual data is a tedious and time consuming task because of the voluminous informal, short and unstructured nature of the data. Past studies have used Twitter data to build an NER model to extract location information Yenkar and Sawarkar (2021); Stavrianou et al. (2014). For low resource languages such as Hausa, there is a shortage of relevant linguistic corpus due to the lack of sufficient annotated corpora, part of speech (POS) tagger, morphological analyser, chunker, and parser. Noting the need to facilitate downstream tasks in LRL, past studies were geared towards developing useful Hausa corpus Suleiman et al. (2019); Oyewusi et al. (2021); Inuwa-Dutse (2021). Developing multilingual NER model is used as an alternative for languages lacking large corpora. For instance, the work of Oyewusi et al. (2021) developed a multilingual NER model on 5 Nigerian languages (English, Pidgin English, Igbo, Yoruba and Hausa). Cross-lingual transfer learning methods have been used for low resource languages for the sake of transferring knowledge from high to low resource language Enghoff et al. (2018); Mbouopda and Melatagia Yonta (2020); Oyewusi et al. (2021).

3 Methodology

The aim of this research is to analyse tweets made in Hausa language using Information Extraction (IE) for violence (threat) detection. This requires extraction of relevant keywords (entities) that may lead to violence from the tweet and train a machine learning model to classify the dataset. In this section, We describe the approach followed in the study.

3.1 Study Data

We collected datasets for this research from Twitter. As a platform, Twitter enables access to its data for research and other purposes using its Application Programming Interface (API). Table 3.1 shows the keywords used in retrieving the study data from Twitter.

Using the provided API, Algorithm 1 depicts the process we followed in searching and retrieving relevant data from the Twitter.

Table 3.1 Keywords use for Data Collection from Twitter

S/N	Keyword	Description
1	bbchausa	Tweets from bbchausa page.
2	bikin sallah	Tweets related to events occurring within hausa language native speakers
3	siyasa	Tweets related to politics
4	zanga zanga	Tweets related to riots, violence
5	hausa	Tweet related to hausa language generally

Algorithm 1: Steps for Collecting Hausa Language Tweets from Twitter Platform

Input: $S = [S_i]_{i=1}^{n}$, list of search keywords, n: number of search keywords, M: number of search iterations, N: number of tweets wanted.

1: Initialize M and N
2: Initialize $T = [T_t]_{t=0}^{N}$, the collected tweets.
3: for each iteration m = 1 to M do
4: for each query in S above, i = 1 to n do
5: Search tweets data with keyword S_i, from $D = [D_j]_{j=1}^{K}$, where D is list of tweets on Twitter
6: If S_i matches D_j do
7: Get its date of post
8: Get its ID
9: Get its retweet count
10: Get its favourite count
11: Get its full text
12: Get its screen name
13: Get its urls
14: End If.
15: Append D_j to T
16 Check for N, if satisfies maximum **break**
17: End for
18: End for
19: Print T
20: Create a Data Frame based on T.
Output: A Data Frame consisting list of tweets $[T_t]_{t=1}^{N}$

3.2 Data Cleaning

The data collected from the Twitter is noisy with many duplicates, retweets, hyperlinks in text, which need to be removed. Out of the 3001 instances collected, 586 instances are retweets and 641 instances are found to be duplicates data. These instances were removed and the total instances reduced to 1,774. The hyperlinks, hashtags (#) and at () symbols found in some tweets were removed using a sub-string matching of regular expression. To remove tweets that does not carry meaningful information, the number of unique words per tweet was calculated and all the tweets with distinct words less than 8 were filtered. This is to have concise data rich in information for model training and evaluation. Similarly, some tweets were found to be written in English words and some combination of English and Hausa (Engausa). Tweets posted in English were removed from the final collection. After the pre-processing, a total of 806 data instances was created and ready for annotation.

Stopwords Removal Stopwords are words that do not add much meaning to a collection of data and their removal will not change the interpretation or meaning of the sentence much. In addition to the standard list of stopwords, words and characters such as *'a', 'ni', 'to', 'su','.', '?' and '/'* have been flagged and removed accordingly. The final set of words in the collection was changed to lower case.

