Supported by ASDS

Proceedings of Digital Avenues for Low-Resource Languages of Sub-Saharan Africa (DASSA'2025)

Yaoundé, Cameroon, 27-28 May 2025

Edited by : Paulin Melatiaga[1] , Damien Nouvel[2] , Sarah Valentin[3]

[1] University of Yaoundé I, Yaoundé, Cameroon
[2] Inalco, France
[3] CIRAD, UMR TETIS, Montpellier, France

Table of Contents

Preface

The following papers were submitted for peer-reviewing and selected for oral presentation at the Digital Avenues for Low-Resource Languages of Sub-Saharan Africa 2025 (https://asds.africa/dassa2025/). This workshop, held at the University of Yaoundé I (Yaoundé, Cameroon), aimed to explore the intersections between linguistics, data science, machine learning, and language models to address the challenges of the under-representation of sub-Saharan languages. The workshop featured six keynote speakers, six oral presentations, and two roundtable discussions.

Préface

Les articles suivants ont été soumis et sélectionnés par un comité de lecture pour être présentés oralement lors de la conférence « Digital Avenues for Low-Resource Languages of Sub-Saharan Africa 2025 » (https://asds.africa/dassa2025/). Cet atelier, qui s'est tenu à l'Université de Yaoundé I (Yaoundé, Cameroun), avait pour objectif d'explorer les les approches à l'intersection entre la linguistique, la science des données, l'apprentissage automatique et les modèles de langue afin de relever les défis liés à la sous-représentation des langues subsahariennes. L'atelier comprenait six conférenciers principaux, six présentations orales et deux tables rondes.

Lexique médical numérique camerounais, langues locales (fulfulde et/ou ewondo)/langues officielles : étude processuelle de collecte de données

Richard Bertrand Etaba Onana

Université de Yaoundé II-ESSTIC, Yaoundé, Cameroun
ribeon777@gmail.com

Abstract. In our previous work, it was proven that, in Cameroonian primary and secondary referral hospitals, patient care faces language barriers. Some patients only speak their local language, which is often not understood by doctors. Others, on the other hand, are afraid to express certain realities of their illness because of the linguistic taboo. It has been proposed to establish a Cameroonian medical lexicon where one will find the scientific name / common name / local name / ethnic name of sacred objects and subjects and even of certain symptoms of diseases in order to overcome these communication deficits. One can start from the local name to the common name (in French and English) and to the scientific name (in French and English). Its implementation requires in-depth research with a view to establishing correspondences (phonetic and syntactic) between natural languages and specialized languages. The objective is not to make communication completely transparent, but rather to bring the doctor and the patient towards a convergence of referents and views. In the context of this workshop, the target languages are Fulfulde, which we have identified as the lingua franca of the East, North, Adamawa and Far North regions of Cameroon and / or Ewondo, a language spoken in the central region of Cameroon. Based on variationist sociolinguistics, the aim is to collect oral data with a view to setting up an application that can automatically generate all information relating to disease symptoms. The Speech-to-text or Speech -to- Speech system can be used. Concretely, this will involve bringing together native Fulfulde and Ewondo speakers in the different regions concerned and studying the realizations of certain sounds as well as their variations. The general alphabet of Cameroonian languages could be used as a basis.

Keywords: Cameroonian digital medical lexical · Local languages · Data collection.

1 Introduction

La prise en charge des patients dans les institutions hospitalières est sous-tendue par la capacité du patient à expliquer clairement son problème au médecin et aussi à celle du médecin à comprendre son patient. Il arrive souvent, lors des consultations médecins des situations assez complexes où le patient et son médecin

ne peuvent pas échanger suite à de nombreuses barrières dues à la langue. Parfois, les appartenances culturelles de certains patients les amènent à ne pas exprimer dans un langage clair certaines maladies dont ils souffrent, soit par la honte d'en parler soit à cause du tabou linguistique [1]. Toutes ces barrières linguistiques et culturelles rendent la communication difficile entre le soignant et le soigné et l'obtention d'un bon diagnostic passe par de multiples examens complémentaires qui entraînent d'énormes dépenses au pauvre patient. Nos recherches antérieures [2,3] ont montré que pour surmonter ces barrières, l'une des solutions est la traduction automatique de ces expressions utilisées par le patient en langage technique, compréhensible au médecin que nous appelons *Lexique médical numérique*, qui est une solution d'intelligence artificielle. Dans cette communication, nous appuyant sur la sociolinguistique variationniste et la théorie de régulations de Paul Zang Zang (2013) [4], nous expliquerons les fonctionnalités de cette application et les données linguistiques utiles pour sa mise en service. Avant d'y parvenir, disons un mot sur notre cadre théorique.

2 Cadre théorique

La théorie de base sur laquelle est assise notre réflexion est la théorie de régulation de Paul Zang Zang et le variotionnisme de Labov en est une théorie connexe.

La théorie variationniste de Labov [5] repose sur plusieurs piliers conceptuels essentiels. D'abord, elle considère que toute langue connaît des variantes linguistiques, c'est-à-dire des façons différentes de dire une même chose, qui coexistent dans un même espace social. Ces variantes concernent la prononciation, la syntaxe, le lexique ou encore la morphologie. Ensuite, Labov démontre que l'usage de ces variantes est corrélé à des variables sociales : l'âge, le sexe, la classe sociale, l'origine géographique ou encore le niveau d'instruction. Cette variation n'est pas une dégradation ou un dysfonctionnement du système linguistique : elle est au contraire constitutive de la langue vivante. La capacité à varier est une compétence sociolinguistique que tout locuteur possède et utilise stratégiquement.

La théorie des régulations, comme le définit Zang Zang, repose sur l'idée que la langue est un espace dynamique de tensions et de transformations, influencé par des forces sociales, culturelles et institutionnelles. La langue n'est pas un système statique, mais un champ d'interactions qui évolue constamment sous l'effet de ces forces. Ces forces de régulation peuvent être classées en cinq catégories majeures : institutionnelles (les pratiques institutionnelles à travers le pouvoir que détiennent le législateur et les institutions qui font usage de la langue), sociales (les pratiques effectives et les représentations que l'on se fait de la langue au plan social), culturelles (influence qu'exerce l'environnement culturel sur les conduites des individus et sur le fonctionnement du système langue), structurelles (concernent le corpus) et par rétroaction (désignent l'interaction entre les usagers de la même langue au niveau individuel, des groupes et des États). Ces 5 régulations

nous permettront de mieux établir les correspondances entre les différents choix linguistiques.

3 L'application Lexique numérique médical camerounais

Du point de vue sociolinguistique, c'est une application qui est destinée aux échanges entre médecins et patients lorsque ceux-ci ne parlent que leurs langues maternelles. Le modèle Speech -To-Speech (STS) ou Speech – To -Text (STT) sera utilisé. C'est une application qui sera beaucoup plus utilisée dans les hôpitaux de 1ère et 2ème références. Les données orales d'une langue locale A sont traduites en données orales d'une langue cible B et converties dans un vocabulaire technique. Le patient est en face d'un médecin, il explique en sa langue les symptômes de sa maladie. Un tableau clinique[1] est ainsi dressé. L'ensemble des éléments contenus dans un tableau clinique peuvent permettre au médecin de poser un diagnostic. Le patient peut les expliquer en sa langue native. Il est question de les rendre compréhensibles au soignant. On pourra donc avoir pour chaque maladie, l'ensemble des symptômes exprimés dans la langue du patient, traduits aussi en français et en anglais, le nom de la maladie correspondant dans un langage courant et le nom scientifique correspondant. Le tableau suivant en est une illustration :

[1] Dresser un tableau clinique d'un malade signifie, c'est *établir la liste de tous les symptômes et signes cliniques du patient. Les symptômes (ex : troubles du rythme cardiaque, difficultés à respirer, céphalées, rougeurs cutanées) peuvent varier beaucoup d'une maladie à l'autre, certains étant très caractéristiques de certaines pathologies. Le tableau clinique peut ainsi se révéler un précieux outil de diagnostic.*

Noms correspondants / Maladies	Nom courant	Nom local	Nom ethnique	Symptômes
Gonococcie	Blennorragie / gonorrhée	Chaude-pisse ou le tuyau est cassé	Megnolok (ewondo), [Transcription phonétique des différentes variations en fulfulde / ewondo]	Écoulement urétral, douleurs intenses au moment de la miction, sensations de picotements [transcription phonétique des différentes variations en fulfulde / ewondo]
Ophtalmomycose ou Ophtalmopathie	Conjonctivite	Appolo 12	Mefiat, apolo (ewondo) ; apolo (fulfulde) [transcription phonétique des différents variations en fulfulde / ewondo]	Rougeurs des yeux ; yeux qui collent [transcription phonétique des différentes variations en fulfulde / ewondo]
Paludisme	Malaria / paludisme	Paludisme	Nsong ntit/ ntit meki (ewondo) ; pabbohdjé (fulfulde) [transcription phonétique des différentes variations en fulfulde / ewondo]	Fièvre, anorexie, vomissement, maux de tête [transcription phonétique des différentes variations en fulfulde / ewondo]

Table 1. Projet du lexique numérique médical camerounais.

Nous avons dans la première colonne les noms scientifiques des pathologies, à la deuxième colonne des noms tels qu'ils sont connus dans le langage courant, la troisième colonne regroupe des traductions littérales de ces pathologies dans certaines localités du Cameroun et à la quatrième colonne on retrouve les symptômes exprimés en langues locales. Dans cette colonne aussi, l'on pourra aussi faire des transcriptions phonétiques des différentes prononciations possibles de ces symptômes en langue locale (fulfulde ou ewondo) par différents locuteurs.

4 Le processus de collecte de données pour la mise sur pied de l'application

Les données à recueillir ici sont les données orales du fulfulde et / ou de l'ewondo. Le patient s'exprime en une de ces langues et ses propos sont enregistrés via un dictaphone approprié. Avec l'aide des chercheurs en sciences numériques, il sera question de recueillir auprès des locuteurs natifs fulfulde ou ewondo, les symptômes des maladies. La collecte peut se faire dans les hôpitaux où ces patients sont reçus ou bien auprès des locuteurs natifs. Les aspects suivants seront pris en compte :

- Les variantes phonétiques : il sera enregistré les différentes réalisations du même son par des locuteurs natifs fulfulde et /ou ewondo. Par exemple pour le mot *pabbohdjé* qui signifie paludisme en fulfulde. Ce mot sera prononcé par divers locuteurs natifs fulfulde à l'effet d'identifier les différentes variations phonétiques que l'application pourra reconnaître.
- Les différentes appellations de cette pathologie dans ces langues-là ;
- Transcription phonétique des différents symptômes en fulfulde/ewondo ainsi que leurs variantes phonétiques ;
- La traduction en langue locale des symptômes des différentes pathologies.

L'objectif est de recueillir une base de données des symptômes en langues locales, de traduire ces symptômes des maladies en français ou anglais et de les regrouper par pathologie. Les noms desdites pathologies seront traduits en français et en anglais.
De manière pratique, la collecte suivra les étapes suivantes :
Étape 1 : Identification des pathologies les plus récurrentes dans les hôpitaux de district : au cours de cette étape, il est question de dresser une liste des pathologies rencontrées dans ces institutions hospitalières. On pourra de servir d'une recherche documentaire ou des entretiens avec des médecins des régions concernées. Si possible, trouver déjà l'équivalence de ces pathologies en français ou anglais courant et aussi un tableau de leurs symptômes.
Étape 2 : Identification des locuteurs natifs connaissant bien ces pathologies dans leur langue locale ainsi que leur manifestation. Au cours de cette étape, les locuteurs natifs pourront dire ces pathologies en leur langue ainsi que leurs symptômes.
Étape 3 : Constitution d'une équipe de personnels médicaux spécialisés dans la prise en charge des pathologies identifiées et si possible parlant le fulfulde et/ou

l'ewondo.
Étape 4 : Analyse phonétique des différentes prononciations desdites pathologies en langue locale. Le but ici est d'analyse comment les différentes maladies sont prononcées en langue locale. Pour y parvenir, les locuteurs natifs seront choisis en fonction de leur zone géographique. Le but est d'étudier les variantes au niveau des sons prononcés. Pour un mot, l'on aura les différentes prononciations.
Étape 5 : Analyse phonétique des différentes prononciations des symptômes desdites pathologies en langue locale. Le but ici est d'analyse comment les différents symptômes sont prononcés en langue locale. Pour y parvenir, les locuteurs natifs seront choisis en fonction de leur zone géographique. Le but est d'étudier les variantes au niveau des sons prononcés. Pour un mot, l'on aura les différentes prononciations.
Étape 6 : Faire des tableaux de correspondance qui pourra aider les spécialistes en sciences numériques à traduire ces données en langage informatique.
Étape 7 : Procédé à des vérifications avec le personnel médical à l'effet de valider les adéquations entre les symptômes et les pathologies ainsi répertoriées.

