ONLINE THREATS DETECTION IN HAUSA LANGUAGE

Anonymous authors
Paper under double-blind review

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 and develop a detection system to help in curtailing security risks by informing decisions on tackling insecurity and related challenges. We employ and train three machine learning algorithms: Random Forest (RF), XGBoost and Multilayer Perceptron (MLP) to classify the annotated dataset. The result shows an accuracy score of 85% for both RF and MLP. However, MLP outperforms RF in precision and recall values.

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 analysis have been widely studied, especially English, information extraction in Hausa language based on data from Twitter and Facebook will enrich the downstream tasks in LRL. In this study, we propose to leverage NER and 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 and Facebook. We create relevant Hausa datasets and build a model trained on the collected dataset to detect threatening and inciteful online content. Essentially, the study contributes the following:

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

- Annotated corpus of threatening posts and terms in Hausa language. The collection will offer a rich set of Hausa lexicons that can be use for various downstream tasks.
- A detection model capable of detecting threatening posts in Hausa language. The model will classify post based on the identified entities as a threat-containing post 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 post made in Hausa language on social media using Information Extraction (IE) for violence (threat) detection. This requires extraction of relevant keywords (entities) that may lead to violence from posts written in Hausa language and train a machine learning model to identify those entities when presented with a post previously not known to the model. We describe the approach followed in the study.

3.1 Study Data

We collected datasets from Twitter and Facebook. As a platform, Twitter enables access to its data for research and other purposes. 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 Twitter.

The study also utilises data from Facebook. Unlike Twitter, the data is quite limited due to the use of manual collection method. A total of 284 post made in Hausa language on different topics were collected and readied for annotation. The data consist of two attributes, id (identity of the post) and text (post). After preprocessing, the data from the two platforms, Twitter and Facebook were merged to build a reliable data corpus of 801 instances.

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.

```
Initialize M and N
Initialize T = [T_t]_{t=0}^{N}, the collected tweets.
for each iteration m = 1 to M do
    for each query in S above, i = 1 to n do
        Search tweets data with keyword S_i, from D = [D_j]_{j=1}^{K}, where D is list of
        tweets on Twitter
        If S_i matches D_j do
            Get its date of post
            Get its ID
            Get its retweet count
            Get its favourite count
            Get its full text
            Get its screen name
            Get its urls
        End If.
    Append D_j to T
    Check for N, if satisfies maximum break
    End for
End for
Print T
Create a Data Frame based on T.
```

Output: A Data Frame consisting list of tweets $[T_t]_{t=1}^{N}$

3.1.1 Data Cleaning

The data collection, especially the dataset from 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. The pre-processed data from twitter was then merged with the data collected from Facebook and a total of 801 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.

3.2 DATA ANNOTATION

To achieve the desired research objectives of classifying a post from social media and detect some entities, 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 3.2.

Table 3.2 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

3.3 SURVEY DATA

In addition to the raw data collected from Twitter and Facebook, we conducted a survey to capture public perception about inciteful/threatening online content. 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
- 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

3.4 PRELIMINARY ANALYSIS

As mentioned earlier, the data for this research were collected from two social media platforms, Twitter and Facebook. Out of the 801 instances collected, 517 instances were collected from Twitter and the remaining from Facebook. Figure 1 shows relative composition of the dataset from the two data sources.

From Figure 2, it can be seen that more than 45% of identified threat resulted from the themes of insecurity such as kidnapping and insurgency; about 35% were from political activities such as campaign, election, riots and hoodlums solely associated with politicians. While only 15% of the threats were resulted from socio-economic activities such as corruption, refugees among others. Figure 3 depict a word cloud showing important threat words that appeared most in the dataset. It can be seen that these words are majorly associated with insecurity and politics as explained earlier. Additionally, Figure 4 shows the various locations were these threat mostly arise. At large, Nigeria was identified to be the most frequent word where most threats are associated to. This is followed by Zamfara, Borno, Abuja, Kaduna, Gombe, Nijar and Sokoto in that order.