3.3 DATA ANNOTATION

To achieve the desired research objectives of classifying a tweet, some set of labels has to be learned by the machine learning model. These labels can be learned by pre-defining them through data annotation. The data annotation process was carried out manually to label the named entities in Hausa language based on the entities (labels) and description shown in Table 3.3. During the annotation, words associated with a threat, abuse, violence or threat-object and a location mentioned in a tweet are identified and extracted. Using these words, a tweet is classified as a threat-containing if it contains words related to threat or violence, and not-threat if it does not. The annotation was carried out by team of three annotators who were native and professional speakers of Hausa language. The first annotator identified threat-containing words in each tweet and labeled the tweet as a threat-containing or not, depending on the identified words. And the two annotators together use the result of the first annotator and ensure correct annotation was made for each tweet.

Table 3.3 Labels use for Data Annotation

S/N	Label	Description
2.	LOC	Name of a location, a place. E.g. Zamfara
3.	ABUSE	Word(s) that indicates abuse. E.g. Shege, ubanka
4.	VIOLENCE	Word(s) that indicates an act of violence or conflict. E.g. Zanga-zanga, kisa, farmaki
5.	THREAT	Word(s) that may result to violence. E.g. Hari, yan bindiga
6.	THREAT_OBJECT	An object that may be a threat or may be use during violence. E.g., bindiga, makamai

3.4 WORD EMBEDDING

Word embedding is a technique for mapping words to vectors in space. This technique enables machine learning algorithm to work on textual data by providing real numbers representation of words in a text. The semantic, syntactic and relationship with each word in a document can be captured using word embedding technique. In this research, a word2vec method was used to map the annotated text to real numbers to enable machine learning algorithms works on the data. There are two methods for word2vec, common bag of words (CBW) and skip n-gram. To achieved the research objective of this thesis of identifying words in a post signifying a threat CBW was used to map the words to real numbers. In CBW, a fixed-length vector representation of a text is created by counting the number of times each appears in the text.

3.5 SURVEY DATA

In addition to the raw data collected from Twitter, we conducted a survey to capture public perception about threatening online content, its harmful effect within a society and its usefulness in insecurity related decisions. This is to ensure that the research is guided by real-life facts as perceived by the public. The survey questionnaire was designed using *SurveyMonkey* [3] consisting of 10 questions (see Section A). A link to the survey was shared to volunteers via online channels. The survey targeted social media users and therefore all the respondents were ensured to be social media users having account on either Twitter, Facebook or both. The questions focus on the following three domains:

- the use of Hausa language in sharing information amongst social media users

[3] https://www.surveymonkey.com

- the use and effect of social media as a tool for disseminating abusive/threatening information online
- the benefits of information extraction on social media contents to identify security threat that may lead to violence before their occurrence

3.6 Survey Result

We received 60 responses from the survey we administered for the study. Out of the 60 respondents, 90% read and post information on social media in Hausa language with 60% of frequent encounter of post made in Hausa language. Of the 60 respondents, 36 believe that people use social media to post abusive, threatening or violent content; 54 of the respondents reported that threatening contents are prevalent in discourse related to politics, ethnicity and religion that could lead to crisis or violence as shown in Figure 1. Similarly, 70% of the respondents believe that extracting relevant information such as a name or a location mentioned in a post on social media can help identify and resolve some crises before they occurred (see Figure 2).