5 Résultats attendus

Obtention d'une base de données des différentes maladies comprenant :

- Les différentes appellations des maladies rencontrées en fulfulde et/ou en ewondo ;
- Les différentes prononciations possibles desdites maladies en langue fulfulde/ewondo ;
- Transcription phonétique de ces différentes prononciations ;
- La liste des différents symptômes de ces pathologies en langues fulfulde et /ou en ewondo ;
- Les différentes prononciations possibles des symptômes en fulfulde ou ewondo ;
- Transcription phonétique de ces différentes prononciations ;
- La traduction des différents symptômes en français ou en anglais ;
- La traduction de ces maladies en français courant ou anglais courant ;
- La traduction desdites maladies et leurs symptômes en langage médical.

6 Conclusion

Dans cette communication, il a été question de constituer une base de données dans la perspective de mettre sur pied une application pouvant aider le personnel médical dans la prise en charge des patients qui ne parlent que leurs locales. Nous avons jeté notre dévolu sur le fulfulde et l'ewondo, langue bantoue codée A70 dans la classification de Guthrie. Le protocole de collecte consistera donc à recueillir, auprès des locuteurs natifs (fulfulde ou ewondo) et du personnel

soignant des informations relatives sur les diverses pathologies et leurs équivalences en français ou anglais et aussi leurs dénominations scientifiques. Il en sera de même de leurs symptômes. Ce qui permettra sans doute aux spécialistes des sciences numériques de transformer ces données en langage informatique. Les langues choisies ne sont que des modèles pour tester l'application, elles peuvent s'étendre sur les autres langues locales du Cameroun.

References

1. Zang Zang, P., Etaba Onana, R. B., Les tabous linguistiques, mi-figue mi-raisin au cours des consultations médicales au Cameroun. Revue internationale d'études en langues modernes appliquées (10 /2017), 27-41(2017), lett.ubbcluj.ro/rielma/.
2. Zang Zang, P., Etaba Onana, R. B., Problèmes linguistiques dans les milieux hospitaliers au Cameroun : cas de l'Hôpital Général de Yaoundé et de l'Hôpital Gynéco-Obstétrique et Pédiatre de Yaoundé. Annales de la Faculté des Arts, Lettres et Sciences Humaines, Université de Yaoundé I (16), 139-165 (2014).
3. Etaba Onana, R. B., Mambo Tamnou, N. G., Communication interpersonnelle et interprétariat dans les institutions hospitalières publiques camerounaises. Revue internationale d'études en langues modernes appliquées (14/2021), 42-52(2021), lett.ubbcluj.ro/rielma/.
4. Zang Zang, P., Linguistique et émergence des nations : essai d'aménagement d'un cadre théorique, Lincom Europa, Munich (2013).
5. Labov, W., Sociolinguistique, Paris, Édition de Minuit (1976).

AbuseBERT-WoFr: refined BERT model for detecting abusive messages on tweets mixing Wolof-French codes

Ibrahima Ndao[1,*], Khadim Dramé[1], Gorgoumack Sambe[2] and Gayo Diallo[3]

[1]Assane Seck University of ziguinchor, Computer Science and Engineering Laboratory for Innovation, Ziguinchor, 27001, Senegal

[2]Cheikh Hamidou Kane Digital University, Dakar, 10000, Senegal

[3]Bordeaux Population Health Inserm 1219 & LaBRI, Univ. Bordeaux, F-33000, Bordeaux, France

i.ndao20150570@zig.univ.sn ; khadim.drame@univ-zig.sn ; gorgoumack.sambe@unchk.edu.sn ; diallo.gayo@gmail.com

Abstract. Pre-trained models have proved effective for many tasks in automatic language processing, but are limited when subjected to mixed-code data. The use of two or more languages in the same sentence is known as mixed code. This communication habit is more common in social networks. The latter are subject to increasing waves of abusive messages, requiring effective measures particularly in low-resource languages. In this article, we present AbuseBERT-WoFr, the first model for abusive message detection on tweets of mixed Wolof-French codes. This model is trained on a large dataset of 144225 mixed-code tweets collected from twitter. We evaluate the model's performance on an abusive message detection task on a corpus of 2022 tweets annotated with Wolof-French mixed codes, and compare its results with state-of-the-art language models. The results are highly competitive.

Keywords: abusive tweets, low-resource languages, code mixing, pre-trained models, natural language processing (NLP), machine learning (ML), deep learning (DL)

1 Introduction

Platforms like Twitter are gaining in popularity every day. They offer users great freedom to express their opinions. This freedom is leading to a growing proliferation of abusive messages. So, to maintain harmony on social networks, coupled with the growing number of users, automatic and effective moderation is needed. This article contributes to the search for solutions to this problem.

An abusive message is one that is excessive or attacks one or more people on the basis of characteristics such as race, ethnicity, sexual orientation, etc. The detection of abusive messages is considered as a classification task consisting in classifying a tweet

into a well-defined class [1]. Among the classes studied are: hate [2] [3], offensive [4] [5], cyberbuilding [6] [7], racist [8] [9], sexist [10] [11], abusive [12] [13], etc. The choice of classifier is crucial to training an effective model. However, researchers have been interested in the use of classical classifiers [14] [8] [15] [16], Deep Learning(DL) models [17] [18] and pre-trained models [19] [20]. Recently, the trend has been towards refining pre-trained models on other languages [21] [22] [23] or other domains [24] [3]. Most of this research has focused more on high-resource languages such as English [25] [1], Arabic [26] [27], French [28], etc. This field is becoming more and more difficult. Moreover, this field becomes more difficult when dealing with mixed code texts.

Code mixing is the mixing of several languages in a message or communication [21]. Numerous studies have focused on this aspect, such as Hindi-English [23] [29], English-Urdu [30], and so on. The use of code-mixing is increasing in social networking texts, particularly in the tweets of Senegalese. Senegalese make up a large online community, where the use of code-mixing between Wolof and French is particularly noticeable. Like all users or communities of users, it is not immune to the upsurge in abusive content. This was particularly noticeable during the numerous protests between 2021 and 2023, which led to restrictions on social networks. However, a great deal of research has focused on the detection of abusive messages in mixed-code tweets [12] [23], but to our knowledge little or no research has focused on Wolof-French mixed-code tweets [31]. We are collecting a large dataset from Twitter that we will make available to researchers. We present an annotated data subset of 2022 Wolof-French mixed-code tweets from the collected data. The remainder of the unannotated dataset is used to train our AbuseBERT-WoFr model. We demonstrate the effectiveness of AbuseBERT-WoFr on this annotated dataset and on the downstream task.

2 Related work

Due to cultural diversity on social networks, the phenomenon of code-mixing is on the increase. The resurgence of this phenomenon has prompted interest in developing models capable of overcoming this problem to improve the detection of abusive messages in social networks.

Early research focused on building specialized corpora to facilitate the implementation of effective methods. Thus, [32] produced an annotated Hindi-English code-mix dataset to study aggression on Twitter and Facebook. The corpus consists of 18,000 Twitter tweets and 21,000 Facebook comments. They define a set of hierarchical tags for annotation: three top-level tags and ten level-2 tags. [33] presents the first Hindi-English code-mix annotated corpus on sarcasm and irony detection. This corpus consists of 5250 tweets, 504 of which were annotated as sarcastic or ironic. [34] produced a dataset of 4575 tweets of mixed Hindi-English code for hate detection. A corpus of mixed Hindi-English code for humor detection is presented in [35]. Meanwhile, [12] produces EkoHate, a dataset of mixed English, Yoruba and Naija codes. It consists of 3398 tweets for the detection of abusive language and political hate speech.

These data sets are useful for building robust models. Several classifiers have been explored in the literature. For the identification of hate speech on Hindi-Angali code-mix tweets, [36] shows the effectiveness of FastText embeddings over Word2Vec and Doc2vVec. He uses the FastText model refined on 10000 tweets to provide vectorizations to an SVM and an RBF model. It also demonstrates that character-level embeddings can improve results on code-mix data. [37] studies the detection of abusive messages on Hindi-English code-mix tweets by comparing three deep learning models: a one-dimensional CNN model, an LSTM model and a Bi-LSTM model. Specific word extensions were trained on the corpus of code mixture data used. This study showed that the use of domain-specific word integrations would help improve model performance. [38] uses two LSTM architectures to identify hate speech on Hindi-English texts. The first LSTM model uses subword-level embeddings while the second uses an attention mechanism based on phonemic subwords. [29] shows the importance of data pre-processing for the detection of hate speech. It performs a ten-step data cleaning and then trains two classifiers: a CNN model and a CNN-LSTM model combination. The CNN-LSTM model resulted in a 2-point increase in accuracy and a reduction in runtime.

New studies have focused on the use or adaptation of pre-trained models for the detection of abusive messages on code-mix data. [23] conducts a comparative study on several pre-trained models for the detection of hate messages on Hindi-English tweets. It uses four pre-trained models on code-mixed or multilingual data: HingBERT, HingRoBERTa, HingRoBERTa-Mixed and mBERT; and three pre-trained models on unmixed data: AlBERT, BERT, RoBERTa. Evaluation results showed that the HingBERT model trained specifically on code-mixing tweets outperformed all the others. This study identified the weaknesses of pre-trained models such as BERT on code-mix data. [39] studies the impact of language identification on the detection of hate messages in Hindi-English code-mix tweets. It defines a data preprocessing and language identification pipeline. It uses word-level extensions and an end-of-sentence linguistic information bin to train BERT models and refined BERT models on Hindi-English tweets. These experiments show an increase in the performance of all models and the importance of having a substantial code-mix dataset. [40] trains four pre-trained models on code-mix data, notably Hindi-English and English-Slovene. It compares their performance against several other mono- and multilingual models. However, the bilingual models performed best.

The language models perform well on several tasks in automatic language processing, notably the detection of abusive messages. But they fail on social network texts (Tweets), especially if several languages are used in the message [21]. What's more, they don't cover low-resource languages such as Wolof.

3 Methodologies

Pre-trained models have demonstrated their effectiveness on several NLP tasks. Training or adapting these models requires a large corpus of carefully processed data. Acquiring such a corpus is not an easy task, especially when dealing with data from

mixtures of Wolof-French codes. To our knowledge, little or no such data exists. In response to this lack, we are producing one of the first data sets of mixed Wolof-French codes.

3.1 Collection and pre-treatment

Collecting a large set of data is tedious work. This difficulty is exacerbated when dealing with low-resource languages. So, to set up these datasets, we're using the Twint[1] python library. This tool is a scraper of tweets from twitter. The choice of this tool is motivated by its ease of use and its ability to retrieve tweets without restriction. The following figure shows an example of the code used for scraping.

```
[ ]  import twint
     import csv

     # Configuration de la recherche
     c = twint.Config()
     c.Search = "Sénégal"
     #c.Lang = "fr"  # Limite les résultats aux tweets en français
     c.Since = "2021-01-01"  # Date de début de la recherche (YYYY-MM-DD)
     c.Until = "2023-05-31"  # Date de début de la recherche (YYYY-MM-DD)
     c.Store_csv = True  # Active la sauvegarde des données au format CSV
     c.Output = "/content/tweets_senegal.csv"  # Chemin du fichier de sortie CSV

     # Exécution de la recherche
     twint.run.Search(c)
```

Fig. 1. Example of scrapping code

Several queries were launched including numerous metadata such as names of influential people, country names, keywords, languages, etc. We collected a large set of tweets from Twitter over the period January 1, 2021 to May 31, 2023. The total data collection is around 150k tweets and is stored on a csv file.
Faced with the need for a quality corpus, the data were pre-processed to reduce noise and focus on the Wolof-French code-mix aspect.
A series of pre-processing steps were carried out to remove certain information that was not relevant to the downstream task, such as URLs, emojis, mentions and so on.
The ability of pre-trained language models to capture new linguistic features relevant to other domains or languages makes them more popular. For example, we are adapting the BERT pre-trained model on tweets of Wolof-French code mixtures to detect abusive messages.

