THE FIRST LARGE SCALE COLLECTION OF DIVERSE HAUSA LANGUAGE DATASETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hausa language belongs to the Afroasiatic phylum, and with more first-language speakers than any other sub-Saharan African language. With a majority of its speakers residing in the Northern and Southern areas of Nigeria and the Republic of Niger, respectively, it is estimated that over 100 million people speak the language. Hence, making it one of the most spoken Chadic language. While Hausa is considered well-studied and documented language among the sub-Saharan African languages, it is viewed as a low resource language from the perspective of natural language processing (NLP) due to limited resources to utilise in NLP-related tasks. This is common to most languages in Africa; thus, it is crucial to enrich such languages with resources that will support and speed the pace of conducting various downstream tasks to meet the demand of the modern society. While there exist useful datasets, notably from news sites and religious texts, more diversity is needed in the corpus. We provide an expansive collection of curated datasets consisting of both formal and informal forms of the language from refutable websites and online social media networks, respectively. The collection is large and more diverse than the existing corpora by providing the first and largest set of Hausa social media data posts to capture the peculiarities in the language. The collection also consists of a parallel dataset, which can be used for tasks such as machine translation with applications in areas such as the detection of spurious or inciteful online content. We describe the curation process – from the collection, preprocessing and how to obtain the data – and proffer some research problems that could be addressed using the data.

1 INTRODUCTION

Hausa language is one of the most widely spoken Chadic language, which is part of the Afroasiatic phylum. With its majority speakers residing in the Northern and Southern areas of Nigeria and the Republic of Niger, respectively, it is estimated that over 100 million people speak the language. Many countries in West Africa and sub-Saharan countries adopt Hausa as the *lingua franca*, hence making it one of the dominant Chadic language spoken across the continent (8). Major Hausa regions in Nigeria include Kano, Sokoto, Jigawa, Zaria, and Katsina. Figure 1 shows a visual depiction of the phylogenetic evolution of Hausa language. The figure shows the linguistic family and orthographic origin of the language. Hausa language relies on Arabic and Latin alphabets (both consonants and vowels). *Ajami* is the *Arabic* version of written Hausa that has been widely in existence in pre-colonisation Hausa land, and the *Boko* script is based on *Latin* alphabet, which has been popularised since the down of colonisation in Hausa land.

Motivation In the quest to make computers more amenable to natural language through useful computational tools, relevant datasets are being collected to support various downstream tasks. As noted earlier, Hausa is considered a well-studied and documented language among the sub-Saharan African languages (8), but a low resource language from NLP perspective due to limited resources to undertake NLP-related tasks. Because this is common to most languages in Africa (9), it is crucial to enrich such languages with resources that will support and speed the pace of conducting various downstream tasks to meet the demand of the modern society. Although Google LLC[1] has access

[1] https://www.google.com/

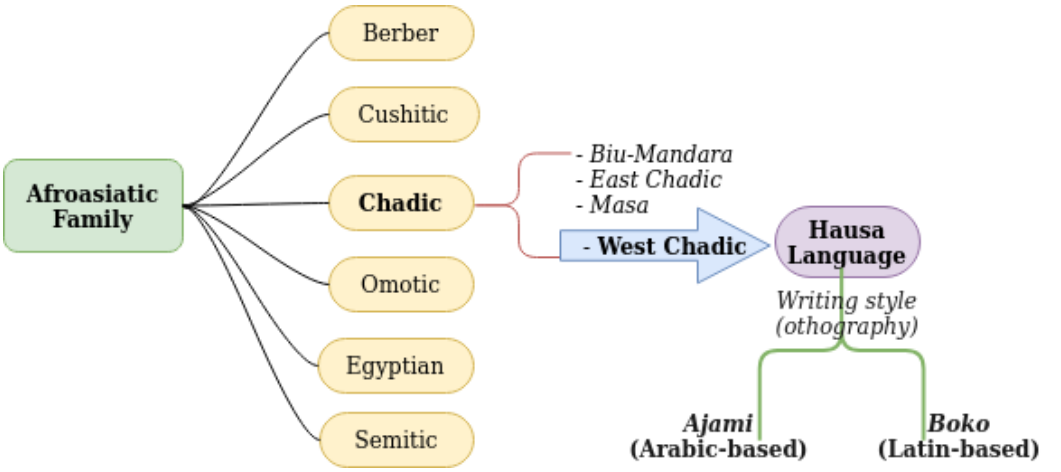

Figure 1: Linguistic family and orthographic origin of Hausa language. The *ajami* is a form of written Hausa text that is based on Arabic alphabets and the *boko* script is based on Latin alphabets.