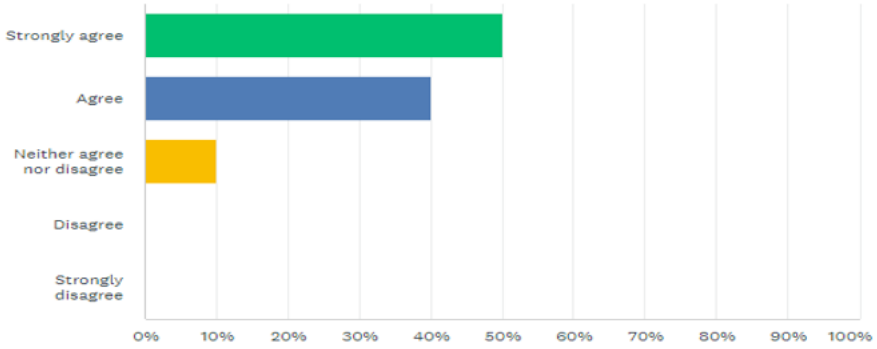

Figure 1: Votes on How Abusive Posts on Social Media Lead to Crisis and Violence

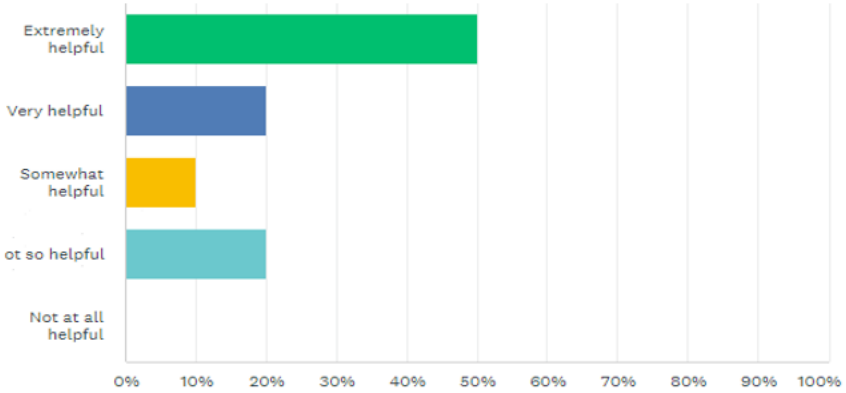

Figure 2: Votes on Use of Information Extraction for Violence Prevention

3.7 Preliminary Analysis

As mentioned earlier, the data for this research was collected from Twitter. Out of the 806 instances collected, 656 instances consist of only Hausa words representing 81% of the data collected while 150 instances consist of engausa words representing 19% of the entire dataset and which were removed during the models training and testing.

From Figure 3, it can be seen that more than 40% of the identified threats resulted from the theme of insecurity such as kidnapping and insurgency; about 30% were from political activities such as campaign, election, riots and hoodlums solely associated with politicians. While 28% of the threats were resulted from social activities such as abuses and discussions amongst social media users based on trending topics. And only 2% of the threats were a result of economic conditions such as corruptions, refugees, among others.

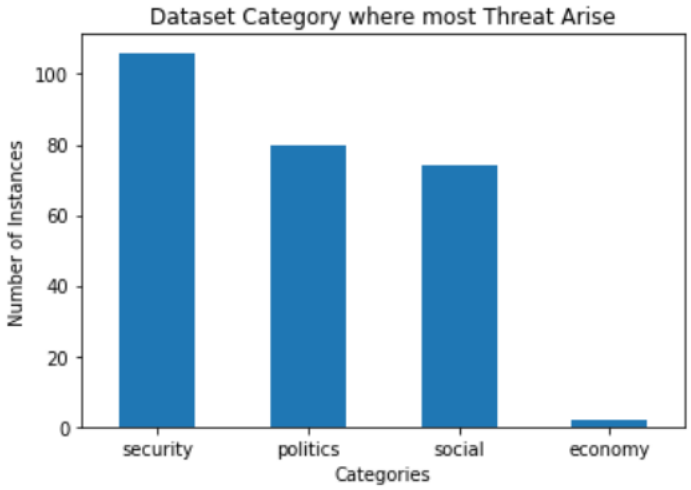

Figure 3: Proportion of Threat Categories in the Dataset

Figure 4 depict a word cloud showing important threat words that appeared most in the dataset. It can be seen that these words majorly associated with insecurity such as 'yan bindiga', 'yan ta'adda' and politics such as 'zanga zanga', 'bangar siyasa' were shown importance . Additionally, Figure 5 shows the various locations were these threat mostly arise. It can be seen that Kaduna was identified to be the most frequent word where most threats are associated to. This is followed by Kano, Zamfara, Borno, Neja, Sokoto, and Gombe in that order. Moreover, Figure 6 and 7 show word importance of abuse and violence words identified in the dataset respectively.

