Science World Journal Vol. 19(No 1) 2024 https://dx.doi.org/10.4314/swj.v19i1.13
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Full Length Research Article

TWITTER SENTIMENT ANALYSIS FOR HAUSA ABBREVIATIONS AND ACRONYMS

Habeeba Ibraheem Abdullahi, *Muhammad Aminu Ahmad, and Khalid Haruna

Department of Computer Science, Faculty of Computing, Kaduna State University, Kaduna

*Corresponding Author Email Address: muhdaminu@googlemail.com

ABSTRACT
The use of natural language processing, to identify, extract and organize sentiment from user generated texts in social networks, blogs or product review of text is known as sentiment analysis or opinion mining. Hausa language belongs to one of the major well-spoken languages in Africa and one of the three major Nigerian languages. Now investigating into such a language will have significant influence on social, economic business political and even educational services and settings. Some of these Hausa texts are abbreviated and some in acronym format which is a challenge to researchers as such comments are in an unstructured format and needs normalization to get further understanding of that text and also there is scarcity of sentiment analysis on Hausa abbreviation and acronym. Abbreviation is a shorten form of a word while acronym is an abbreviation formed from the initial letters of other words and pronounced as a word. This research aims to develop an improved Hausa Sentiment Dataset for the enhancement of sentiment analysis with abbreviation and acronyms. This is achieved by adapting to the approach for Hausa Sentiment Analysis based on Multinomial Naïve Bayes (MNB) and Logistic Regression algorithms using the count vectorizer, along with python libraries for NLP. This research affirmed that the improved dataset with abbreviation and acronym outperforms the plain Hausa dataset by 4% in accuracy using Multinomial Naïve Bayes. The result shows that in addition to normal preprocessing techniques of the social media stream, understanding, interpreting and resolving ambiguity in the usage of abbreviations and acronyms lead to improved accuracy of algorithms with evidence in the experimental result.

Keywords: Sentiment Analysis, Hausa abbreviation and acronym, Count Vectorizer, Machine Learning.

INTRODUCTION

The use of natural language processing, to identify, extract and organize sentiment from user generated texts in social networks, blogs or product review of text is known as sentiment analysis or opinion mining (Tang et al, 2015). Recently there is a huge rate of code-mix (combination of languages) text available on social media. Some of these texts are abbreviated and some are transliteration of one text into another which needs normalization to get further understanding of that text. Individuals make comments on twitter which involves perception, emotion and unexpected behavior. This feedback could be of value to those involved to create new or revise existing services and solutions. Such data can be easily extracted from various social-media platforms like Twitter, Facebook, web blogs etc. Abbreviation is a shorten form of a word while acronym is an abbreviation formed from the initial letters of other words and pronounced as a word.
Just like there is ambiguity in the use of normal language there is also ambiguity in the usage of slang, abbreviation and acronym because they often have context-based meanings, which must be rightly interpreted in order to improve the results of social media analysis.
Sentiment analysis origin can be traced to 1950's, then, sentiment analysis was mainly used on written paper documents (kdnuggets, 2015). Dewaele (2010) claims that 'strong emotional arousal' increases the frequency of code-mixing. Mantyla et al (2017) reported that the sentimental analysis outburst that are computer-based only transpired due to the presence of web based subjective text. The adoption of social media platforms, such as Twitter has made SA of tweets an important area of research in customer feedback, public opinion polls, advertisement etc. Contrary to popular assumption, African languages have not been given much attention in corpus linguistics despite the existence of electronic resources (Zakari et al 2021). The analysis also suggests a dearth of research to create a new feature representation that suits the characteristics of the Hausa Language. Furthermore, by expanding the work for Hausa Natural Language Processing across many domains and creating reference structures across several disciplines.
The study by Hassonah et al (2020) postulated for the adoption of a hybrid machine learning strategy in 2020 to enhance sentiment analysis given that an SVM classifier was used to develop a classification approach that is based on "Positive, Negative, and Neutral" classes and the MVO and Relief models included two feature selection methods. The proposed approach was also assessed using Twitter data. The results of the studies showed that the suggested method performed better than more well-known methods.