[1] https://github.com/twintproject/twint

3.2 Model training

Before refining the BERT model on our corpus, we selected a subset of 2022 tweets from our datasets for annotation. The selection was based mainly on the code-mix aspect, in other words, we selected 2022 tweets where there was the presence of a mix of Wolof and French terms. These were meticulously annotated into two distinct classes: the non-abusive class noted "0" and the abusive class noted "1". The rest of the dataset is used to refine the BERT-base-uncased[2] model. We used the hugginface library to load the original BERT model and the Pytorch library for fitting. The principle of this adjustment is contextual learning to enable the BERT model to learn the relationships between (cross-linguistic) tokens. The resulting model is a BERT model refined on tweets data of Wolof-French code mixes called AbuseBERT-WoFr.

4 Experiments and results

We use our AbuseBERT-WoFr model to train it on an abusive message detection task using the annotated dataset of 2022 annotated tweets. It consists of 950 'abusive' tweets and 1072 'non-abusive' tweets. A fairly balanced distribution of classes compared with existing corpora in the literature, which are mostly unbalanced [24]. This dataset is subdivided into 80% for training and 20% for testing on the downstream task. We trained our model over 10 epochs with a batch size of 32, a learning rate of 2e-5 and the Adam optimizer.

We apply the same configuration to the other language models. The models are evaluated in terms of precision, recall, f-measure per class and accuracy. The evaluation results confirm the effectiveness of the approach. We compare the performance of AbuseBERT-WoFr with that of other models on the annotated corpus. The evaluation results are presented in Table 1.

Table 1. Results of the experiment

Models	Precision		Recall		F-measure		Accuracy
	0	1	0	1	0	1	
BERT-base-uncased	0.75	0.49	0.46	0.60	0.58	0.55	0.60
BERT-based-cased	0.72	0.52	0.45	0.59	0.57	0.54	0.59
Camembert	**0.80**	0.40	0.42	0.57	0.56	0.51	0.54
Davlan/Bert	0.76	0.51	0.58	0.45	0.68	0.48	0.62
RoBERTa	0.77	0.52	0.55	0.54	0.59	0.53	0.58
AbuseBERT-WoFr	0.73	**0.60**	**0.65**	**0.68**	**0.69**	**0.64**	**0.69**

The AbuseBERT-WoFr model outperforms all models in terms of precision, recall and f-measure for class 1 'abusive' and in terms of recall and f-measure on class 0 'not abusive'. This demonstrates the effectiveness of AbuseBERT-WoFr in capturing

[2] https://huggingface.co/google-bert/bert-base-uncased

nuances in the language mix and detecting abusive messages. The results show better performance in both classes, particularly in the abusive class, demonstrating the importance of BERT model pre-training on code-mix tweets. This finding confirms expectations of further improvement in this aspect of code-mixing.

Our AbuseBERT-WoFr model has two special features: the first is its ability to model the linguistic variability of the Wolof-French code mix: its training on a large dataset of tweets; and the second is its application domain: the detection of abusive messages.

5 Conclusion and future work

This paper presents AbuseBERT-WoFr, a model refined on the BERT base uncased model for abusive message detection on tweets of Wolof-French code mixtures. We produce a downstream task dataset and have annotated a subset for validation of AbuseBERT-WoFr. The results show that AbuseBERT-WoFr is more efficient and robust than language models on this task. Although we have focused on the detection of abusive messages, and our compelling results confirm empirical evidence, our future work will be directed not only at expanding the annotated dataset while focusing on fine-grained annotation, but also at exploring new sentence integration models for language boundary identification. In addition, we intend to test our model against mixed-code models.

References

1. C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, et Y. Chang, « Abusive Language Detection in Online User Content », in Proceedings of the 25th International Conference on World Wide Web, Montréal Québec Canada: International World Wide Web Conferences Steering Committee, avr. 2016, p. 145-153. doi: 10.1145/2872427.2883062.
2. Anjum et R. Katarya, « Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities », Int. J. Inf. Secur., vol. 23, no 1, p. 577-608, févr. 2024, doi: 10.1007/s10207-023-00755-2.
3. K. Mnassri, P. Rajapaksha, R. Farahbakhsh, et N. Crespi, « BERT-based Ensemble Approaches for Hate Speech Detection », 15 septembre 2022, arXiv: arXiv:2209.06505. Consulté le: 18 mars 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2209.06505
4. A. Joshi et R. Joshi, « Harnessing Pre-Trained Sentence Transformers for Offensive Language Detection in Indian Languages », 3 octobre 2023, arXiv: arXiv:2310.02249. Consulté le: 13 octobre 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2310.02249
5. N. Hamad, M. Jarrar, M. Khalilia, et N. Nashif, « Offensive Hebrew Corpus and Detection using BERT », in 2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA), Giza, Egypt: IEEE, déc. 2023, p. 1-8. doi: 10.1109/AICCSA59173.2023.10479258.
6. C. Perez et S. Karmakar, « An NLP-Assisted Bayesian Time Series Analysis for Prevalence of Twitter Cyberbullying During the COVID-19 Pandemic », 28 février 2023, arXiv: arXiv:2208.04980. Consulté le: 18 mars 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2208.04980

7. M. S. Akter, H. Shahriar, et A. Cuzzocrea, « A Trustable LSTM-Autoencoder Network for Cyberbullying Detection on Social Media Using Synthetic Data », 2023.
8. I. Kwok et Y. Wang, « Locate the Hate: Detecting Tweets against Blacks », Proc. AAAI Conf. Artif. Intell., vol. 27, no 1, p. 1621-1622, juin 2013, doi: 10.1609/aaai.v27i1.8539.
9. C. Arcila-Calderón, J. J. Amores, P. Sánchez-Holgado, L. Vrysis, N. Vryzas, et M. Oller Alonso, « How to Detect Online Hate towards Migrants and Refugees? Developing and Evaluating a Classifier of Racist and Xenophobic Hate Speech Using Shallow and Deep Learning », Sustainability, vol. 14, no 20, p. 13094, oct. 2022, doi: 10.3390/su142013094.
10. Y. Zhang et L. Wang, « HHS at SemEval-2023 Task 10: A Comparative Analysis of Sexism Detection Based on the RoBERTa Model », in Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), A. Kr. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, et E. Sartori, Éd., Toronto, Canada: Association for Computational Linguistics, juill. 2023, p. 963-968. doi: 10.18653/v1/2023.semeval-1.133.
11. D. Grosz et P. Conde-Cespedes, « Automatic Detection of Sexist Statements Commonly Used at the Workplace », in Trends and Applications in Knowledge Discovery and Data Mining, vol. 12237, W. Lu et K. Q. Zhu, Éd., in Lecture Notes in Computer Science, vol. 12237. , Cham: Springer International Publishing, 2020, p. 104-115. doi: 10.1007/978-3-030-60470-7_11.
12. C. E. Ilevbare, J. O. Alabi, D. I. Adelani, F. D. Bakare, O. B. Abiola, et O. A. Adeyemo, « EkoHate: Abusive Language and Hate Speech Detection for Code-switched Political Discussions on Nigerian Twitter », 28 avril 2024, arXiv: arXiv:2404.18180. Consulté le: 13 mai 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2404.18180
13. N. Cécillon, « Exploration de descripteurs de plongements de graphes pour la détection de messages abusifs », juill. 2019, Consulté le: 17 juillet 2023. [En ligne]. Disponible sur: https://dumas.ccsd.cnrs.fr/dumas-04073337
14. Z. Waseem, « Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter », in Proceedings of the First Workshop on NLP and Computational Social Science, Austin, Texas: Association for Computational Linguistics, nov. 2016, p. 138-142. doi: 10.18653/v1/W16-5618.
15. P. Burnap et M. L. Williams, « Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making », Policy Internet, vol. 7, no 2, p. 223-242, 2015, doi: 10.1002/poi3.85.
16. T. Davidson, D. Warmsley, M. Macy, et I. Weber, « Automated Hate Speech Detection and the Problem of Offensive Language », 11 mars 2017, arXiv: arXiv:1703.04009. Consulté le: 22 novembre 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1703.04009
17. B. Gambäck et U. K. Sikdar, « Using Convolutional Neural Networks to Classify Hate-Speech », in Proceedings of the First Workshop on Abusive Language Online, Vancouver, BC, Canada: Association for Computational Linguistics, août 2017, p. 85-90. doi: 10.18653/v1/W17-3013.
18. Z. Zhang, D. Robinson, et J. Tepper, « Hate Speech Detection Using a Convolution-LSTM Based Deep Neural Network », 2018.
19. A. G. D'Sa, I. Illina, et D. Fohr, « BERT and fastText Embeddings for Automatic Detection of Toxic Speech », in 2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies" (OCTA), Tunis, Tunisia: IEEE, févr. 2020, p. 1-5. doi: 10.1109/OCTA49274.2020.9151853.
20. S. Kalra, K. N. Inani, Y. Sharma, et G. S. Chauhan, « Applying Transfer Learning using BERT-based models for Hate Speech Detection », 2021.

21. R. Nayak et R. Joshi, « L3Cube-HingCorpus and HingBERT: A Code Mixed Hindi-English Dataset and BERT Language Models », 18 avril 2022, arXiv: arXiv:2204.08398. Consulté le: 20 mai 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2204.08398
22. T. Chavan, S. Patankar, A. Kane, O. Gokhale, et R. Joshi, « A Twitter BERT Approach for Offensive Language Detection in Marathi », 20 décembre 2022, arXiv: arXiv:2212.10039. Consulté le: 17 mars 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2212.10039
23. A. Patil, V. Patwardhan, A. Phaltankar, G. Takawane, et R. Joshi, « Comparative Study of Pre-Trained BERT Models for Code-Mixed Hindi-English Data », in 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), avr. 2023, p. 1-7. doi: 10.1109/I2CT57861.2023.10126273.
24. T. Caselli, V. Basile, J. Mitrović, et M. Granitzer, « HateBERT: Retraining BERT for Abusive Language Detection in English », 4 février 2021, arXiv: arXiv:2010.12472. Consulté le: 15 novembre 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/2010.12472
25. Z. Waseem, T. Davidson, D. Warmsley, et I. Weber, « Understanding Abuse: A Typology of Abusive Language Detection Subtasks », 30 mai 2017, arXiv: arXiv:1705.09899. Consulté le: 28 décembre 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1705.09899
26. H. Mubarak, A. Rashed, K. Darwish, Y. Samih, et A. Abdelali, « Arabic Offensive Language on Twitter: Analysis and Experiments », undefined, 2021, Consulté le: 22 novembre 2022. [En ligne]. Disponible sur: https://www.semanticscholar.org/reader/1e3170311bf21a70b6f974e40b8932fbc0f3052b
27. I. Bensalem, M. Mout, et P. Rosso, « Offensive Language Detection in Arabizi », in Proceedings of ArabicNLP 2023, H. Sawaf, S. El-Beltagy, W. Zaghouani, W. Magdy, A. Abdelali, N. Tomeh, I. Abu Farha, N. Habash, S. Khalifa, A. Keleg, H. Haddad, I. Zitouni, K. Mrini, et R. Almatham, Éd., Singapore (Hybrid): Association for Computational Linguistics, déc. 2023, p. 423-434. doi: 10.18653/v1/2023.arabicnlp-1.36.
28. N. Cecillon, R. Dufour, V. Labatut, et G. Linares, « Tuning Graph2vec with Node Labels for Abuse Detection in Online Conversations », in 11ème Conférence Modèles & Analyse de Réseaux : approches mathématiques et informatiques (MARAMI), Montpellier, France, oct. 2020. Consulté le: 13 novembre 2022. [En ligne]. Disponible sur: https://hal.archives-ouvertes.fr/hal-02993571
29. K. Al-Hussaeni, M. Sameer, et I. Karamitsos, « The Impact of Data Pre-Processing on Hate Speech Detection in a Mix of English and Hindi–English (Code-Mixed) Tweets », Appl. Sci., vol. 13, no 19, Art. no 19, janv. 2023, doi: 10.3390/app131911104.
30. G. I. Ahmad et J. Singla, « (LISACMT) Language Identification and Sentiment analysis of English-Urdu 'code-mixed' text using LSTM », in 2022 International Conference on Inventive Computation Technologies (ICICT), juill. 2022, p. 430-435. doi: 10.1109/ICICT54344.2022.9850505.
31. D. Mbaye, M. Diallo, et T. I. Diop, « Low-Resourced Machine Translation for Senegalese Wolof Language », 2023, doi: 10.48550/ARXIV.2305.00606.
32. R. Kumar, A. N. Reganti, A. Bhatia, et T. Maheshwari, « Aggression-annotated Corpus of Hindi-English Code-mixed Data », 2018.
33. S. Swami, A. Khandelwal, V. Singh, S. S. Akhtar, et M. Shrivastava, « A Corpus of English-Hindi Code-Mixed Tweets for Sarcasm Detection », 30 mai 2018, arXiv: arXiv:1805.11869. Consulté le: 24 décembre 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/1805.11869
34. A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, et M. Shrivastava, « A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection », in Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, M. Nissim, V. Patti, B. Plank, et C. Wagner, Éd., New Orleans, Louisiana,