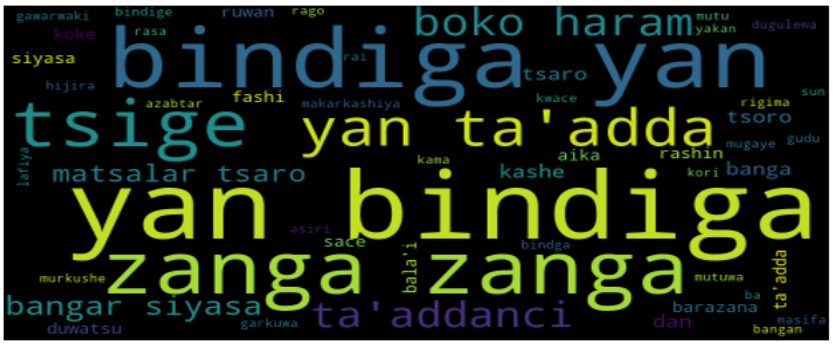

Figure 4: Word Cloud Depicting Threat Words in the Dataset

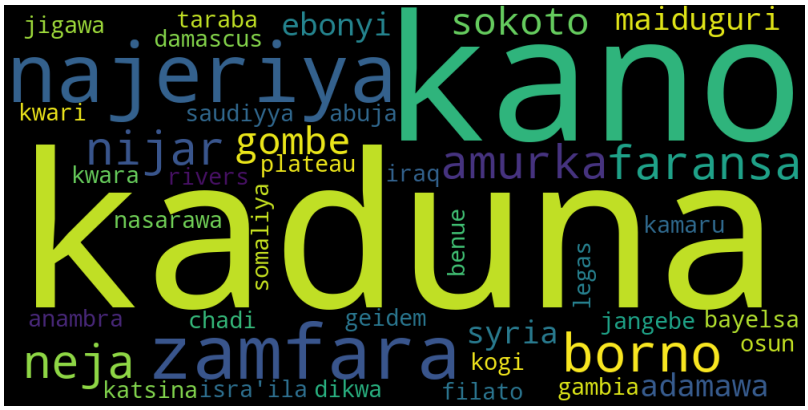

Figure 5: Word Cloud Depicting Locations Associated with Threat in the Dataset

Figure 6: Word Cloud Depicting Abuse Words in the Dataset

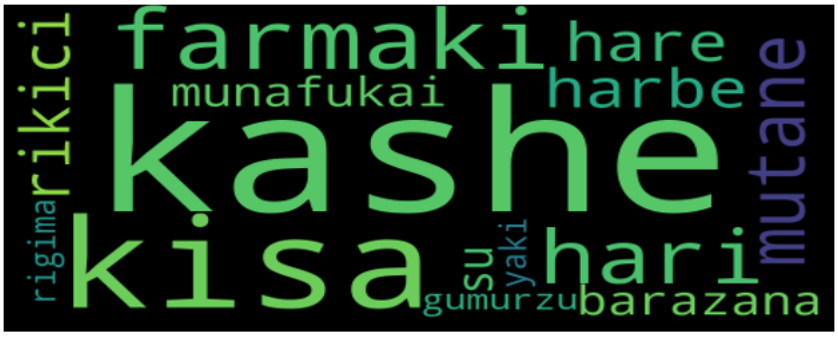

Figure 7: Word Cloud Depicting Violence Words in the Dataset

4 Result and Discussion

4.1 Model Building

This section describes the development and training of the machine learning algorithms to classify our annotated dataset. Four machine learning algorithms were trained and tested. These are: Random Forest, XGBoost and Decision Tree and Naive Bayes. The algorithms performance was evaluated after been trained with the 70% of the original dataset having the size of 459 instances and the remaining 30% having 197 instances was used for the evaluation.

To achieve the desired result, two classes were identified for each instance of the data based on the entities manually annotated from the tweet. A tweet can be a threat or not and the goal is to classify the instances to either one of the two classes.

4.2 PERFORMANCE METRICS

Performance metrics are indicators used to measure the performance of machine learning algorithms. In this study, four performance metrics were used to evaluate the performance of the machine learning algorithm trained. These are precision, recall, accuracy and confusion matrix.

Precision: Precision is a performance metrics used to measure the accuracy of a prediction made by a machine learning model. Precision shows how a model predict a correct result as correct. The maximum value for precision is 1. And a precision result close to 0 indicates poor accuracy by a model. Equation 1 below shows a formula use to obtain a precision of machine learning algorithm.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad \ldots\ldots\ldots\ldots \quad (1)$$

Where True Positive is the positive prediction predicted correctly while False Positive is the negative prediction predicted correctly.

