LEXICON BASED SENTIMENT CLASSIFICATION FOR SELECTED AFRICAN LANGUAGES

DEEP LEARNING INDABA

BOWEN UNIVERSITY
Excellence & Godliness

Precious Ooreoluwanimi Ayinde & Halleluyah Oluwatobi Aworinde

MOTIVATION

- In the modern era of social media, where online reviews, discussions, and comments are abundant, people and businesses heavily depend on content from these platforms to inform their choices.
- Analyzing and extracting valuable information from extensive amounts of text can be a challenging endeavor, this is where automated sentiment analysis systems come to the forefront.
- However, Nigerian languages are low-resource in the context of the sentiment classification task.
- Using deep learning-based approaches such as pre-trained language models which have been shown to have gotten impressive performance on different NLP is computationally expensive and resources for training those kinds of models are not readily available in Nigeria.
- Furthermore, the sentiment lexicon has been curated for these languages but not used for developing sentiment classifiers for these languages.

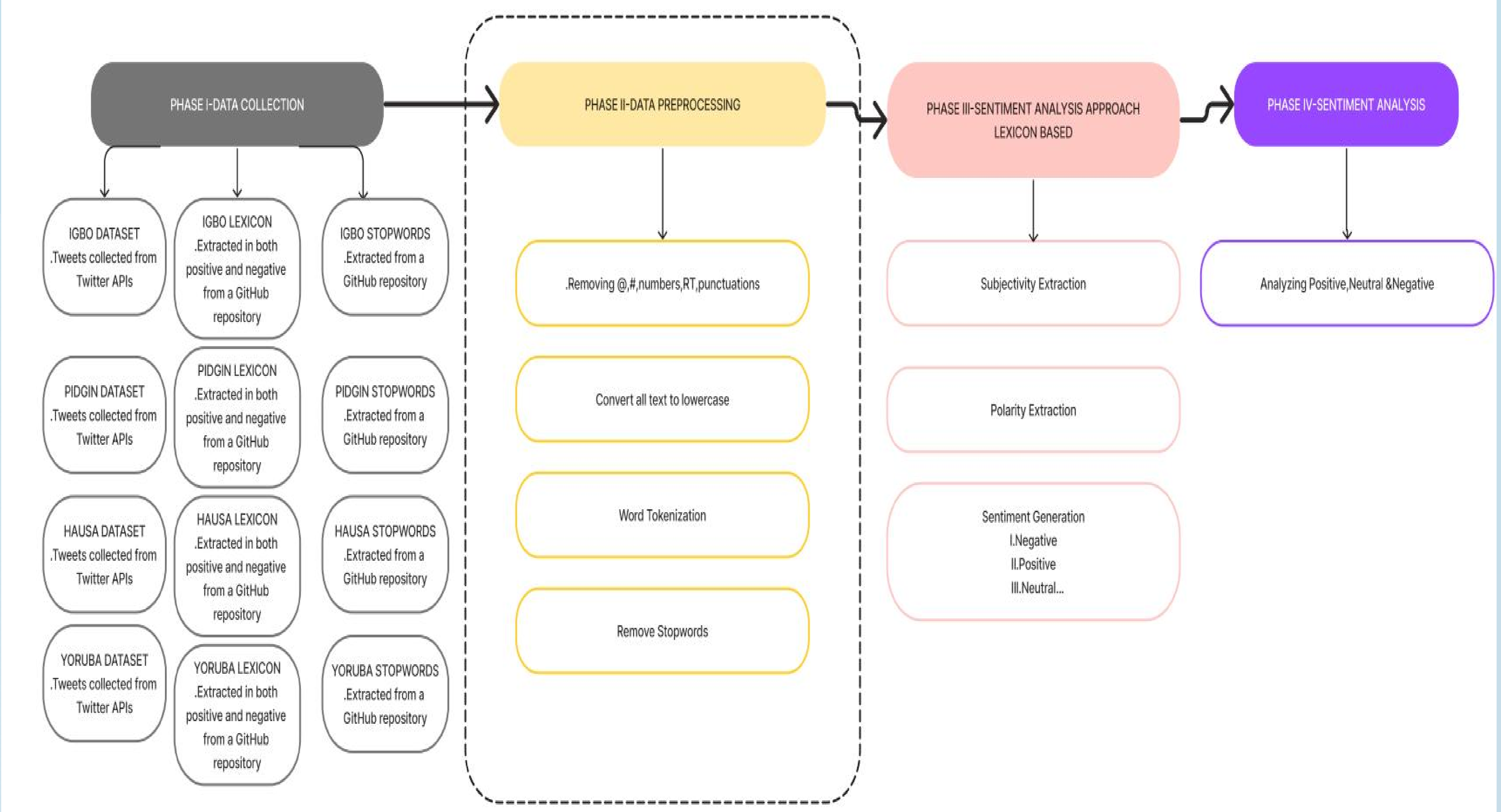
OBJECTIVES

i. Lexicon-based sentiment classification were performed for 4 African languages {Nigerian Pidgin, Yoruba, Igbo, Hausa}.

ii. Investigation on how preprocessing techniques and other lexicon-based approaches influence the performance of sentiment classification models specifically designed for African languages.

iii. Comparison of the result of lexicon-based approach with machine learning approach of sentiment classification for African languages.