The work by Muhammad et al (2022), introduced the first large-scale human-annotated Twitter sentiment dataset for the four most widely spoken languages in Nigeria—Hausa, Igbo, Nigerian-Pidgin, and Yoruba—consisting of around 30,000 annotated tweets per language, including a significant fraction of code-mixed tweets. They proposed text collection, filtering, processing, and labeling methods that enabled them to create datasets for these low-resource languages. They evaluated a range of pre-trained models and transfer strategies on the dataset and found out that language-specific models and language-adaptive fine-tuning generally perform best.

Sani et al (2022) developed a text classification framework for Hausa sentiment analysis based on Multinomial Na'ive Bayes (MNB) and logistic Regression algorithm using the count vectorizer and TF-IDF methods. A significant result was achieved in text categorization performance with the help of the model. Their conclusion suggested that for future work on sentiment analysis, it is necessary to work on abbreviations and acronyms, as most

Science World Journal Vol. 19(No 1) 2024 https://dx.doi.org/10.4314/swj.v19i1.13
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

internet users tends to write comments, replies or opinions by shortening some words which may reduce the sentiment polarity.
For the purpose of this research, the dataset was generated from BBC Hausa Twitter handle and then preprocessed using polyglot. Polyglot is a python package that can process a lot of languages without the need to translate them to English before acting upon. Naïve Bayes algorithm and Logistic Regression were used as classifier. Naïve Bayes is considered as the easiest or not too complex and quickest classifier as this has offered novel results to many researchers adopting it. Natural Language Processing is useful in enhancing grammar correctness, speech to text conversion, and language translation automation. It also analyzes text allowing machines to understand how human speak. (Anupama et al, 2020)
According to the findings of this study, abbreviation and acronym haven't gotten much attention in corpus linguistics despite the availability of electronic resources. This study aims to develop an improved Hausa Sentiment Dataset for the enhancement of sentiment analysis with abbreviation and acronyms.
This research developed an improved Sentiment Dataset for the enhancement of sentiment analysis by enhancing Hausa Sentiment dataset with abbreviation and acronym text and developing a text classification model for Hausa sentiment analysis using Multinomial Na'ive Bayes (MNB) and logistic Regression algorithm using the count vectorizer where at the end the result was evaluated for accuracy.
The remainder of this paper is organized as follows: Section 2 presents the methodology of our proposed approach. Section 3 shows the result obtained from the study and section 4 presents the conclusion.

MATERIALS AND METHODS

The approach of this research is a model for sentiment analysis that uses Multinomial Naïve Bayes (MNB) and Logistic Regression (LR) classification algorithm to analyze Hausa text lexicon using count vectorizer method by adopting the work of Sani et al (2022). The twitter dataset was annotated as positive, negative and neutral by annotators. The updated dataset created contains Hausa words in abbreviations and some acronyms.
The polyglot preprocessing package of python was used for the preprocessing. The dataset was further trained using machine learning Multinomial Naïve Bayes and logistic regression algorithm and in order to evaluate the effectiveness and usefulness of the classifiers the result was evaluated for accuracy.

Figure 1 shows an overview of the approach used for twitter sentiment analysis for Hausa abbreviated and acronym text.

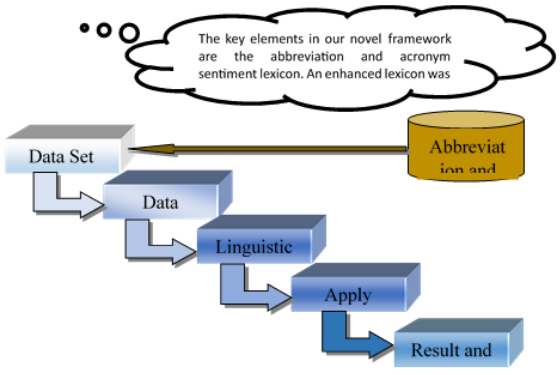

Figure 1: Overview of the Sentiment Analysis Framework.