Recall: Recall metric measure the ratio of positive prediction predicted correctly to total correct predictions. Recall can be obtain using the formula in equation 2 below.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad \ldots\ldots\ldots\ldots \quad (2)$$

Accuracy: Accuracy assess the performance of a model in making prediction by dividing the number of classifications a model predicts correctly with the total number of predictions made.

Confusion Matrix: Confusion matrix is a representation of the performance of classification algorithm usually presented in form of a table showing the outcomes of the prediction. Table 4.2 is an example of how confusion matrix is tabulated. To understand the table better, imagine the classification is binary with classes 1 and 0. True Positive (TP) represent the instances of class 1 correctly classified as class 1 and False Positive (FP) are instances of class 1 wrongly classified as class 0. Similarly, False Negative (FN) are instances of class 0 wrongly classified as class 1 and True Negative (TN) are instances of class 0 correctly classified as 0.

Table 4.2 Illustration of Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

4.3 CLASSIFICATION RESULTS

Table 4.3(a) shows the result of the performance of the four algorithms in classifying the tweets. From the table, it can be seen that XGBoost and Random Forest have the same recall values but different precision values with XGBoost having the highest. Decision Tree and Naive Bayes achieve very low precision values, however Naive Bayes produce higher recall values than all the three algorithms. The classification result shows that XGBoost algorithm produce an accuracy of 72% and therefore outperforms the remaining three algorithms in precision and recall values. Table 4.3(b) is

a confusion matrix for the four algorithms. XGBoost has the highest True Positive (TP) and lowest False Positive (FP). This indicates the ability of the model to correctly classified the tweets and therefore the most suitable model to achieve the objective of this research.

Table 4.3(a) shows the performance of the models used in the study.

Metric	Random Forest	XGBoost	Decision Tree	Naïve Bayes
Precision	0.628	0.659	0.532	0.407
Recall	0.397	0.397	0.368	0.515
Accuracy	0.711	0.721	0.670	0.574
Accuracy (%)	71	72	67	57

Table 4.3(b) Confusion Matrix of the Four Algorithms

	Random Forest		XGBoost		Decision Tree		Naïve Bayes	
	TP	FP	TP	FP	TP	FP	TP	FP
Predicted Positive	113	16	115	14	107	22	78	51
	FN	TN	FN	TN	FN	TN	FN	TN
Predicted Negative	41	27	41	27	43	25	33	35

5 Concluding Remarks

The use of social media to spread information can be beneficial to the humanity in various purposes ranging from security, economy to climate change. Security is one of the sector that information plays an important role. Analysis of the information spread on social media can guide so many security decision making and may prevent crises and violence before their occurrence. In this study, we developed a machine learning model that can classify a tweet written in Hausa language as a threat containing information or otherwise. Essentially, we contributed the following:

- a useful collection of dataset in Hausa language containing threatening terms and events extracted from Twitter. The dataset is also annotated to support relevant downstream tasks.
- a model to classify the annotated dataset expressed in Hausa language.

In the future work, we aim to develop and train machine learning model to understand the context of Hausa language vocabularies. Our developed model classify a tweet based on the presence of threat or violence related word in a tweet. A tweet containing word like 'bindiga' is classified as a threat-containing tweet by our model. However, that might not always be true for all cases and hence the limitation of this research.

References

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Leonardo Neves, Vítor Silva, and Francesco Barbieri. Twitter topic classification. *arXiv preprint arXiv:2209.09824*, 2022.

Ms Mrunali D Chaure, Jayant P Mehare, et al. Text classification and analysis with social media platform. *IJRAR-International Journal of Research and Analytical Reviews (IJRAR)*, 6(4):276–280, 2019.

Zi Chen, Badal Pokharel, Bingnan Li, and Samsung Lim. Location extraction from twitter messages using bidirectional long short-term memory model. In *GISTAM*, pages 45–50, 2020.

Jan Vium Enghoff, Søren Harrison, and Željko Agić. Low-resource named entity recognition via multi-source projection: Not quite there yet? In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 195–201, 2018.