USA: Association for Computational Linguistics, juin 2018, p. 36-41. doi: 10.18653/v1/W18-1105.

35. A. Khandelwal, S. Swami, S. S. Akhtar, et M. Shrivastava, « Humor Detection in English-Hindi Code-Mixed Social Media Content : Corpus and Baseline System », 14 juin 2018, arXiv: arXiv:1806.05513. Consulté le: 24 décembre 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/1806.05513
36. K. Sreelakshmi, B. Premjith, et K. P. Soman, « Detection of Hate Speech Text in Hindi-English Code-mixed Data », Procedia Comput. Sci., vol. 171, p. 737-744, 2020, doi: 10.1016/j.procs.2020.04.080.
37. S. Kamble et A. Joshi, « Hate Speech Detection from Code-mixed Hindi-English Tweets Using Deep Learning Models », 13 novembre 2018, arXiv: arXiv:1811.05145. Consulté le: 27 décembre 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/1811.05145
38. T. Y. S. S. Santosh et K. V. S. Aravind, « Hate Speech Detection in Hindi-English Code-Mixed Social Media Text », in Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, Kolkata India: ACM, janv. 2019, p. 310-313. doi: 10.1145/3297001.3297048.
39. G. Takawane, A. Phaltankar, V. Patwardhan, A. Patil, R. Joshi, et M. S. Takalikar, « Leveraging Language Identification to Enhance Code-Mixed Text Classification », 8 juin 2023, arXiv: arXiv:2306.04964.
40. S. Yadav, A. Kaushik, et K. McDaid, « Leveraging Weakly Annotated Data for Hate Speech Detection in Code-Mixed Hinglish: A Feasibility-Driven Transfer Learning Approach with Large Language Models », 4 mars 2024, arXiv: arXiv:2403.02121. Consulté le: 10 août 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2403.02121

Diversity Text Generation via Adversarial Network in low-resource languages

Charnelle YANHAMO[1 [0000−1111−2222−3333]] and Norbert TSOPZE[2,3 [1111−2222−3333−4444]]

University of Yaounde 1, Computer Science Department
charnelle.yanhamo@facsciences-uy1.cm, tsopze.norbert@gmail.com
https://facsciences.uy1.cm/

Abstract. Automatic text generation is a technique that enables the production of new texts from existing ones by leveraging machine learning tools. It plays a crucial role in fields such as machine translation and data augmentation. Among the most commonly used approaches for this task, generative adversarial networks (GAN) stand out due to their remarkable performance in data generation. Low resource languages such as Ghomala' could benefit from GAN models for generating text. In this paper, we introduce the DTextGAN (Diversity Text Generation via Adversarial Network in low resource languages) model, designed to generate texts that are both realistic and diverse. The adopted approach is based on a hierarchical generator and a specific reward calculation method, aimed at promoting diversity in the generated texts. By assigning low rewards to less diverse generations and leveraging the hierarchical structure, our method encourages the creation of coherent, richer, and more varied content. Experiments were conducted on a local dataset consisting of sentences in Ghomala' (a local language of Cameroon), comparing our approach against four state-of-the-art GAN-based text generation models. The results demonstrate that our model achieves the best balance between linguistic quality (BLEU-4: 0.86) and diversity (8% repetition rate), outperforming baselines that either generate incoherent content or exhibit high repetition. Expert human evaluation confirms that generated sentences achieve equivalent quality to real sentences across diversity, fluency, and relevance criteria, thereby illustrating the effectiveness of our approach in specific and varied linguistic contexts.

Keywords: Text Generation · Low-Resource Languages · Generative Adversarial Networks · Reinforcement Learning

1 Introduction

Recent advancements in data processing have enabled the development of models capable of extracting knowledge from large datasets. However, in the context of low-resource languages, where corpora are limited, this requirement presents a major challenge. Approaches for generating synthetic data [6] aim to overcome this issue by generating new samples from existing data, making them suitable

for underrepresented languages. However, text generation remains complex due to the grammatical and syntactic constraints inherent to each language [5].

Generative Adversarial Networks (GANs)[1], originally designed for continuous data like images, have been adapted to discrete data, notably text. Their ability to generate realistic samples from limited data is particularly relevant for low-resource languages. Approaches combining GANs with reinforcement learning, such as RankGAN[4] and LeakGAN [2], have addressed the challenges posed by the discrete nature of text.

In this work, we explore text generation in low-resource languages using a GAN-based architecture, integrating a hierarchical generator coupled with a discriminator based on cross-entropy. This approach optimizes the generation of coherent and diverse texts, even with limited corpora. Our main contributions are: (1) an adaptation of hierarchical GAN architecture for low-resource language text generation, (2) a discriminator that promotes diversity through entropy-based rewards, and (3) comprehensive experimental validation against existing GAN-based approaches with expert human evaluation.

2 Related Work and Background

2.1 Low-Resource Languages Challenges

Low-resource languages, defined as languages with limited digital corpora and computational resources, face significant challenges in natural language processing applications. These languages, spoken by millions worldwide, often lack the extensive datasets required for training robust language models. Traditional approaches that work well for high-resource languages like English struggle with the data scarcity typical of low-resource scenarios. This digital divide creates barriers to technological inclusion and cultural preservation. Synthetic data generation emerges as a promising solution to bridge this gap by augmenting limited corpora with artificially generated text that maintains linguistic authenticity while expanding dataset size.

2.2 GAN-based Text Generation

Text generation consists of automatically producing texts that resemble those written by humans, relying on machine learning models known as language models. The probability of generating a token w_t, given the previous n tokens, is given by $P(w_t|w_{t-1}, w_{t-2}, \ldots, w_{t-n})$. The text generation process involves: (1) input of the initial sequence, (2) probability distribution calculation, (3) selection of the next word, and (4) updating the sequence.

Generative Adversarial Networks[1] have been adapted for text generation despite the discrete nature challenge. When applied to text, GANs are reformulated as reinforcement learning problems [7], where the generator acts as an

agent constructing sequences incrementally. The goal is to optimize the generator's policy $\pi(a|s,\theta)$ to maximize expected reward $J(\theta)$:

$$\theta = \theta + \alpha \nabla J(\theta) = \theta + \alpha E_{\pi_\theta} \left[\sum_{t=0}^{T-1} R(s_t, a_t) \right] \tag{1}$$

Several notable approaches have emerged: SeqGAN pioneered the application of GANs to sequence generation using policy gradient methods. RankGAN [4] employed ranking-based rewards instead of binary classification. LeakGAN [2] introduced hierarchical reinforcement learning with a manager-worker architecture for long text generation. More work includes and GSGAN[3] applying Gumbel-Softmax for differentiable sampling and TextGAN[8] employs feature matching for stable training. However, existing GAN approaches often prioritize realism over diversity, leading to mode collapse and repetitive outputs—particularly problematic for low-resource languages where corpus expansion requires maximum lexical and syntactic variety.

3 Proposed Architecture

We propose DTextGAN, an architecture based on a hierarchical generator coupled with a discriminator based on cross-entropy. The main goal is to leverage the generator's ability to produce structured sequences while using a reward signal that emphasizes semantic diversity in low-resource contexts. Figure 1 provides an overview of our text generation framework

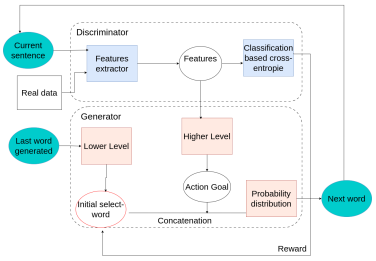

Fig. 1. Illustration of DTextGAN

Hierarchical Generator
Building on the approach by Guo et al. [2], we propose a generator composed of two levels:

– The **manager** (higher level) is responsible for defining semantic sub-goals from latent representations. It receives information from the discriminator and conveys it to the worker in the form of objectives at each time step. This helps address the issues related to the non-informativeness and scarcity of reward signals in classical GANs.

– The **worker** (lower level) executes these sub-goals, generating the tokens that compose the text, one sequence at a time, ensuring that each sub-goal is respected within the generation process.

This hierarchical structure enhances the lexical and syntactical diversity of the generated text, which is crucial for low-resource languages where corpora are often limited.

Discriminator based on Cross-Entropy

Our main methodological contribution lies in the discriminator design. Unlike binary discriminators in classical GANs, our discriminator assigns continuous reward based on the divergence between the distribution of generated words and that of real texts, measured through cross-entropy. The reward is computed at both the word level and the sentence level. Formally, let $Y_{1:T} = (y_1, ..., y_t, ..., Y_T)$, with $y_t \in \Lambda$, Λ being the set of generated sentences, and $y_t = (y_{t,1}, ..., y_{t,K})$ representing the t^{th} sentence, where $y_{t,k}$ is the k^{th} word of the t^{th} sentence. The reward for $y_{t,k}$ is defined as:

$$R(y_{t,k}) = -\log D_\phi(y_{t,k}) \tag{2}$$

And for sentence y_t::

$$R(y_t) = -\frac{1}{K}\sum_{k=1}^{K} \log D_\phi(y_{t,k}) \tag{3}$$

where $D_\phi(y_{t,k})$ is the output of the discriminator for the k^{th} word of the t^{th} sentence. Entropy plays a crucial role in evaluating the lexical diversity of sequences. Low-entropy sequences, often synonymous with repetition or excessive predictability, receive a low reward. In contrast, sequences with higher entropy, representing greater diversity, are rewarded.

4 Experiments and Results

Dataset and Experimental Setup: Our experiments use a corpus of Ghomala' sentences, a local language from Cameroon. Due to the severe limitation of resources for this low-resource language, the Bible translation was the only accessible substantial text corpus in Ghomala', making it our sole viable data source. We sampled 2000 short sentences (2-10 words) for initial training, following a three-step process: (1) pre-training the generator, (2) training the discriminator with cross-entropy evaluation, (3) reinforcement learning optimization. We generated 3000 sentences for evaluation.

For comparison, we evaluated DTextGAN against four representative GAN-based text generation models: GSGAN, TextGAN, RankGAN, and LeakGAN, allowing comprehensive assessment of our diversity-oriented approach.

Evaluation Metrics: We evaluated using: **BLEU Score** (n-gram overlap with reference texts), **Coverage Rate** (overlap between real and generated sentences), **Sentence Repetition Rate** (proportion of repeated sentences), and

Word Repetition Rate (consecutive word repetitions). For effective data augmentation, we seek high BLEU scores (indicating linguistic quality) combined with low repetition rates (indicating diversity):balancing coherence with variety.