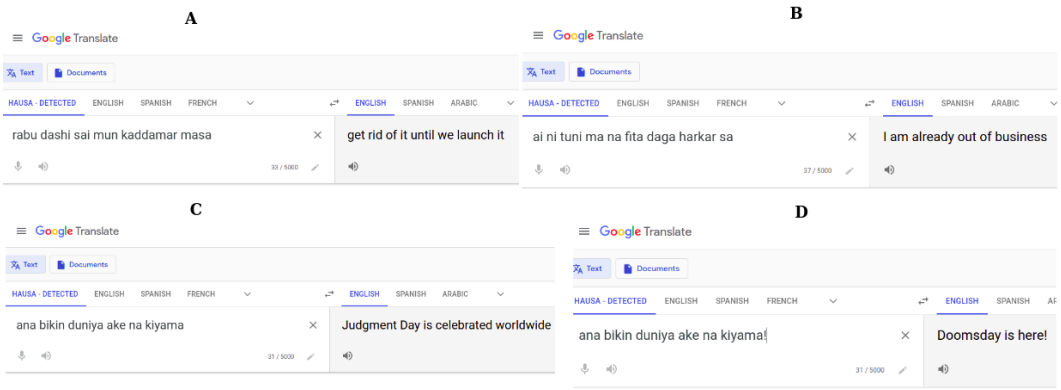

Figure 2: Some examples of translated Hausa texts using Google's translation engine. The texts in *panels A, B, C* and *D* have been incorrectly translated. The use of an exclamation mark (!), *motsin rai* in Hausa, in *panel D* drastically changed the message.

to tones of diverse data, anecdotal evidence suggests that its transcription system embedded in the video-communication (*Meet*[2]) performs poorly in recognising Hausa speech.

For a basic assessment of the efficacy of Google's translation engine[3] on Hausa language, Figures 2 and 3 show some examples of translated Hausa texts using the translation system. The discrepancies in the figures (2 and 3) can be attributed to the lack of sufficient and appropriate data to train relevant language models. Some of the informal or day to day terms used in the language can only be obtained via online social media, not news websites or religious texts. Among the implications of the incorrect translation is misinterpretation or out of context interpretation of events. One of the reasons of the above challenges is the lack of diverse datasets from various sources in larger quantities that will enable the development of relevant resources and tools with a wide range of applications in uncovering meanings in data. The following constitutes the motivations for undertaking this study:

- the existing NLP resources for downstream tasks in Hausa language relies mostly on formal text collections notably from news sites, such as BBC Hausa Service[4] and VOA Hausa Service[5]. Others include parallel datasets from religious texts such as (3) and *Tanzil*[6] for tasks such as automatic speech recognition, named entity recognition (NER) and part of speech

[2] https://meet.google.com/
[3] See https://translate.google.co.uk/
[4] https://www.bbc.com/hausa/
[5] https://www.voahausa.com/
[6] http://tanzil.net/docs/tanzil_project