Isa Inuwa-Dutse. The first large scale collection of diverse hausa language datasets. *arXiv preprint arXiv:2102.06991*, 2021.

RP Kusumawardani and MH Basri. Topic identification and categorization of public information in community-based social media. In *Journal of Physics: Conference Series*, volume 801, page 012075. IOP Publishing, 2017.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*, 2020.

Michael Franklin Mbouopda and Paulin Melatagia Yonta. Named entity recognition in low-resource languages using cross-lingual distributional word representation. *Revue Africaine de Recherche en Informatique et Mathématiques Appliquées*, 33, 2020.

László Nemes and Attila Kiss. Information extraction and named entity recognition supported social media sentiment analysis during the covid-19 pandemic. *Applied Sciences*, 11(22):11017, 2021.

Yuyang Nie, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai. Named entity recognition for social media texts with semantic augmentation. *arXiv preprint arXiv:2010.15458*, 2020.

Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, Ifeoma Okoh, Vitus Onuigwe, Mary Idera Salami, Opeyemi Osakuade, Sharon Ibejih, and Usman Abdullahi Musa. Naijaner: Comprehensive named entity recognition for 5 nigerian languages. *arXiv preprint arXiv:2105.00810*, 2021.

Arya Roy. Recent trends in named entity recognition (ner). *arXiv preprint arXiv:2101.11420*, 2021.

Anna Stavrianou, Caroline Brun, Tomi Silander, and Claude Roux. Nlp-based feature extraction for automated tweet classification. *Interactions between Data Mining and Natural Language Processing*, 145, 2014.

MA Suleiman, Muktar M Aliyu, and SI Zimit. Towards the development of hausa language corpus. *Int. J. Sci. Eng. Res*, 10:1598–1604, 2019.

Yulia Tsvetkov. Opportunities and challenges in working with low-resource languages. *Slides Part-1*, 2017.

Pranali Yenkar and SD Sawarkar. Gazetteer based unsupervised learning approach for location extraction from complaint tweets. In *IOP Conference Series: Materials Science and Engineering*, volume 1049, page 012009. IOP Publishing, 2021.

Feng Yi, Bo Jiang, Lu Wang, and Jianjun Wu. Cybersecurity named entity recognition using multi-modal ensemble learning. *IEEE Access*, 8:63214–63224, 2020.

Kudzai Zishumba. *Sentiment Analysis Based on Social Media Data*. PhD thesis, 2019.

A Sample Survey Questions

- What is your gender?
  - ○ Male
  - ○ Female
- Do you have a social media account?
  - ○ Yes
  - ○ No
- Which of the following social media platforms do you have an account?
  - ○ Twitter
  - ○ Facebook

- ○ Twitter and Facebook

- Which of the following social media platforms do you use most?
  - ○ Twitter
  - ○ Facebook
- Do you post or read information in Hausa language from your social media platform?
  - ○ Yes, I do
  - ○ No, I don't
- How often do you encounter posts on social media written in Hausa language?
  - ○ Many
  - ○ Few
  - ○ Very Few
- How satisfied or dissatisfied are you with the statement: *People within our community use social media to spread news that may serve as a threat for insecurity or violence within our community.*
  - ○ Very satisfied
  - ○ Somewhat satisfied
  - ○ Neither satisfied nor dissatisfied
  - ○ Somewhat dissatisfied
  - ○ Very dissatisfied
- How much do you agree with the statement: *The information spread by individuals and organisations on social media platforms if used by security agencies can help prevent a lot of crises and violence before their occurrence.*
  - ○ Strongly agree
  - ○ Agree
  - ○ Neither agree nor disagree
  - ○ Disagree
  - ○ Strongly disagree
- Abuses made by people on social media regarding political, religious or ethnic differences have in one way or another lead to crisis and violence.
  - ○ Strongly agree
  - ○ Agree
  - ○ Neither agree nor disagree
  - ○ Disagree
  - ○ Strongly disagree
- Some information such as a name, a location, mentioned in a post on social media platforms can help identify and resolve some crises before their occurrence.
  - ○ Extremely helpful
  - ○ Very helpful
  - ○ Somewhat helpful
  - ○ Not so helpful
  - ○ Not at all helpful