Results: Tables 1 and 2 show the comparison with baseline GAN models for 3000 generated sentences.

Table 1. BLEU Scores Comparison

Model	BLEU-2	BLEU-3	BLEU-4
GSGAN	0.08	0.06	0.05
TextGAN	0.65	0.09	0.08
RankGAN	0.69	0.40	0.25
LeakGAN	0.87	0.71	0.56
DTextGAN	**0.86**	**0.69**	**0.55**

Table 2. Diversity Metrics (%)

Model	Coverage Rate	Sentence Rep. Rate	Word Rep. Rate
GSGAN	0	99.96	100
TextGAN	13	21	8
RankGAN	19	30	4
LeakGAN	12	13	7
DTextGAN	**7**	**8**	**10**

Analysis: Table 1 and Table 2 reveal important insights about the relationship between BLEU scores and diversity. Models with very low BLEU scores (GSGAN: 0.05, TextGAN: 0.08) exhibit extremely high repetition rates (99.96%, 21%), indicating they generate repetitive or incoherent content rather than diverse text. In contrast, DTextGAN achieves both high BLEU score (0.86) and the lowest repetition rate (8%), demonstrating effective generation of linguistically coherent and diverse content. This shows that meaningful diversity requires maintaining linguistic quality:a crucial balance for low-resource language data augmentation

Human Evaluation: To assess the quality of generated text, we conducted a blind human evaluation. We randomly sampled 20 sentences from both the real corpus and DTextGAN-generated sentences. A linguistic expert specialized in Ghomala' language evaluated each sentence across three criteria: diversity, fluency, and relevance. The evaluator had no knowledge of the data source. Generated sentences achieved equivalent scores to real sentences across all three

criteria, confirming that our model produces coherent and linguistically valid text while maintaining the desired diversity.

5 Conclusion

We proposed DTextGAN, a GAN-based framework for generating coherent and diverse text in low-resource languages. Our hierarchical generator with cross-entropy discriminator successfully generates original sentences without copying training data, evidenced by low overlap and repetition rates. Comparative evaluation against four baseline GAN models demonstrates superior diversity performance, confirmed by expert human evaluation showing equivalent quality to real sentences. **Limitations include:** reliance on Biblical text corpus due to resource constraints, focus on short sentences, and limited scale of human evaluation. **Future work will explore:** incorporating cultural and tonal features specific to Ghomala', transfer learning from better-resourced languages, expanding to longer sentences, applying the approach to other low-resource African languages to assess generalizability and comprehensive human evaluation for semantic coherence assessment

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
2. Guo, J., Lu, S., Cai, H., Zhang, W., Yu, Y., Wang, J.: Long text generation via adversarial training with leaked information. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
3. Kusner, M.J., Hernández-Lobato, J.M.: Gans for sequences of discrete elements with the gumbel-softmax distribution. arXiv preprint arXiv:1611.04051 (2016)
4. Lin, K., Li, D., He, X., Zhang, Z., Sun, M.T.: Adversarial ranking for language generation. Advances in neural information processing systems **30** (2017)
5. de Rosa, G.H., Papa, J.P.: A survey on text generation using generative adversarial networks. Pattern Recognition **119**, 108098 (2021)
6. Sahal, N., Krishnamoorthy, R., Singh, N.: Generating synthetic text using generative adversarial networks. In: 2024 International Conference on Optimization Computing and Wireless Communication (ICOCWC). pp. 1–7. IEEE (2024)
7. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018)
8. Zhang, Y., Gan, Z., Fan, K., Chen, Z., Henao, R., Shen, D., Carin, L.: Adversarial feature matching for text generation. In: International conference on machine learning. pp. 4006–4015. PMLR (2017)

The hybrid CNN+LSTM+SVM based architecture for multilingual speech emotion recognition in low-resource African language using radio data.

Go Issa Traoré[1,2], Borlli Michel Jonas Some[1,3]

[1] Université Nazi BONI, Laboratoire d'algèbre, de Mathématiques discrètes et d'Informatique, Bobo-Dioulasso, Burkina Faso
[2] goissatraore@yahoo.fr
[3] sborlli@gmail.com

Abstract. In low-resource African languages, speech emotion recognition (SER) is an important yet unexplored area of artificial intelligence. Convolutional neural networks (CNN), long short-term memory networks (LSTM), and support vector machines (SVM) are combined in this study to create a novel hybrid machine learning architecture to address the significant challenges of multilingual speech emotion recognition and acoustically diverse communication contexts. This suggested CNN+LSTM+SVM architecture performs competitive results when managing the complex acoustic variances found in African languages. We used the Mel-frequency Cepstral Coefficients (MFCC) technique to extract speech features. The model was trained and validated using radio broadcast recordings from three African tonal languages(Mooré, Dioula and Fulfuldé), encompassing a total of 8522 audio samples across four emotional categories: sadness, satisfaction, anger and neutral. The experimental results reveal an accuracy of 61.11% for emotion recognition when combining data from these three languages, outperforming traditional machine learning approaches by 0.85% on the same dataset.

Keywords: Speech Emotion Recognition · Multilingual Recognition · Deep Learning · African Languages · Hybrid Machine Learning.

1 Introduction

A key area of artificial intelligence and human-computer interaction is speech emotion recognition, which allows for more complex and sympathetic communication between people and technology systems. Even though speech emotion recognition (SER) for commonly spoken languages has advanced significantly, little is known about African low-resource languages, which constitute a complex and linguistically diverse communication ecology. Due to a variety of dialectal variances and recording circumstances, these languages frequently face high acoustic variability, scarce annotated datasets, and restricted digital resources.

This paper presents a hybrid architecture that combines CNN, LSTM, and SVM to recognize multilingual speech emotion recognition in low-resource African language. Through the use of radio broadcast data, a rich and easily accessible source of spoken language, we hope to create a framework for emotion recognition that is flexible and transferable to many linguistic and acoustic contexts. Our suggested methodology aims to achieve three main goals. Fisrtly, Create a strong model for multilingual speech emotion recognition that can function with less training data. Secondely, Show how well a CNN+LSTM+SVM hybrid architecture captures nuanced emotional representations. At last, make a contribution to the developing field of speech technology for linguistically marginalized groups. The rest of the paper is organised as follows: in section 2 we describe the CNN+LSTM+SVM hybrid architecture. In section 3 we present the context of the studied languages. The data collection, processing and annotation step is explain in section 4. The section 5 present the results of our experimentation.

2 Description of the CNN+LSTM+SVM Architecture

The CNN+LSTM+SVM architecture is an innovative machine learning approach that combines three different models to enhance data analysis and classification. Each component plays a specific role in processing information:

Convolutional Neural Networks (CNNs): Extract local characteristics and complex patterns from input data. It use convolutional and pooling layers to create hierarchical representations and reduce data dimensionality while preserving key features.

Long Short-Term Memory (LSTM): Once spatial features[4] are extracted by the CNN, they are passed to the LSTM layers. These layers capture temporal and sequential dependencies in the data. LSTM cells have gates (input, forget, output) that regulate the flow of information, allowing the model to memorize relevant informations across multiple temporal steps while forgetting irrelevant informations. At the end of the learning process, the outputs of the last LSTM layer (called hidden states) are used as features that will be sent to the SVM.

Support Vector Machines (SVMs): Perform final classification of the extracted features. It excel at handling non-linear problems using kernel techniques. It project data into higher-dimensional spaces to enable linear separation. The architecture works sequentially: CNNs first extract spatial features, LSTMs then process these features to understand temporal and sequential relationships, and finally, SVMs classify the resulting feature vectors into predefined categories. This multi-stage approach allows for more sophisticated and accurate data analysis and classification. The functional architecture of this approach is represented through the figure 1.

[4] https://www.sciencedirect.com/science/article/abs/pii/S1047320320300675

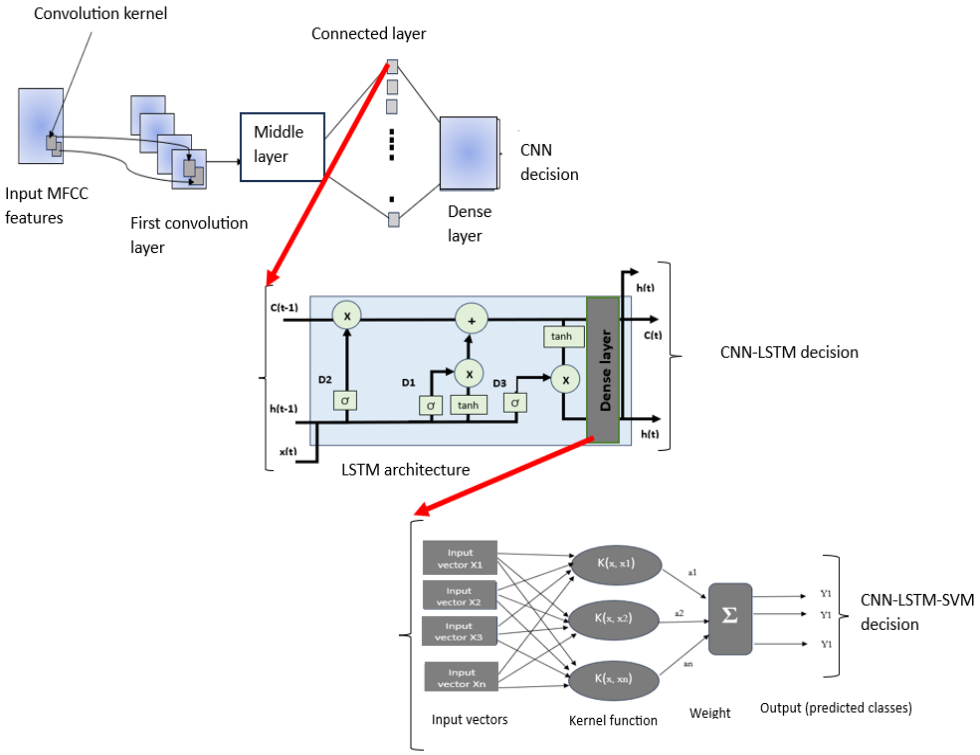

Fig. 1: Functional architecture of CNN+LSTM+SVM

3 Language background

Mooré is a low-resourced language spoken mainly in some West Africa contry as a mother tongue. According to linguistic statistics from the fifth general population and housing census in Burkina Faso [1], Mooré is spoken as the first language by more than 52.9% of the Burkinabe population and is the main language in the capital, Ouagadougou. Several linguistic aspects characterize Mooré. It is a tonal language characterize by high tone, low tone, and middle tone. Not respecting the tones leads to confusion, misunderstandings, or nonsense. A tonal variation in Mooré creates two different words with distinct meanings. For example, the pronunciation of the word "ka" (high tone, means nail) and "ka" (low tone, means here). "Sida" (low tone, means truth) and "Sida" (high tone, means husband).

Dioula is a Mandingo language primarily used in West Africa as a vehicular language. In this region, Dioula occupies a significant place, both for its commercial importance and its social and cultural role. Dioula is a tonal language[5]. We distinguishes the high tone and the low tone. The high tone is marked by

[5] https://journals.openedition.org/corela/4586

the acute accent and the low tone by the grave accent, and to avoid certain ambiguities, tones are mandatory in some cases. For example, 'a' (low tone, means he/she), while 'á' (high tone, means you)

Fulfuldé (also called Peul) is an African language spoken across several countries in West and Central Africa[6]. Peul is a tonal language and generally has three main tones: the high tone (elevated), the middle tone (or normal tone, often considered neutral), and the low tone (grave or low). These tones influence the meaning of words. The tonal system can differ depending on the various Peul dialects. Some dialectal variations can include additional tonal nuances, such as ascending or descending tones, consequently affecting the elocution and meaning of terms [2]. Beyond its lexical role, tone in Peul has a grammatical importance, particularly concerning verb conjugation, plural construction, and other morphological elements. Therefore, a tone variation can signal a specific verbal element or grammatical link [3].