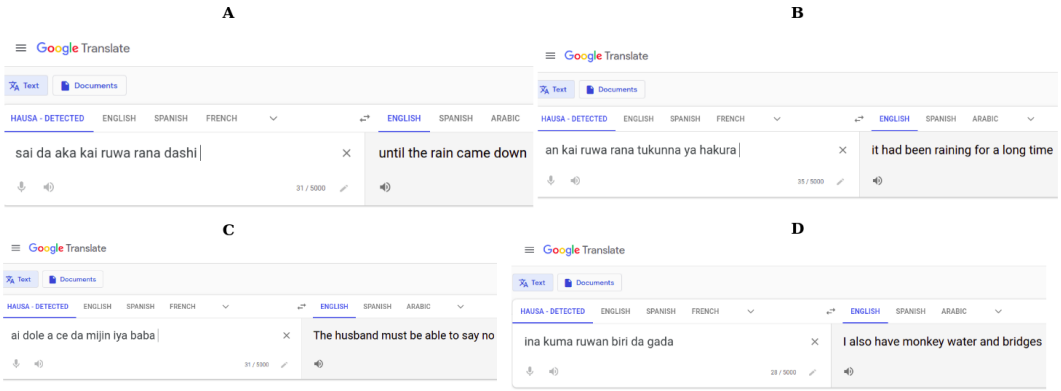

Figure 3: More examples of Hausa texts that have been incorrectly translated. The texts, mostly proverbial, in *panels A, B, C* and *D* are examples of common conversational setting.

tagging. Thus, the existing Hausa corpora suffer from limited diversity and availability of curated ground-truth data for large scale downstream tasks.

- no known existing collection of curated Hausa datasets from social media platforms, hence, the need for diverse datasets to support various NLP-related tasks.
- Other languages have received substantial contributions when compared to the most dominant Chadic language (Hausa). For instance, the HLT for African Languages (2) is focused on notably Yoruba and Igbo.

This paper aims at addressing the aforementioned gap by making available relevant datasets that will facilitate downstream tasks.

Contributions The main focus of this study is to contribute diverse datasets to support NLP research that relies on Hausa language. Noting the above issues, the study offers the following contributions:

- to enrich the language with useful resources that will facilitate downstream tasks, we provide an expansive collection of curated datasets[7] consisting of both formal and informal forms of the language from refutable websites and online social media networks, respectively. Based on the collection sources, the corpus consists of two broad categories: *Social Stream Category* (Section 3.1) and *Website and Blog Category* (Section 3.2). The *Social Stream* category consists of Twitter data from users interacting in Hausa language, and the *Website and Blog* category is obtained through the scrapping of relevant websites and blogs. The collection also consists of parallel data, which can be used for natural language processing and machine translation tasks with a wide area of applications.
- because the data is from various sources – news and government web sites, blog posts and social media platforms – the collection spans many aspects of social life. Moreover, the inclusion of social media data will help in capturing the peculiarities in the language and how colloquial expressions manifest in the language.
- we make available additional parallel dataset to support downstream tasks such as machine translation and automatic speech recognition systems.
- we present a process to streamline the process of obtaining more diverse datasets in Hausa from social media platforms. We describe how to retrieve the hydrated version of the data and proffer some research problems that could be addressed using the data.

The remaining of this paper is structured as follows. Section 2 presents a summary of relevant studies and Section 3 presents a detailed description of the data collection and processing. Section 4 concludes the study and discusses some future work.

[7]See https://github.com/ijdutse/hausa-corpus for details.

2 RELATED WORK

As the major spoken language among the sub-Saharan African language, Hausa language is estimated to have over 100 million speakers. Geographically, the major dialects of the language include Eastern Hausa (e.g. Kano), Western Hausa (e.g. Sokoto) and the Aderanci dialects from Niger. Hausa is considered a well-studied and documented language among the sub-Saharan African languages (8), however, from NLP perspective, it is viewed as a low resource language due to limited resources for computational linguistic. This is common to most languages in Africa. There is a growing body of studies exploring various aspects of the language. Of interest to this study is the prevailing efforts to enrich the language along the following dimensions.