RESAERCH METHODOLOGY

As show in figure 1 above this work comprises of four stages: the dataset, data pre-processing, linguistic data processing using NLP, apply the machine learning algorithm and finally the result and analysis, which are explained below.

1. Dataset: The Hausa dataset was generated from BBC Hausa Twitter handle texts then manually updated with the abbreviation and acronym text and further manually annotated as positive, negative or neutral. The dataset created contains Hausa words in abbreviations and some acronyms.

2. Data Pre-processing: This is where the extracted data is modified to eradicate the inconsistencies and upgrade the quality of the data. The data was filtered and became noise free at this stage. This phase was achieved by the use of a python library.

- ***Polyglot:*** Is a natural language pipeline that supports very large multilingual applications. It offers a comprehensive analysis and has a practical application in compatibility. It is an open source python package used for manipulating text and extracting useful information from it. Some of its features are; language detection, tokenization, name entity recognition, part of speech tagging, sentiment analysis and lots more.

3. Linguistic Data Processing using NLP: Analyzing data requires some basic steps; first, the text was prepared in a proper text format. The second step involves tokenization/feature extraction (which means dividing the data into different set of statement so that the computer understands very well). The third step is detection and negation which involves targeting the keyword in the data, if the word is found return "True" to verify else "False" for negated. The dependency parser analyzes the grammatical structure of the sentence if the value is "True". Co-reference parser analyzes the expression and it is the main object in NLP. The last step involves analyzing the result.

Feature Extraction: The feature consists of language information and features such as words, word type, positive and negative sentiment score. All this data can be transformed to various other formats for the purpose of fitting it according to the requirements of the machine learning algorithm in use. The output of this process is then fed to the classifier which is able to perform the sentiment classification of the given document.

1. **Applying Machine Learning Algorithm:** Classification of text are tested against two classification algorithms, the Multinomial Naïve Bayes and Logistic Regression.
   - ***Naïve Bayes:*** is a probabilistic machine learning algorithm that can be used in a wide variety of classification task. It is a commonly used method for text classification due to its effective grading assumptions, quick and easy implementation. The classifiers of Naïve Bayes (NB) are a family of classifiers based on Bayes' popular probability theorem. The Multivariate Bernoulli Naïve Bayes model (BNB) is done to classify documents and treat the absence of each word as a logical attribute as in one of the initial statistical models of language. It is well thought out but it only focuses on the appearance of words which make it a baseline for text classification. In BNB, when a word appears in the document, the value of the attribute equivalent to that word is written either as one, otherwise zero. The

Multinomial Naïve Bayes (MNB) was proposed as an improved method of BNB. The MNB assumes that the document is a bag of words and takes word frequency and information into account.

- ***Logistic Regression:*** is a supervised machine learning algorithm method that help to predict the probability of events that have binary outcome. It is part of the regression family that involves predicting outcomes based on quantitative relationships between variables. It computes the sum of input features and calculates the logistic of the result. Generally, the result of the response variable is called success and failure, yes or no, or true and false represented by 1 and 0 respectively.

System Setup

The experimental setup was implemented with Windows 10 operating system using jupyter notebook. The goal is to figure out how to adapt to the approach for Hausa Sentiment Analysis based on MNB and Logistic Regression algorithm using the count vectorizer, along with python libraries for NLP.

RESULTS

The outcome of this research is an enhanced Hausa dataset with acronyms and abbreviations (i.e., adapting the Dataset), then adopt the evaluation procedure of Sani et al. (2022)

The accuracy of a classifier on a given dataset is the percentage of the correctly classified tuples by the classifier.