4 Data Collection and Preprocessing

We collecte data from Rural Radio which is the main public information channel in Burkina Faso dedicated to local language broadcasts. These kind of data source are considered the most efficient for realistic emotions recognition instead of simulated or experimental datasets. For each language, we first grouped audio files by program, and files with fewer musical transitions were kept. Unwanted parts with noise were removed, and the remaining audio was divided into 5-second segments using the Python library "FFMPEG" [4]. These segments were then annotated using the PLAVIDA tool [5]. we added a label called "interaction" to identify sound segments containing the voice of more than one speaker. After annotation, files labeled as "interaction" were excluded from the dataset. The annotated data came from 41 interactive radio programs, covering various topics and involving participants from Burkina Faso and beyond. Annotation was carried out by students from Nazi BONI University (UNB), who are familiar with the languages they worked on, specifically Mooré, Dioula, and Fulfuldé. The annotators were mainly modern literature students, trained in local languages spoken in Burkina Faso, including the three target languages. Table 1 presents the annotation statistics for each language at the end of the annotation process.

Table 1: Audio annotation state

Language	**Number of annotated files**	**Number of usable files**	**Rate**
Mooré	10830	6017	55.56%
Dioula	4668	1902	40.74%
Fulfuldé	3241	603	18.60 %
Total	18739	8522	45.47 %

[6] https://www.inalco.fr/langues/peul

5 Results

The emotions considered are neutral, satisfaction, anger, sadness. The results of the emotional corpus fusion are presented in Table 2. This table shows the performance of 3 other classifiers compared to the proposed architecture. Features were extracted using the Mel-frequency Cepstral Coefficients (MFCC) technique.

Table 2: Algorithm performance for each type of data fusion

Dataset	**Modèle**	**Accuracy %**	**Précision %**	**Rappel %**	**F1-score%**
Dioula et Mooré	CNN	56.21	52	52	52
	LSTM	62.42	60	59	60
	SVM	62.82	62	60	60
	CNN+LSTM+SVM	63.60	60	58	58
Fulfuldé et Mooré	CNN	55.77	52	53	51
	LSTM	61.76	61	58	59
	SVM	59.26	70	52	50
	CNN+LSTM+SVM	60.35	59	56	57
Fulfuldé et Dioula	CNN	64.02	56	55	55
	LSTM	70.79	67	67	67
	SVM	66.92	67	54	54
	CNN+LSTM+SVM	71.36	67	68	67
Fulfuldé Mooré, Dioula	CNN	54.98	51	50	51
	LSTM	60.26	58	57	57
	SVM	59.09	60	55	54
	CNN+LSTM+SVM	61.11	58	58	58

Seeking to understand these results by analyzing emotions, we note that using the CNN+LSTM+SVM classifier on the fusion of emotional corpora of these three languages, the emotions anger and neutral are the best recognized with precisions of 69% and 63% respectively (Table 3), which would suggest that these two emotions have similar characteristics for these three languages. The satisfaction emotion gave a precision of 58%. Overall performance is negatively impacted by the recognition of the sadness emotion, with a precision of 41%. The sadness emotion could be considered the one whose oral manifestation differs most in these three languages. By analyzing the recognition rates by emotions for each type of corpus fusion, we realize that when fusing the Dioula and Fulfuldé corpora, the precisions are around 75%, 70%, 45%, and 45% respectively for neutral, satisfaction, anger, and sadness emotions (Table 3). The fusion of Mooré and Dioula corpora gives precisions of 70%, 58%, 56%, 54% respectively for anger, satisfaction, sadness, and neutral with the CNN+LSTM+SVM (Table 3). When fusing the Mooré and Fulfuldé corpora, the precisions are 67%, 62%, 60%, and 55% for anger, neutral, sadness, and satisfaction emotions respectively. This could explain the proximity of oral emotion expression in Dioula and Fulfuldé compared to the two other language combinations.

Table 3: Performance of the CNN+LSTM+SVM classifier on the fusion of the Mooré, Dioula and Fulfuldé corpora

Emotion	**Fusion**							
	Mooré, Dioula, Fulfuldé		**Dioula, Fulfuldé**		**Mooré, Dioula**		**Mooré, Fulfuldé**	
	Precision	F1-score	Precision	F1-score	Precision	F1-score	Precision	F1-score
Anger	69%	68%	45%	51%	70%	72%	67%	70%
Neutral	63%	64%	75%	75%	54%	57%	62%	69%
Sadness	41%	42%	45%	36%	56%	43%	60%	47%
Satisfaction	58%	57%	70%	68%	58%	59%	55%	51%

6 Conclusion

This research presents the potential of hybrid machine learning architectures in SER technologies for low-resource African languages, demonstrating that these kind of architectures can be use in solving challenges in multilingual emotional communication analysis.

One of the advantages of this architecture is its ability to efficiently process complex data. Moreover, the combination of these models can outperform the individual use of these models. A major drawback of such an architecture lies in computational complexity as the combination of three different architectures can make training more costly in terms of resources and time, particularly in computational and memory requirements. Another significant disadvantage is the difficulty in hyperparameter tuning: the number of CNN layers, LSTM parameters, and SVM regularization can be delicate. Future research will focus on expanding the dataset to include more linguistic regions and emotional variations, exploring additional deep learning models and feature extraction methods, testing the hybrid model across diverse environments and acoustic conditions, and incorporating more emotions into the dataset.

References

1. Résultats du 5E Recensement Général de la Population et de l'Habitation | INSD (2019), https://www.insd.bf/fr/file-download/download/public/2071
2. Ard, J.: A comparative and historical study of locative-based periphrastic verbal forms in fula dialects. Studies in African linguistics **10** (2010), https://api.semanticscholar.org/CorpusID:125208153
3. Arnott, D.: The nominal and verbal systems of fula (1970), https://api.semanticscholar.org/CorpusID:141610784
4. Dosani, P., Parekh, T., Harishankar, C.V., Mulla, N.: Media file descriptor using deep learning. In: 2021 Asian Conference on Innovation in Technology (ASIANCON). pp. 1–7 (2021). https://doi.org/10.1109/ASIANCON51346.2021.9544642, https://doi.org/10.1109/ASIANCON51346.2021.9544642
5. Traoré, G.I., Some, B.M.J., Ouédraogo, O., Kalmogo, L.: Plavida, an annotation tool for audio and video in african languages. In: Towards new e-Infrastructure and e-Services for Developing Countries. pp. 141–153. Springer Nature Switzerland, Cham (2025), https://link.springer.com/chapter/10.1007/978-3-031-81573-7$_1$1

A Triphone Hidden Markov Model for Forced Alignment of Nda' Nda' Speech

Tchaheu Tchaheu Dimitri[1], Kana Azeuko Sherelle[1], and Melatagia Yonta Paulin[1,2]

[1] Department of Computer Science, University of Yaounde I, Yaounde, Cameroon
dimitri.tchaheu@facsciences-uy1.cm,
{kanaazeukosherelle,paulinyonta}@gmail.com
[2] IRD, UMMISCO, F-93143, Bondy, France

Abstract. Forced alignment is a technique for automatically synchronizing text and an audio recording. In this work, the aim was to propose a model to improve automatic forced alignment of speech in poorly endowed languages, in particular for the Nda' Nda' language, spoken in the West Cameroon region, while taking into account the tonal aspect. To achieve this goal, a triphone Hidden Markov Model (HMM) model was developed, trained with Mel-Frequency Cepstral Coefficients (MFCC) and pitch features, to which delta and delta-delta derivatives were added. A phonetic decision tree was used when training the triphone model, particularly during state fusion, with two groups of questions concerning phonemes with the same tones, tones with the same base vowels, and sound categories that describe the articulatory and acoustic characteristics of the phonemes; nasals and fricatives were taken into account. For the experiments, four models were trained: HMM monophone, HMM triphone, HMM triphone + Speaker Adaptive Training (SAT), and a hybrid HMM - Deep Neural Network (DNN) model. The best model was the HMM triphone, with a Word Error Rate (WER) of 8.92% and a median Phone Boundary Error (PBE) of 75.5 millisecond (ms).

Keywords: HMM triphone · Forced alignment · Poorly endowed languages · WER · PBE

1 Introduction

A poorly resourced is one with few computerized linguistic resources (lexicons, spell checkers, parsers, etc.). Some of these languages present an additional challenge, notably tonal languages, where the pitch or tone used can change the meaning of a word. Nda' Nda' falls into this category, and is also considered an endangered language [1]. It is a language spoken in the western region of Cameroon. In 1990, the number of speakers was estimated at 10,000 [10]. It has four tones: "high" (˝), "low" (ˋ), "low high" (ˇ) and "high low" (ˆ).

Forced alignment plays a crucial role in Automatic Speech Recognition (ASR), as it contributes to the creation of aligned speech corpora, the creation of synchronized subtitles, and so on. There are several approaches to forced speech

alignment. Three main families of approaches can be distinguished: approaches based on machine learning, notably Hidden Markov Models (HMM) [4, 8, 7]; another family is deep learning [5]. The last approach is hybrid HMM-DNN [11, 9] or HMM- Support Vector Machine (SVM) [6]. All these approaches work well for highly endowed languages, such as French or English. The aim of this work is to propose an alignment model that takes into account tone and low data volume. To achieve this goal, a triphone HMM model was trained using MFCCs and pitch features [3], with delta derivatives [2] added, and a phonetic decision tree [12] was used for state fusion, incorporating tonal and articulatory information.

The rest of the article is structured as follows: Section 2 presents the different existing approaches, Section 3 describes the proposed approach, Section 4 discusses the experiments, Section 5 outlines the results obtained, and finally, Section 6 concludes this work.

2 Related work

[9] Present an approach that combines Hidden Markov Models (HMM) with Deep Neural Networks (DNN). The DNN replaces the GMM in order to avoid the assumption that acoustic features follow normal laws. The experiments were carried out on a database of connected digit recordings. The HMM-DNN model obtained a WER of 2.5%, while the HMM recorded a WER of 3.8%. [6] propose an improved method for automatic phoneme segmentation, using a two-stage architecture integrating HMM and SVM. The first stage performs a forced HMM alignment to associate a phoneme sequence with its acoustic signal, based on a minimum boundary error criterion. The second stage refines the initial boundaries using an SVM. Evaluation of the HMM-SVM models on the TIMIT database showed an average word boundary error of 6.75 ms, compared with 7.14 ms for the simple HMM model. To facilitate the training of HMM-based forced alignment models, [4] have developed an automatic speech alignment system, called Prosodylab-Aligner (PLA), using the HTK toolbox. In addition to this toolbox, which exploits transcribed audio files to develop acoustic models, other architectures, such as those based on RNNs, have been proposed. [5] proposes an RNN architecture consisting of five layers of hidden units, the first three of which are non-recurrent, with a CTC (Connectionist Temporal Classification) loss function. This model was trained on 300 hours of English corpus, including Switchboard and Fisher. The RNN achieved a WER of 16 %, while the HMM-DNN achieved 18.4 %. Although the RNN architecture has proven its effectiveness for English, other languages, such as Vietnamese, which is a tonal language, require specific approaches to capture their linguistic peculiarities. With this in mind, [7] propose four models, the two best of which are: a triphone HMM trained with the SAT method [2] and a hybrid model combining HMM with DNN. All these models are trained with MFCCs, pitch, and tonal aspects through questions relating to phones with the same tone and tones with the same base vowel. Experimental results on a database of 16 hours of Vietnamese recordings show a WER of 12.13 % for the triphone HMM model and

9.48 % for the hybrid model. Complementing these acoustic models, [8] introduce Montreal Forced Aligner (MFA), an open-source system for speech-text alignment based on the Kaldi toolbox, which updates PLA. In addition to the features offered by PLA, MFA includes training methods for speaker adaptation [2].