Existing Corpus From a computational linguistic point of view, Hausa is categorised as low-resource language owing to the limited resources to computationally analyse the language and support downstream tasks. Other languages have received substantial contribution when compared to the most dominant Chadic language (Hausa) widely spoken in Africa. Generally, there are limited linguistic resources to support NLP tasks involving African languages. The commonly used parallel datasets for NLP tasks, such as Machine Translation, include (1) Tanzil dataset[8], which is a translation of the Holy Qur'an in various languages including Hausa. (2) JW300 dataset (3) consisting of multilingual parallel sentences, and is used in the *Masakhane* Project (9) as the default dataset for open implementation. So far, the *Masakhane* Project[9] is one of the biggest collaborative attempts to support African language developments through the development and support for open-source linguistic resources and tools for African languages (9). The work of (4) noted the lack of adequate parallel and diverse dataset for language translation. This led the authors to embark on a project aimed at improving machine translation capability for low-resource language. The data consists of *Hausa–English* parallel corpus, which is useful for translation and as evaluation benchmark for *English–Hausa* Neural Machine Translation (NMT).

Language Models The work of (1) contributes a Hausa-based words embedding model that is trained on formalised texts from news web sites and religious texts to support downstream tasks. Noting the inadequacy of models trained using high-resource languages (such as English) to be transferred for training low resource languages (such as Hausa), the work of (6) is based on multilingual transformer models to improve accuracy and effectiveness of low resource language models. The use of more diverse datasets from various sources such as social networks will help in capturing the peculiarities in the language, thus improving the language models.

Downstream tasks in low resource languages There are relevant parallel datasets in place to support downstream tasks such as machine translation and automatic speech recognition (ASR) systems. Noting how the limited resources available to facilitate Human Language Technology (HLT), the work of (2) is focused on speech technology for some dominant indigenous languages in Nigeria (2). Similarly, the work of (10) presents a database of text and speech, transcribed speech resources, in Hausa language. The speech collection is based on the GlobalPhone project ((11)), and the textual part collected from Hausa-based websites, mostly from news sites. Using the collected datasets, the authors develop an ASR system to improve pronunciation and acoustic modelling in Hausa language.

Taking the aforementioned resources into account, the diversity of the datasets utilised can be improved by incorporating various sources that capture more dialects and social media data texts. Thus, the datasets in this study can be used for developing and improving relevant resources for downstream tasks.

3 DATASETS

In the quest to make computers more amenable to natural language through useful computational tools, numerous collection of diverse datasets are being collected. Based on the collection sources, the corpus is categorised into the following: *Social Stream* and *Website and Blog* categories. Accordingly, the collection consists of two sub-categories of Twitter data, from users interacting in

[8] http://tanzil.net/docs/tanzil_project
[9] https://www.masakhane.io/

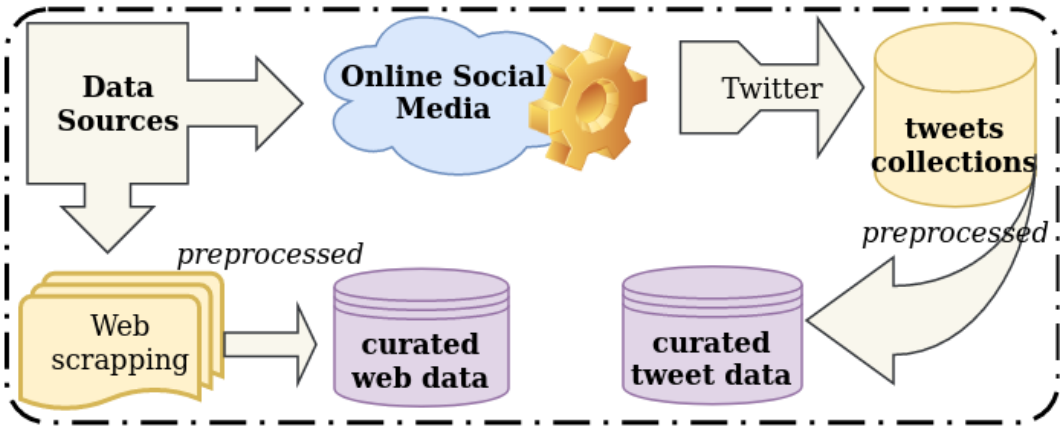

Figure 4: A summary of the data collection pipeline.

Hausa language, and two sub-categories of web scraping data from various Hausa-based language sites and historic online files. The collection also consists of parallel data, which can be used for downstream tasks such as machine translation. Figure 4 shows a summary of the data collection pipeline.