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

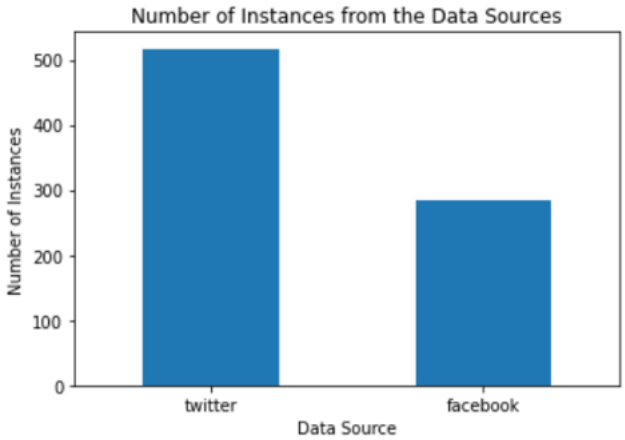

Figure 1: Data Sources

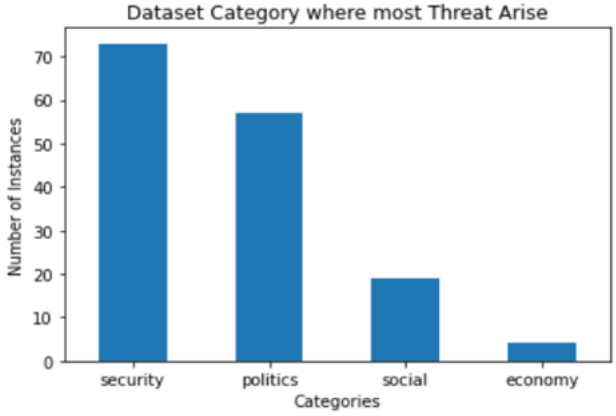

Figure 2: Proportion of Threats in the Dataset

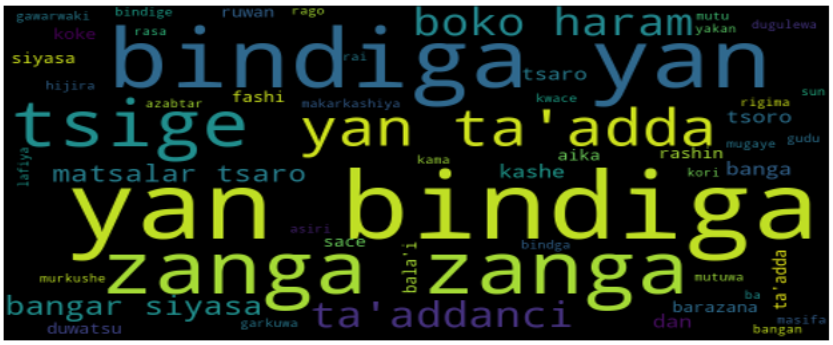

Figure 3: Word Cloud Depicting Threat Words in the Dataset

Moreover, Figure 5 and 6 show word importance of abuse and violence words identified in the dataset respectively.

Figure 4: Word Cloud Depicting Locations Associated with Threat in the Dataset

Figure 5: Word Cloud Depicting Abuse Words in the Dataset

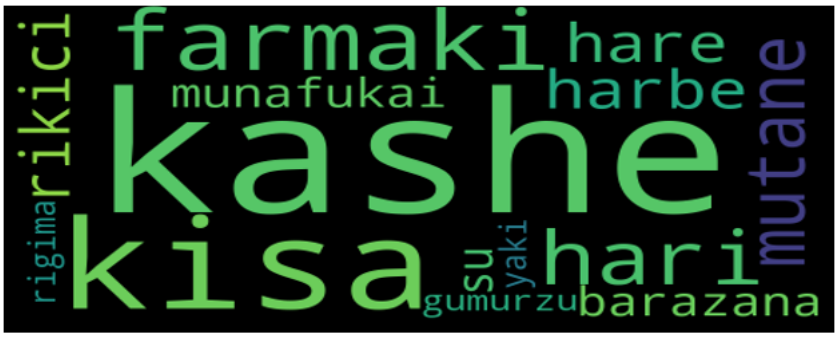

Figure 6: Word Cloud Depicting Violence Words in the Dataset

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. Three machine learning algorithms were implemented. These are: Random Forest, XGBoost and Multilayer Perceptron. The algorithms performance was evaluated after been trained with the 70% of the original dataset having the size of 560 instances and the remaining 30% having 241 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, three performance metrics were used to evaluate the performance of the machine learning algorithm trained. These are precision, recall and accuracy.