FRAMEWORK



RESULTS

HAUSA			
Overall accuracy: 0.5347313237221494			
	POSITIVE	NEGATIVE	NEUTRAL
Precison	0.543723	0.694749	0.469392
Recall	0.574215	0.376250	0.644561
F-1 Support	0.558553	0.488141	0.543204
Support	3281.000000	3200.000000	3438.000000

IGBO			
Overall accuracy: 0.5893733352025795			
	POSITIVE	NEGATIVE	NEUTRAL
Precison	0.601452	0.601695	0.575829
Recall	0.652456	0.468132	0.616165
F-1 Support	0.625917	0.526576	0.595315
Support	2158.000000	1820.000000	3155.000000

Result generated using Tweets in Original Language(Hausa, Igbo)

HAUSA			
Overall accuracy: 0.3466075209194475			
	POSITIVE	NEGATIVE	NEUTRAL
Precision	0.0	0.0	0.346608
Recall	0.0	0.0	1.000000
F-1	0.0	0.0	0.514786
Support	3281.0	3200.0	3438.000000

IGBO			
Overall accuracy: 0.4423103883359035			
	POSITIVE	NEGATIVE	NEUTRAL
Precision	0.0	0.0	0.442310
Recall	0.0	0.0	1.000000
F-1	0.0	0.0	0.613336
Support	2158.0	1820.0	3155.000000

Result generated without preprocessing dataset.

DATA COLLECTION AND VISUALIZATION

	ID	tweet	label	preprocessed
0	ig_train_04335	@user @user @user @user @user @user Ji lu ile ...	neutral	[lu, ile, guo, eze, onu, ocho, lu, “, if, cap,...
1	ig_train_03810	Eze! Ka o di iche https://t.co/bdp0K0eZ8D	neutral	[eze, iche, httpstcobdpkezd]
2	ig_train_02389	Ndi isi mgbaka juru na ebe a, so not surprised...	negative	[mgbaka, juru, so, not, surprised, one, bit, h...
3	ig_train_04932	@user That's the content I am here for. Isuru ...	neutral	[thats, content, am, here, for, isuru, ekwulob...
4	ig_train_04668	Mmiri enweghi ilo.	neutral	[mmiri, enweghi, ilo]

```
cp_train
```

	ID	tweet	label	preprocessed
0	ig_train_04335	@user @user @user @user @user @user Ji lu ile ...	neutral	[lu, ile, guo, eze, onu, ocho, lu, “, if, cap,...
1	ig_train_03810	Eze! Ka o di iche https://t.co/bdp0K0eZ8D	neutral	[eze, iche, httpstcobdpkezd]
2	ig_train_02389	Ndi isi mgbaka juru na ebe a, so not surprised...	negative	[mgbaka, juru, so, not, surprised, one, bit, h...
3	ig_train_04932	@user That's the content I am here for. Isuru ...	neutral	[thats, content, am, here, for, isuru, ekwulob...
4	ig_train_04668	Mmiri enweghi ilo.	neutral	[mmiri, enweghi, ilo]
...	...	...	...	...
7128	ig_train_00466	@user @user @user Nne, ujo a na atu gi? Gozie ...	negative	[ujo, atu, gozie, is, my, managernya, jee, nal...
7129	ig_train_03598	Good Morning Coal City! 🌑 Ndị Enugu kedụ ka os...	neutral	[good, morning, coal, city, 🌑, ndị, enugu, ked...
7130	ig_train_04093	@user @user @user Enweghi m ego. 😢😔😫	neutral	[enweghi, 😢😔😫]
7131	ig_train_01633	Umu nwoke na tuwita, ịịrikwa udo ke do ada obi...	negative	[umu, tuwita, ịịrikwa, udo, ke, do, ada, gi, n...

```
array(['gò', 'ese', 'aisodødo', 'abirun', 'wahala'], dtype=object)
```

```
array(['a', 'an', 'bá', 'bí', 'bẹ̀rẹ̀'], dtype=object)
```

Source:https://github.com/afrisenti-semeval/afrisent-semeval-2023/tree/main/sentiment_lexicon

Language code	Name	Language family	Number of speakers (approx.)	Number of Tweets
ha	Hausa	Chadic	120000000	14172
ig	Igbo	Niger-congo	25000000	10192
pcm	Naija pidgin	English creole	78000000	5123
yo	Yoruba	Niger-congo	30000000	8522

Tweets used for the work was obtained from Afrisenti Datatset

ENGLISH			
Overall accuracy: 0.49413145539906106			
	POSITIVE	NEGATIVE	NEUTRAL
Precision	0.645010	0.421875	0.450247
Recall	0.388620	0.329519	0.713563
F-1	0.485016	0.329519	0.552117
Support	2478.000000	1311.000000	0.552117

The result of the classification carried out on preprocessed translated datasets

ENGLISH			
Overall accuracy: 0.3646881287726358			
	POSITIVE	NEGATIVE	NEUTRAL
Precision	0.0	0.0	0.364688
Recall	0.0	0.0	1.000000
F-1	0.0	0.0	0.534464
Support	2478.0	1311.0	2175.000000

Result of the classification carried out on un-preprocessed translated datasets

CONCLUSION

To improve the precision and efficiency of lexicon-based sentiment categorization, the researcher's study emphasizes the necessity for more comprehensive lexicon resources for African languages. For accurate sentiment analysis in African languages, language-specific lexicons must still be developed and made available, even though the machine learning approach shows promise in the absence of enough lexicons.