3.1 Social Stream Category

The proliferation of various online information on social media platforms such as Twitter and Facebook makes it possible to harvest relevant posts. With the evolving nature of communication in which majority of the populace obtain useful information from the web and social media platforms, it is pertinent to retrieve as much as possible relevant online datasets on low resource languages. In view of this, the *Social Stream* collection contains *tweet objects*[10] retrieved from Twitter using the platform's application programming interface (API). The collection activity starts on 11/12/20 to date (13/02/21) and consists of the following sub-categories.

- **account-based collection:** this relies on specific usernames, mostly of news channels, to collect data posted in Hausa language. Because of the presence of news channels, the category usually consists of formal posts on news items that attract other users to engage through *comment, reply, retweet* or *like*. This allows for the retrieval of more data on users' engagements using slang, memes and other colloquial or informal posting styles in the local language.
- **miscellaneous collection:** the miscellaneous set mostly consists of users who engage or interact with the content posted by users in the *account-based collection*. For a more diverse dataset, the miscellaneous set also consists of post collected using relevant *hashtags* and *keywords* (see Table 1). The use of *hashtags* and *keywords* offers a generic collection of daily tweets spanning numerous topics posted in Hausa language.

We define *main tweet*, as a post or tweet that attracts engagements by other users via replies. For annotation and sharing purpose, each main tweet (prefixed with MT) is followed by its corresponding reply, prefixed with MTR. Under this setting, the main tweet is not unique but the reply is.

Getting the hydrated tweets The full version of the *social stream category* cannot be made available due to Twitter's policy on content redistribution[11]. Instead, we provide relevant IDs of the *tweet-object* to be used in retrieving the complete data, including the meta-information about each tweet[12]. Moreover, we include a short *Jupyter notebook* with a description of how to retrieve the

[10] a *tweet object* is a complex data object composed of numerous descriptive fields, which enables the extraction of various features related to a post

[11] see https://developer.twitter.com/en/developer-terms/agreement-and-policy

[12] The data presented in this paper are available at https://github.com/ijdutse/hausa-corpus

Table 1: Summary of data collection sources and corresponding descriptions

Social Stream	*account-based collection* from Twitter
Description	tweets from specific account handles collected via Twitter's API
Handles	@bbchausa, @voahausa, @HausaDw, @RFI_Ha, @freedomradionig, @aminiyatrust, @HausaRadio, @hausaonline, @rariyajarida, @CRIHausaRadio, @TRTHausa, @HausaTranslator, @HausaDictionary, @PTimesHausa,
Social Stream	*miscellaneous collection* from Twitter
Description	tweets collected from Twitter using diverse keywords
Keywords	Hausa language, Yaran Hausa, Dan Hausa, Hausa day, Hausa week, barkan mu da rana, sannu da zuwa, lokaci, gobe da rana, sanyi, yan siyasa, majalisa, shugabanni, labarai, #HausaDay, yan sanda, sojoji, wasan kwallon kafa, #hausaday
Website & Blog	*Major websites for non-tweet-based collection*
Description	more formal texts from major news channels and a chronicle of posts
Some urls	https://www.bbc.com/hausa,https://www.voahausa.com/,https://freedomradionig.com/,http://indigenousblogs.com/ha/,https://www.rumbunilimi.com.ng/,http://fagenmarubuta.blogspot.com/

data, especially the hydrated data, and render it usable. Alternatively, interested users may use the *hydrator package* (12).

3.2 Website and Blog Category

This category comprises data from Hausa-based news sources, blog posts and from historic files from various sources. Using a custom web scrapping[13] crawler, we scrap useful data from the relevant websites; see Table 1 for some of the visited sites and Appendix 4 for more details. Because the content in this category mostly consists of well-written texts, it is possible to examine and compare the formal written form of the language with its informal counterpart from social media platforms.