To analyze the effect of enhancing the data set with abbreviation and acronym on sentiment analysis task, Table 1 shows an estimated predictive result of Multinomial Naïve Bayes performance and table 2 shows the estimated predictive result of Logistic Regression performance. Also figure 3 shows performance comparison of MNB and LR performance for overall accuracy.

Table 1: The summary of performance of the models using both approaches based on multinomial Naïve Bayes.

PARAMETER	WITHOUT ABBREVIATION AND ACRONYM	WITH ABBREVIATION AND ACRONYM
Precision	0.84	0.84
Recall	0.76	0.83
F1-Score	0.80	0.83
Accuracy	80%	84%

Table 2: The summary of performance of the models using both approaches based on Logistic Regression.

PARAMETER	WITHOUT ABBREVIATION AND ACRONYM	WITH ABBREVIATION AND ACRONYM
Precision	0.83	0.85
Recall	0.93	0.85
F1-Score	0.88	0.85
Accuracy	86%	85%

The benchmarking is to test the Sentiment of

1. Adopted method + Adopted Dataset
2. Adopted Method + Adapted Dataset

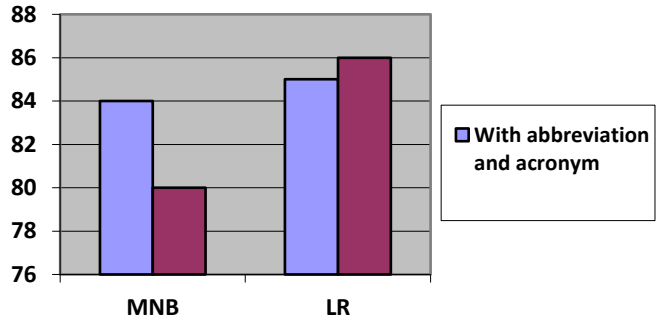

Figure 3: Performance Comparison of MNB and LR based on Accuracy

To evaluate the effectiveness of the classifiers and to produce a more accurate classification procedure, experimental result was analyzed for accuracy. Where the result shows that performance of the models using both approaches. Based on multinomial Naïve Bayes, accuracy is higher by 4% with sentiment conveyed by abbreviations and acronyms affecting the surrounding text.

Conclusion

An enhanced lexicon was created and combined with the existing lexicon. The corpus was annotated manually to rate the abbreviation and acronyms in their lexicon for their associated sentiment. The annotators assigned ratings of positive, negative and neutral. Based on the practical experience from this research, polyglot worked incredibly well in manipulating texts by extracting useful information. The result shows that in addition to normal preprocessing techniques of the social media stream, understanding, interpreting and resolving ambiguity in the usage of abbreviations and acronyms lead to improved accuracy of algorithms with evidence in the experimental result.

REFERENCE

Tang et ai (2015). Sentiment Analysis Wikipedia. https://en.m.wikipedia.org/wiki/sentiment_analysis

Hassonah, M. A., Al-Sayyed, R., Rodan, A., Al-Zoubi, A. M., Aljarah, I., & Faris, H. (2020). An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter. *Knowledge-Based Systems*, *192*, 105353. https://doi.org/10.1016/j.knosys.2019.105353

James Lappeman, Robyn Clark, Jordan Evans, Lara Sierra-Rubia, Patrick Gordan. 2020. 'Online Sentiment Analysis using Human Validation' Volume 7. https://doi.org/10.106/j.mex.2020.100867

Mika V. Mantyla, Daniel Graziotin, Mikka Kuutila (2017) 'The evolution of sentiment analysis - A review of research topics, venue and top cited papers, Computer Science Review', volume 27, pg 16-32, ISSN 1574-0137, https://doi.org/10.1016/j.cosrev.2017.10.002

Muhammad Sani, Abubakar Ahmad and Hadiza S. Abdulazeez (2022). 'Sentiment Analysis of Hausa Language Tweet Using Machine Learning Approach' Journal of Research in Applied Mathematics Volume 8 ~ Issue 9 (2022) pp: 07-16 ISSN(Online): 2394-0743 ISSN (Print): 2394-073.