3 Methodology

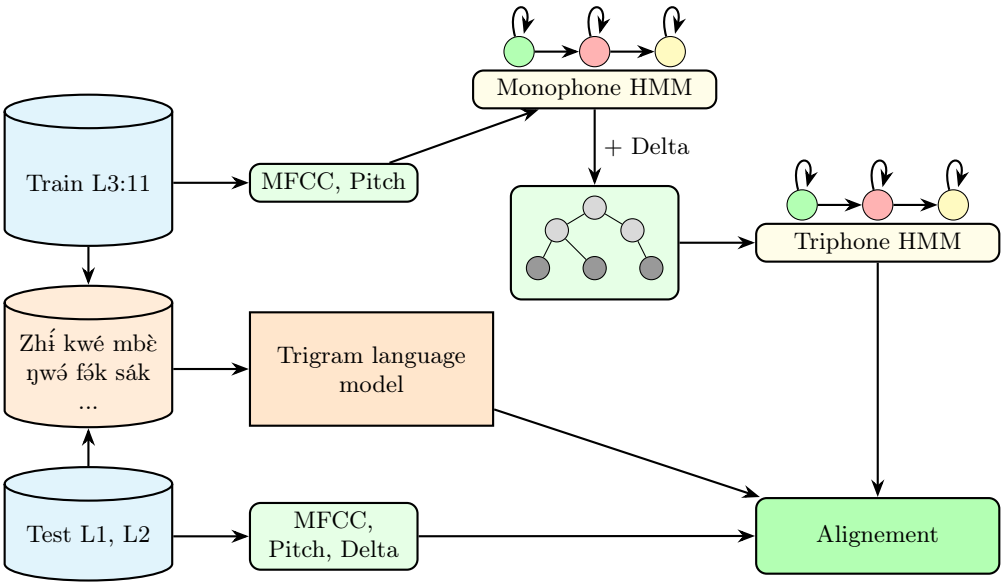

Fig. 1: Model architecture

The proposed model (Fig.1) is inspired by [7], uses two HMMs placed consecutively. A monophone HMM is first trained using MFCC features and pitch [3]. Then, to incorporate the contextual aspect, a triphone HMM is trained using the previous monophone HMM. In addition to the MFCC and pitch used previously, delta-derived coefficients [2] were added for training the triphone model. To overcome the problem of lack of training data [12], a state fusion mechanism using a phonetic decision tree was implemented [12]. Four groups of questions have been proposed to build this tree. The first two groups of questions concern phonemes with the same tone (Tab. 1) and those with the same base vowel (Tab. 2) [7]. It should be noted that of the two existing tone models [7], the data-driven one was used as it simplifies phonological complexity while capturing the essential variations of Nda' Nda'. The second group concerns sound categories, which describe the articulatory and acoustic characteristics of phonemes (Tab. 3). The

sound categories used are nasals and fricatives. Although alignment is performed at phoneme level, it is imperative to control the phoneme sequence being output. Thus a trigram language model was trained with manual transcriptions of training and test utterances.

Table 1: Same tone

◌́		◌̀		◌̌	◌̂
ə́ ɔ́ á é ú ó ɛ́ ɨ́ ɔ́	ɔ̀ ɛ̀ ə̀ ò à	ǎ ǔ	î ê ə̂ ɔ̂		

Table 2: Same base vowel

ɔ	**a**	**ə**	**i**	**o**	**ɛ**	**u**
ɔ́ ɔ̀ ɔ̌	á à	ə̂ ə́ ə̀	í î	ó ò	ɛ́ ɛ̀ ɛ̂	ǔ ú

Table 3: Sound categories

Nasals	**Fricatives**
m n ŋ	f s z

4 Experimental protocol

The experiments were conducted on a corpus consisting of 48 sentences and 11 speakers. Each speaker articulated these sentences, defined by a linguist, according to the subject-verb-complement structure. The recordings were made using a Zoom H6 voice recorder equipped with noise filtering functionality, and took into account various categories of age, gender, and the origins of the speakers (Yaoundé and Bangoua). After pre-processing with Audacity, the total duration of the recordings was reduced to approximately 20 minutes, down from nearly an hour initially. The data from the first nine speakers were manually aligned using Praat software. The data were divided into two parts: the recordings of the first two speakers were used for testing, while the rest were used for training. The dimensions of the feature vector include 13 MFCCs; when combining, 10 MFCCs and 3 pitch coefficients are used. This dimension of 13 was determined empirically, as poor performance has been observed above this threshold. With the addition of delta-derived coefficients, the vector dimension increases to 39. Features were normalized using CMVN [2] to improve the robustness of the recognition models. Four models were trained: a monophone HMM model, a triphone HMM (Fig. 1), an HMM+SAT and a hybrid HMM-DNN based on the architecture [9] but with different parameters. All HMM models contain around 200 Gaussians, and the number of leaves in the phonetic decision tree was set at 10, determined empirically after several tests. For the hybrid configuration, the DNN has a single hidden layer with 300 nodes. The number of parameters is limited due to the small amount of training data. To carry out these experiments[3], 8 CPUs (1.60 GHz each), 8 GB of memory and the Kaldi and Python environments were required.

5 Results and discussion

The tables (Tab. 4 and Tab 5) show that the proposed model is the best with a WER of 8.92 % and a median PBE of 75.9 ms, using MFCC and pitch.

[3] https://github.com/mende237/Nda-Nda-Force-Aligner.git

Adding questions does not improve performance, probably due to the imbalance of triphones (Fig. 2) and tones (Fig. 3) in the corpus. Monophone HMM and HMM+SAT follow with 11.89 % WER; 81.6 ms median PBE and 12.70 % WER; 77.5 ms median PBE, respectively. HMM+SAT performs less well on WER, which shows that speaker-adaptive speech recognition requires a larger number of speakers to be effective, and that the addition of pitch had a positive impact on its performance for PBE. The hybrid HMM-DNN model performs least well, suffering from a lack of data, although it improves its results by combining MFCC, pitch and questions. A common trend is that voice pitch improves performance.

Table 4: WER Results

	MFCC	+Pitch	+Questions
HMM-Mono	12.70	11.89	-
HMM-Tri	9.73	**8.92**	**8.92**
HMM+SAT	12.70	12.97	12.70
HMM-DNN	21.35	19.73	18.65

Table 5: PBE Results in ms (Mean; median)

	MFCC	+Pith	+Questions
HMM-Mono	(122.1; 83.3)	(123.7; 81.6)	-
HMM-Tri	(**115.0**; 76,9)	(117.1; **75.9**)	(117.1; **75.9**)
HMM+SAT	(127.0; 77.5)	(118.0; 81.4)	(132.7; 82.3)
HMM-DNN	(134.5; 88.0)	(124.6; 84.1)	(124.0; 84.6)

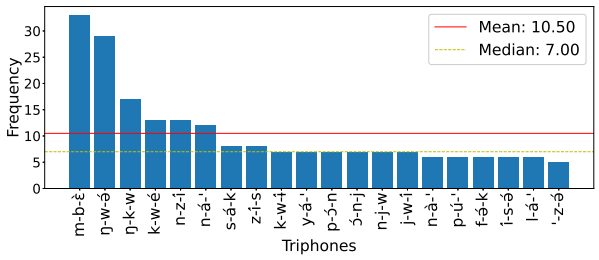

Fig. 2: The 20 most frequent triphones on 117

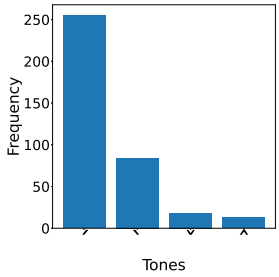

Fig. 3: Tone distribution

6 Conclusion

The aim of this work was to develop an acoustic model for speech alignment in Nda' Nda', a poorly endowed language spoken in the West Cameroon region. An HMM triphone model was proposed, incorporating MFCC and pitch features, as well as delta derivatives to better capture tonal variations. Experiments compared several models, revealing that the triphone HMM performed best, with a WER of 8.92 % and 75.9 ms of median PBE. The results underline the importance of vocal characteristics, such as pitch, in better discriminating the tonal variations essential to Nda' Nda'. However, the impact of integrating linguistic questions when building the decision tree was not felt, due to the imbalance in the distribution of triphones and tones in the corpus. Nevertheless, the limited

volume of data remains a major constraint, suggesting that transfer learning and increased data collection could be avenues for the future.

References

1. Zachée Denis Bitjaa Kody. *La Dynamique des langues camerounaises en contact avec le français: approche macrosociolinguistique*. PhD thesis, Univ. Yaounde, 2004.
2. Mark Gales, Steve Young, et al. The application of hidden markov models in speech recognition. *Foundations and Trends® in Signal Processing*, 1(3):195–304, 2008.
3. Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur. A pitch extraction algorithm tuned for automatic speech recognition. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2494–2498, 2014.
4. Kyle Gorman, Jonathan Howell, and Michael Wagner. Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, 39(3):192–193, 2011.
5. Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.
6. Jen-Wei Kuo, Hung-Yi Lo, and Hsin-Min Wang. Improved hmm/svm methods for automatic phoneme segmentation. In *Interspeech*, pages 2057–2060, 2007.
7. Hieu-Thi Luong and Hai-Quan Vu. A non-expert kaldi recipe for vietnamese speech recognition system. In *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*, pages 51–55, 2016.
8. Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502, 2017.
9. Nelson Morgan and Herve Bourlard. Continuous speech recognition. *IEEE signal processing magazine*, 12(3):24–42, 1995.
10. Emile Gille Nguendjio. *A Descriptive Grammar of Bangwà - a Grassfields Language of Cameroon*, volume 47 of *Grammatical Analyses of African Languages*. Köppe, Cologne, 2014.
11. Xian Tang. Hybrid hidden markov model and artificial neural network for automatic speech recognition. In *2009 Pacific-Asia Conference on Circuits, Communications and Systems*, pages 682–685. IEEE, 2009.
12. Steve J Young, Julian J Odell, and Phil C Woodland. Tree-based state tying for high accuracy modelling. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994.

Pro-TeVA: Prototype-based Explainable Tone Recognition for Low-Resource Language

S.G.B. Bengono OBIANG[1,2,3], Paulin MELATAGIA YONTA[1,3], Norbert TSOPZE[1,3], Tania JIMENEZ[2], and Jean-Francois BONASTRE[2,4] Farida NCHARE[1,3]

[1] Département d'Informatique - Université de Yaoundé 1, Yaoundé, Cameroun
[2] Laboratoire Informatique d'Avignon, EA 4128, Avignon Université, France
[3] Sorbonne Université - IRD - UMMISCO - F-93143, Bondy, France
[4] Inria, Defense and Security dept., Paris, France

Abstract. Many sub-Saharan African languages are tonal, where pitch variations affect meaning. Manual tone annotation is slow and requires expert knowledge, posing challenges for documenting low-resource languages. We introduce Pro-TeVA, a Prototype-based Temporal Variational Autoencoder that combines tone recognition with interpretability. The framework integrates a variational autoencoder for latent representation, a prototype layer to model tonal patterns, and a CTC-based classifier for sequence prediction. Fundamental frequency (F_0) is used both as a learning signal and a transparent explanation tool. Evaluated on Yoruba speech, Pro-TeVA achieves a Tone Error Rate of 17.74% using just 10 prototypes—comparable to black-box models—while offering interpretable outputs through prototype visualisation and F_0 reconstruction. This approach supports linguists in validating model decisions and accelerates tone annotation for low-resource languages.

Keywords: Speech Processing · Tone Recognition · Low-resource · Explainable AI

1 Introduction

Tonal languages, such as Yoruba, use pitch to distinguish between otherwise identical phonetic sequences. For instance, "kí" (to salute), "ki" (thick), and "kì" (to press) differ only by tone [3]. These distinctions pose specific challenges for automatic speech recognition (ASR), especially in low-resource settings where annotated datasets are limited [8].

Several approaches have achieved promising results for tone recognition using features such as MFCC [6], cepstrograms [11], and more recently Wav2vec 2.0 [1], which has achieved tone error rates (TER) as low as 17.72% [3]. However, most of these systems rely on black-box models, limiting their utility for linguistic validation or manual annotation [2].