3.3 Parallel Data

As a first step in creating a parallel corpus (in English) for the corpus (in Hausa), we leverage Google's translation API, *Googletrans*[14], to efficiently create a parallel English corpus for the collected Hausa corpus (using the *Social Stream* category). We sample from the *Social Stream* category for the translation, and subset of the translated segments or sentences is manually validated for correctness. Similar to the *Social Stream* category, the parallel collection consists of *Hausa main tweet* (HMT), *Hausa main tweet reply* (HMTR), *English main tweet* (EMT) for the translated HMT, and *English main tweet reply* (EMTR) for the translated HMTR.

3.4 Preprocessing and meta-analysis

Due to the multifaceted nature of the datasets, we applied the following transformation so that all the collections conform to a unified format for redistribution. This is to improve usability and suitability for sharing and further analysis.

Preprocessing Noting how messy the *Social Stream* category could be and difficult to use directly without cleaning (7), we applied the following rudimentary cleaning process involving tokenisation, stopwords removal and text formatting. For normalisation, we remove *urls, user mentions, emojis* and *non-Hausa posts* from the collection. Both the raw and the cleaned version of the applicable datasets are made available because some of the raw tweets or replies contain only *user mention* or *emojis*, which we designate as stopwords in the cleaned version of the data. After the removal of spaces and special symbols in the *Website and Blog* category, each paragraph or chunk of texts is split into smaller unit punctuated by a *full stop*.

[13]using the *BeautifulSoup* package, available at www.crummy.com/software/BeautifulSoup
[14]https://translate.google.co.uk/

Table 2: A summary of basic statistics in the collected datasets. #DS and #TS denote dataset size and tokens size the collection, respectively; $\mu_{sent.}$ refers to the average sentence size in each data category

	Category	#DS	#TS	$\mu_{sentence\ size}$	Description
Datasets	Main Tweets	4055	42,545	11	Tweets from social stream
	Reply Tweets	10,742	80,025	8	Tweets from social stream
	News and Blogposts	29,371	932,523	32	Data collection from websites and blogs
	Web Documents	12,277	10507117	856	Data collection from websites and blogs
	Merged Datasets	35,706	65754433	1842	Merged datasets for training a language model
	Parallel Main Tweets	23	6331	13	Parallel data from social stream
	Parallel Reply Tweets	492	3399	7	Parallel data from social stream

Figure 5: Some common terms in the tweets collection.

Meta-analysis Because the datasets in the corpus have been collected over time, we conducted longitudinal and exploratory analyses for better understanding. Table 2 shows basic statistics about the collected data, including the number of unique main tweets, replies, mean token size, and the average length of sentences in each category. For the *Social Stream* category, we determine the most common terms through a high-level visualisation of samples from the tweets collection. Accordingly, we provide a summary of the relevant themes in Figure 5 using *word cloud*[15].

3.5 Utility of the Datasets

Because this is the first large scale diverse collection of datasets in Hausa language, many relevant studies can be conducted using the corpus to complement existing resources. Thus, we identify relevant areas and research problems that could harness the curated datasets.

- **detection of online spurious content:** the increasing number of uncensored posts is posing many challenges to online socialisation, especially the problem of identifying genuine information in an ecosystem cluttered with spurious content. Despite the prevention measures taken by social media platforms, many sources of misleading information and rumours still exist, especially in low-resource languages such as Hausa. The collected corpus, especially the *Social Stream* category, could be used to develop a prediction system to ascertain content veracity.
- **propagation of fake news:** noting the prevalence of online fake news, it will be crucial to understand how such fake news and unfounded claims propagate in low resource languages, especially since the existing methods are based on high resource languages. The contributed datasets will be useful in investigating the problem.
- **low resource language model:** the diverse nature of the corpus will improve the training of useful state-of-the-art language models, such as BERT (5). While variants of such models have been used for multilingual setting, the data collection used for training is shallow and lacks diversity. Among other applications, the corpus can be used for downstream tasks such as machine translation, automatic speech recognition system, and sentiment analysis in the low resource language.

[15]the word cloud visualisation is based on the implementation at: https://github.com/amueller/word_cloud

Overall, the contribution of this study will significantly help in reducing the amount of time or effort in building tool-kits for downstream tasks in the low resource language.