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.

Table 4.2 shows the result of the performance of the three algorithms in classifying the posts. From the table, it can be seen that multilayer perceptron and random forest perform equally with an accuracy of 85%, however, multilayer perceptron outperforms random forest in precision and recall values. XGBoost algorithm achieve an accuracy of 83% less than both random forest and multilayer perceptron, however, outperforms them in recall values.

Table 4.2 shows the performance of the models used in the study.

Evaluation Metrics	Random Forest	XGBoost	Multilayer Perceptron
Recall	0.135	0.189	0.142
Precision	0.500	0.412	0.600
Accuracy	0.846	0.834	0.846

4.3 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 7. 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 8).

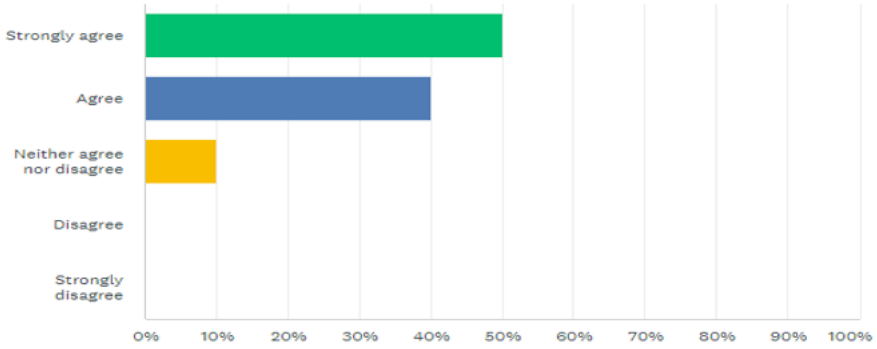

Figure 7: Votes on How Abusive Posts on Social Media Lead to Crisis and Violence

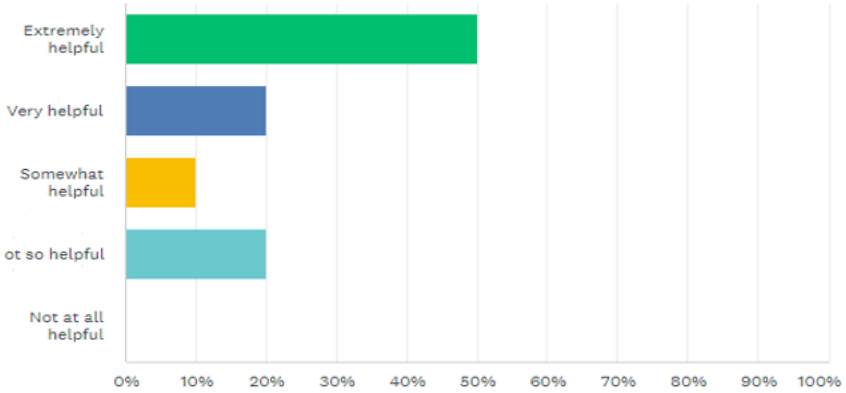

Figure 8: Votes on Use of Information Extraction for Violence Prevention

5 Conclusion and Future Work

This rapid development of online social sites gives people the freedom to express their opinions, share messages with one another and communicate information that is vital within their community Nemes and Kiss (2021). 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 post on social media 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 and Facebook. The dataset is also annotated to support relevant downstream tasks.
- a model to detect threatening posts and associated events expressed in Hausa language.

In the future work, we aim to develop and train machine learning model to understand the semantics of Hausa language vocabularies. The model will be employ to develop entity recognition system to identify words containing threat or violence from the social media data written in Hausa language.

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

- 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