Zakari, R. Y., Lawal, Z. K., & Abdulmumin, I. (2021). A Systematic Literature Review of Hausa Natural Language Processing. *International Journal of Computer and Information*

Technology(2279-0764), *10*(4). https://doi.org/10.24203/IJCIT.V10I4.86

Abegunde, O., Iyanda, A. R., & Ninan, D. O. (2019). *Design Issues in Sentiment Analysis for Yorùbá Written Text*. *3*(1), 12–25.

Ali Hasan, Sana Moin, Ahmad Karim and ShahaboddinShamshirband (2018)" Machine Learning-Based Sentiment Analysis for Twitter Accounts" 2018 by the authors. Licensee MDPI, Basel, Switzerland.

Amitava Das and Bjorn Gamback. 2014. Identifying languages at the word level in code-mixed Indian social ¨ media text. *In Proceedings of the 11th International Conference on Natural Language Processing*, pages 378– 387, Goa, India, December. NLP Association of India.

Anapuma B S, Rakshit D B, Rahul Kumar M, Navaneeth M. (2020) 'Real Time Sentiment Analysis using Natural Language Processing', *International Journal of Engineering Research and Technology (IJERT),* ISSN:2278-0181, Vol 9, issued 07, july 2020 DOI: 10.17577/IJERTV91S070406 <https://www.researchgate.net/publication/343288167_Real_Time_Twitter_Sentiment_Analysis_using_Natural_Language_Processing>

Arouna Konate and Du Ruiying 2018, 'Sentiment Analysis of Code-Mixed Bambara-French Social Media Text Using Deep Learning Techniques', *Wuhan University Journal of Natural Sciences,* Vol. 23 No.3, pg 237-243, Doi: https://doi.org/10.1007/s11859-018-1316-z.

David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. 2018. The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. ACM Trans. Manage. Inf. Syst. 9, 2, Article 5 (August 2018), 29 pages. https://doi.org/10.1145/3185045

Joan-Frances Fondevila-Gascon, Pedro Mir-Bernal, Elena Puiggros-Roman et al 2016, 'Sentiment Analysis as a Qualitative methodology to analyze Social Media: Study Case of Tourism'. *1st International Symposium on Qualitative Research,* Volume 5, ISQR2016 Proceedings.

Kingsley A. Ogudu and Dahj Muwawa Jean Nestor (2019) 'Sentiment Analysis Application and Natural Language Processing for Mobile Network Operators' Support on Social Media', IEEE

Marjan Kamyab, Ran Tao, Mohammad Hadi Mohammadi and Abdul Rasool (2018) 'Sentiment Analysis on Twitter: A Text Mining Approach to the Afganistan Status Reviews' pp23-25 DOI: https://doi.org/10.1145/3293663.3293687.

Masawee Masdisornchote 2015, 'A Sentiment Analysis Framework in Implicit Opinions for Thai Language',*41st Annual Confrerence of the IEEE Industrial Electronics Society,IECON 2015*, pg 00357-00361.IEEE

Mohammed Arshad Ansari and Sharvari Govilkar 2018, 'Sentiment Analysis of Code for the Transliterated Hindi and Marathi Texts', *International Journal on Natural Language Computing (IJNLC),* Vol.7, No.2, DOI: 10.5121/ijnlc.2018.7202.

Öztürk, N., Ayvaz, S. 2017, Sentiment Analysis on Twitter: A Text Mining Approach to the Syrian Refugee Crisis, Telematics and Informatics, doi: https://doi.org/10.1016/j.tele.2017.10.006

Parth Patwa et al (2020) 'SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed tweets' *Procedings of the 14th international workshop on semantic Evaluation,* pp 774-790.