4 CONCLUSION

Hausa language is one of the dominant Chadic languages that is widely spoken in Africa. From the computational linguistic point of view, Hausa, among other African languages, is categorised as low-resource language owing to the limited resources to handle various NLP-related tasks. This limitation slows down the pace of development of the language in terms of computational linguistic and for downstream tasks. In line with the existing studies, the study contributes a diverse Hausa corpus that will significantly help in reducing the amount of time in building tool-kits for various NLP downstream tasks in the language. We contributed the first comprehensive collection of curated datasets to support various NLP-related tasks in Hausa language. Essentially, the corpus consists of two categories according to collection sources: online social network and websites and blogs. Due to the collection sources, the themes in the datasets span many aspects of social life. Moreover, the inclusion of social media data will help in capturing the peculiarities in the language and how colloquial expressions manifest in the language. We provide a basic but useful analysis of the datasets and recommend some relevant problems that could leverage the corpus. We also presented a process to streamline the activity of obtaining more diverse datasets in Hausa from social media platforms, and how to retrieve the hydrated version of the data.

Future work It is pertinent to have a comprehensive collection and streamline the process to harness in enriching the language with the relevant resources. For future study, we hope this will attract research curiosity to further explore the language in-depth for supporting downstream tasks. Future focus will be on the development of a language model using the contributed datasets and other existing ones. The diverse nature of the corpus will improve the training of useful state-of-the-art language models, such as BERT (5). As pointed earlier in Section 3.5, the corpus can be used for downstream tasks such as machine translation, automatic speech recognition system, and sentiment analysis in the low resource language. Thus, it will be useful to develop a more diverse translation system that is equipped to support the informal Hausa content commonly found in online social media platforms such as Facebook and Twitter.

REFERENCES

[1] Idris Abdulmumin and Bashir Shehu Galadanci. hauwe: Hausa words embedding for natural language processing. In *2019 2nd International Conference of the IEEE Nigeria Computer Chapter (NigeriaComputConf)*, pages 1–6. IEEE, 2019.

[2] Tunde Adegbola. Building capacities in human language technology for african languages. In *Proceedings of the First Workshop on Language Technologies for African Languages*, pages 53–58, 2009.

[3] Željko Agic and Ivan Vulic. Jw300: A wide-coverage parallel corpus for low-resource languages. 2020.

[4] Adewale Akinfaderin. Hausamt v1. 0: Towards english-hausa neural machine translation. *arXiv preprint arXiv:2006.05014*, 2020.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Michael A Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. Transfer learning and distant supervision for multilingual transformer models: A study on african languages. *arXiv preprint arXiv:2010.03179*, 2020.

[7] Isa Inuwa-Dutse, Mark Liptrott, and Ioannis Korkontzelos. Detection of spam-posting accounts on twitter. *Neurocomputing*, 315:496–511, 2018.

[8] Philip J Jaggar. *Hausa*, volume 7. John Benjamins Publishing, 2001.

[9] Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*, 2020.

[10] Tim Schlippe, Edy Guevara Komgang Djomgang, Ngoc Thang Vu, Sebastian Ochs, and Tanja Schultz. Hausa large vocabulary continuous speech recognition. In *Spoken Language Technologies for Under-Resourced Languages*, 2012.

[11] Tanja Schultz. Globalphone: a multilingual speech and text database developed at karlsruhe university. In *Seventh International Conference on Spoken Language Processing*, 2002.

[12] Hydrator [Computer Software]. Documenting the now. *https://github.com/docnow/hydrator*, 2020.

APPENDIX

Details about the visited *urls* used in the *Website and Blog* category (Section 3.2) can be found at the project's Github repository available at https://github.com/ijdutse/hausa-corpus/tree/master/data/web-sites. Other relevant resources include the following:

- HugginFace: https://huggingface.co/datasets/hausa_voa_ner#curation-rationale
- Published Hausa Corpora: https://github.com/besacier/ALFFA_PUBLIC/tree/master/ASR/HAUSA
- Library of Congress: https://www.loc.gov/rr/pdf/requestingmaterials.pdf