Post-hoc explicability methods [14, 10], offer some avenues, but are not very well suited to our context. For tonal speech, the fundamental frequency (F_0) remains a key explanatory feature, well understood by linguists [5]. Integrating F_0

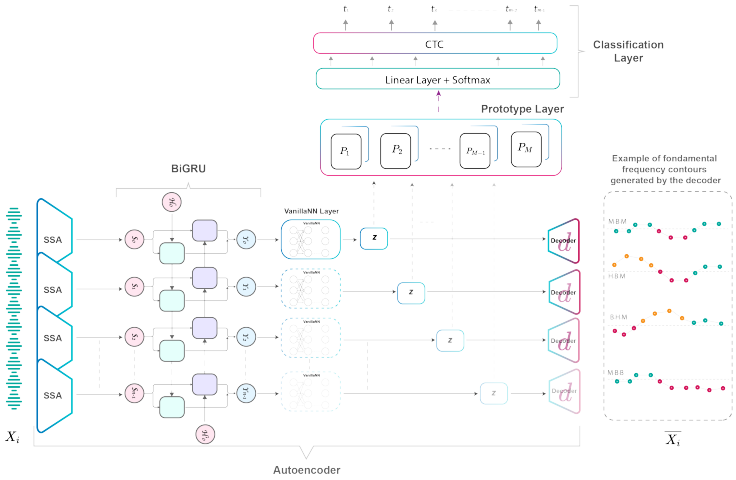

Fig. 1. Pro-TeVA: VAE for latent encoding and F0 reconstruction, prototype layer for interpretability, and CTC for sequence prediction.

into a prototype-based system provides a natural bridge between computational decisions and human interpretability.

In this paper, we introduce Pro-TeVA (Prototype-based Temporal Variational Autoencoder), a framework for explainable tone recognition. Pro-TeVA combines latent-space prototype learning with F_0 reconstruction and alignment-free tone prediction via connectionist temporal classification (CTC) [7]. Applied to Yoruba, Pro-TeVA achieves a TER of 17.74%, matching the best non-explainable models, while providing fine-grained, interpretable outputs that are useful for both annotation and linguistic analysis.

2 Pro-TeVA Framework

Pro-TeVA (Prototype-based Temporal Variational Autoencoder) enhances tone recognition by uniting performance and interpretability. As shown in Figure 1, it combines: (i) a VAE for latent tonal representations. The latent space uses a vanilla neural network (VanillaNN), which is a basic feedforward neural network with fully connected layers and no advanced features.(ii) a prototype layer to anchor tonal categories (M is Middle Tone. B is Low Tone. H is High Tone), and (iii) a CTC classifier for tone sequence prediction.

The encoder maps SSA-Hubert [4] speech features to a probabilistic latent space. Prototypes represent recurring tonal patterns and are compared to encoded frames via negative distances. These similarities feed a CTC classifier to predict tone sequences, enabling alignment-free training.

2.1 Latent Encoding and F0 Reconstruction

The VAE compresses speech into a 512-dim latent space. Its decoder reconstructs both SSA features and F_0 contours (extracted via YIN [5]) to enforce linguistic structure. Reconstructed F_0 provides explanations for tonal predictions and acts as an auxiliary loss.

2.2 Prototype Layer

The prototype layer models contextual tone variants by anchoring latent representations to interpretable reference points. During training, a joint loss encourages the prototypes to be representative, diverse, and well-distributed [13], thereby enhancing interpretability without sacrificing performance.

2.3 CTC Classifier and Loss

Similarity vectors are passed to a fully connected layer and softmax to output tone predictions. The CTC loss [7] trains the model without frame-level alignments.

2.4 Overall Loss

The total training objective combines tone prediction, latent regularization, and prototype constraints:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{CTC}}\mathcal{L}_{\text{CTC}} + \alpha\mathcal{L}_{\text{VAE}} + \beta\mathcal{L}_{\text{proto}}$$

This design offers a transparent, linguistically grounded system for tone analysis in low-resource settings.

3 Experiments

We evaluate Pro-TeVA for tone recognition in Yoruba through a series of quantitative, qualitative, and error analyses.

3.1 Setup and Training

We use the Yoruba speech dataset by Gutkin et al. [8], adapted for tone recognition via syllabification [12] and manual tone verification. The dataset (3h20m) was downsampled to 16 kHz and filtered to utterances $<$6s (3,223 examples). It was split into training (60%), validation (10%), and test (30%) sets. SSA-Hubert features and F_0 values were extracted. The tone inventory includes three tones (H, M, L) and a separator token for CTC.

Training follows a two-stage strategy: (1) prototype initialization prioritising classification; (2) full model fine-tuning with CTC, reconstruction, and alignment losses. We apply SpecAugment, noise injection (RIRS [9]), and speed perturbation. SSA-Hubert features are derived from multilingual pre-trained models.

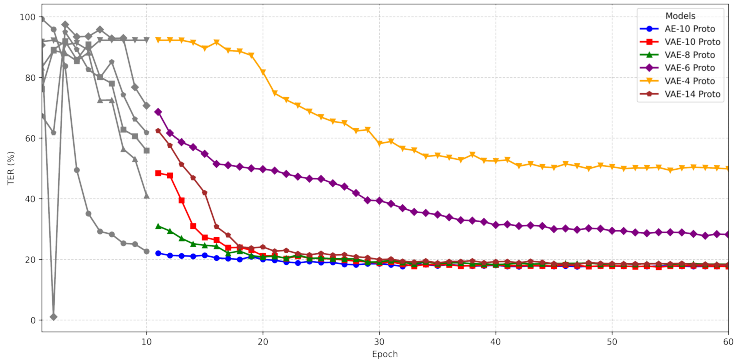

Fig. 2. TER by number of prototypes over training epochs.

3.2 Performance and Explainability

Tone Error Rate (TER) and **Error Analysis** are shown in Tables 1 and 2. Pro-TeVA achieves 17.74% TER, closely matching the best-performing W2V+GRU (17.72%) while offering explainability. Error-wise, Pro-TeVA favours substitution errors, which are easier to post-edit.

Table 1. Comparison of tone recognition performance on Yoruba

Model	TER (%)	Explainable
FB [3]	21.43	No
CS [3]	20.36	No
MFCC	25.66	No
LEAF	21.91	No
FB + CS	19.45	No
W2V (frozen)	22.33	No
W2V (fine-tuned)	18.87	No
W2V + GRU	**17.72**	No
Pro-TeVA (10 prot.)	17.74	**Yes**

Table 2. Error types across models

Model	Insertions	Deletions	Substitutions
FB	367	647	890
CS	306	753	711
FB + CS	286	719	689
W2V	225	477	841
Pro-TeVA	241	503	823

3.3 Interpretability and Prototype Efficiency

Figure 3 shows that Pro-TeVA's predicted tone sequence aligns closely with reconstructed F_0 contours, enabling interpretable error analysis and tone transition explanation. Figure 2 highlights the impact of the number of prototypes on performance: 10 prototypes yield optimal TER; using too few (e.g., 4) degrades performance severely (50.58%).

4 Conclusion

We introduced ProTeVA, a prototype-based temporal variational autoencoder designed for explainable tone recognition in Yoruba, a low resource tonal lan-

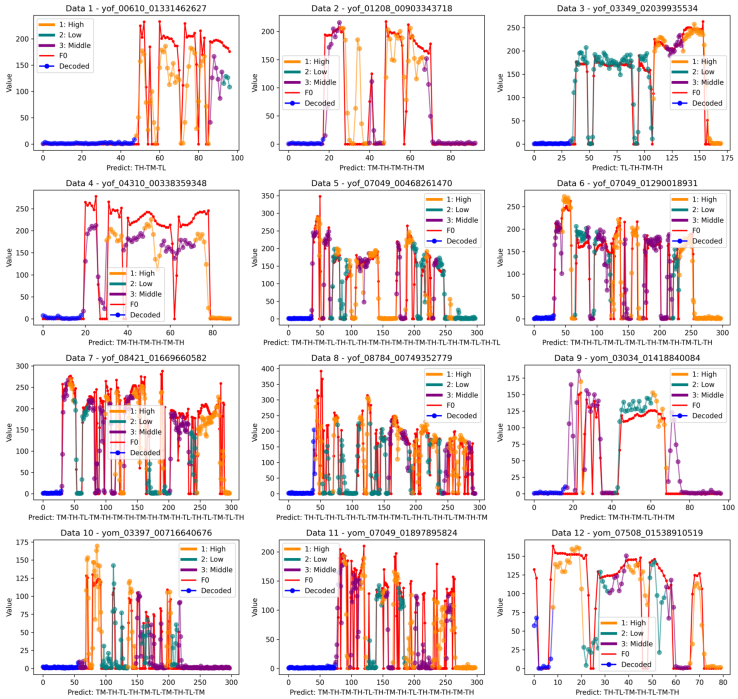

Fig. 3. F_0 reconstruction and tonal predictions on short/long examples; silence marked by final tone.

guage. By combining variational encoding, prototype learning, and F_0 reconstruction, ProTeVA achieves a Tone Error Rate of 17.74%, comparable to state-of-the-art models, while offering clear linguistic interpretability. With only 10 prototypes, the model captures contextual tone variations and generates visual explanations aligned with fundamental frequency contours. Its robustness across utterance lengths and compact architecture makes it a valuable tool for both speech recognition and linguistic analysis.

Future work includes extending ProTeVA to other tonal languages through cross-lingual transfer and integrating it into full speech recognition pipelines to support end-to-end explainable automatic speech recognition.

References

1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 12449–12460. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

2. Becker, S., Ackermann, M., Lapuschkin, S., Müller, K., Samek, W.: Interpreting and explaining deep neural networks for classification of audio signals. CoRR **abs/1807.03418** (2018), http://arxiv.org/abs/1807.03418
3. Bengono Obiang, S.G.B., Tsopze, N., Melatagia Yonta, P., Bonastre, J.F., Jiménez, T.: Improving tone recognition performance using wav2vec 2.0-based learned representation in yoruba, a low-resourced language. ACM Trans. Asian Low-Resour. Lang. Inf. Process. **23**(12) (Nov 2024). https://doi.org/10.1145/3690384, https://doi.org/10.1145/3690384
4. Caubrière, A., Gauthier, E.: Africa-centric self-supervised pre-training for multilingual speech representation in a sub-saharan context (2024), https://arxiv.org/abs/2404.02000
5. de Cheveigné, A., Kawahara, H.: Yin, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America **111**(4), 1917–1930 (04 2002). https://doi.org/10.1121/1.1458024, https://doi.org/10.1121/1.1458024
6. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing **28**(4), 357–366 (1980). https://doi.org/10.1109/TASSP.1980.1163420
7. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning. pp. 369–376. ICML '06, Association for Computing Machinery, New York, NY, USA (2006). https://doi.org/10.1145/1143844.1143891, https://doi.org/10.1145/1143844.1143891
8. Gutkin, A., Demirşahin, I., Kjartansson, O., Rivera, C., Túbọ̀sún, K.: Developing an Open-Source Corpus of Yoruba Speech. In: Proceedings of Interspeech 2020. pp. 404–408. International Speech and Communication Association (ISCA), Shanghai, China (October 2020). https://doi.org/10.21437/Interspeech.2020-1096, http://dx.doi.org/10.21437/Interspeech.2020-1096
9. Ko, T., Peddinti, V., Povey, D., Seltzer, M.L., Khudanpur, S.: A study on data augmentation of reverberant speech for robust speech recognition. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5220–5224 (2017). https://doi.org/10.1109/ICASSP.2017.7953152
10. Li, O., Liu, H., Chen, C., Rudin, C.: Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. pp. 3530–3537. AAAI Press (2018), https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17082
11. Lugosch, L., Tomar, V.S.: Tone recognition using lifters and ctc. In: Proc. Interspeech 2018. pp. 2305–2309 (2018). https://doi.org/10.21437/Interspeech.2018-2198
12. van Niekerk, D.R., Barnard, E.: Tone realisation in a yorùbá speech recognition corpus. In: Proc. 3rd Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2012). pp. 54–59 (2012)
13. Obiang, S.B., Tsopze, N., Bonastre, J.F., Yonta, P.M., Jimenez, T.: Variational autoencoder for a prototype-based explainable neural network (2024). https://doi.org/10.2139/ssrn.4861108, https://ssrn.com/abstract=4861108
14. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144 (2016)