INTERNATIONAL JOURNAL
OF SOFT COMPUTING

SINCE 2000
MEDWELL PUBLICATIONS

Sentiment Analysis on Nigerian Tweet Using Data Mining Techniques

[1]Amaechi Chinedum and [2]Okeke Ogochukwu
[1]*Department of Computer Science, Nnamdi Azikwe University, Awka, Anambra State, Nigeria*
[2]*Department of Computer Science, Chukwuemeka Odumegwu Ojukwu University, Uli, Anambra State, Nigeria*

Key words: Sentiment analysis, pre-processing, opinion mining, Nigerians Tweets, Twitte

Corresponding Author:
Amaechi Chinedum
Department of Computer Science, Nnamdi Azikwe University, Awka, Anambra State, Nigeria

Page No.: 25-28
Volume: 16, Issue 02, 2021
ISSN: 1816-9503
International Journal of Soft Computing
Copy Right: Medwell Publications

Abstract: Probing sentiments in social media poses a task to natural language processing because of the complexity and variability in the different dialect expression, noisy terms in form of local slang, abbreviation, acronym, emoticon and spelling error coupled with the availability of real-time content. Most of the knowledge based approaches for resolving local Nigerian slangs, abbreviation and acronym do not consider the issue of ambiguity that evolves in the usage of these noisy terms. This research implements an improved framework for social media feed pre-processing that leverages on the adapted Lesk algorithm to facilitate pre-processing of social media feeds. The results from the experimental evaluation revealed an improvement over existing methods when applied to supervised learning algorithms in the task of extracting sentiments from Nigeria-Igbo tweets with an accuracy of 90%.

INTRODUCTION

Opinions, feedback and critiques provided by internet users show attitudes and sentiments towards specific topics, products or services[1]. Sentiment Analysis is being used to robotically detect speculations, emotions, opinions and evaluations in social media content[2]. Unlike carefully created news and other literary web contents, social media streams present different difficulties for analytics algorithms because of their extensive scale, short nature and uses of slangs. There are few pre-processing techniques that factor in local Nigerian slangs.

A number of the knowledge-based approaches for resolving these noisy terms do not reflect the subject of ambiguity and culture that evolves in their usage[3]. Furthermore, social media write-ups content a short length of messages, use of enthusiastically evolving, lopsided, informal and abbreviated words. These make it difficult for techniques that build on them to perform effectively and efficiently[4, 5].

Kolajo *et al*.[6] proposed proposes an improved framework for social media feed pre-processing that leverages on the combination of integrated local knowledge bases and adapted Lesk algorithm to facilitate pre-processing of social media feeds.

Related work: The first step to sentiment analysis is the Text preprocessing[7]. Various scholars have considered the effect and influence of pre-processing (which include tokenization, removal of stop-words, lemmatization, fixing of slangs, redundancy elimination) on the accuracy of result of techniques building on them for sentiment analysis and consistently agreed that when social media stream data are well deduced and represented, it leads to substantial improvement of sentiment analysis result).

The preprocessing stages include removal of HTML tags, stop word removal, negation handling, stemming and expansion of abbreviation using pattern recognition and regular expression techniques. The problem here is that representing acronym and slangs based on co-occurrence does not take care of ambiguity[8].

The effect of pre-processing techniques on Twitter sentiment Analysis was explored by Krouska *et al*.[9], data preprocessing is a crucial step in sentiment analysis, since selecting the appropriate preprocessing methods, the suitably classified instances can be improved. Appraisal of sentiment polarity classification methods for Twitter text and the role of text preprocessing in sentiment analysis were discussed.

Their results clarified that with applicable feature selection and representation, sentiment analysis accuracies can be improved. However, they said it is worthy to investigate further the accessible preprocessing options in order to find the optimal settings. In the same vein[10, 4] investigated the role of text pre-processing and established that an appropriate combination of preprocessing tasks advances classification accuracy.

The impact of pre-processing methods on Twitter sentiment classification was explored by Bao *et al*.[11]. They evaluated the effects of URLs, negation repeated letters, stemming and lemmatization by using the Stanford Twitter Sentiment Dataset. The result of the study showed an improvement in accuracy when negation transformation, URLs feature reservation and repeated letters normalization is employed while lemmatization and stemming reduce the accuracy of sentiment classification.

Singh and Kumari[4] focuses to identify the importance of slang words and to measure their impact on sentiment of the tweet. Since, research efforts have not been directed towards the handling all of slang/abbreviation/acronym as well as resolving ambiguity in the usage of noisy terms based on contextual information. Kolajo *et al*.[6] proposed proposes an improved framework for social media feed pre-processing that leverages on the combination of integrated local knowledge bases and adapted Lesk algorithm to facilitate pre-processing of social media feeds.

MATERIALS AND METHODS

Data collection: The dataset was extracted from tweets of twitter users from Nigerians. The dataset focused on politics, on Nigerians' sentiments on the administration of State government of Anambra under Governor Willie Obiano, Anambra Nigeria. A user interface was built around an underlying API provided by Twitter to collect tweets, retweets on the handle @WillieMObiano. The total tweets extracted was 5,000. 80% was used as training data while 20% was used as test data. The General Preprocessing Method (GTPM) and the Improved Preprocessing Method (ITPM) are depicted in Fig. 1 and 2, respectively.

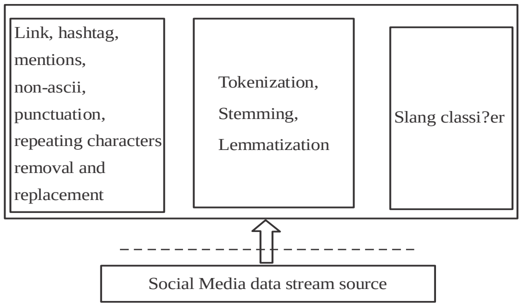

Fig. 1: General textual preprocessing method

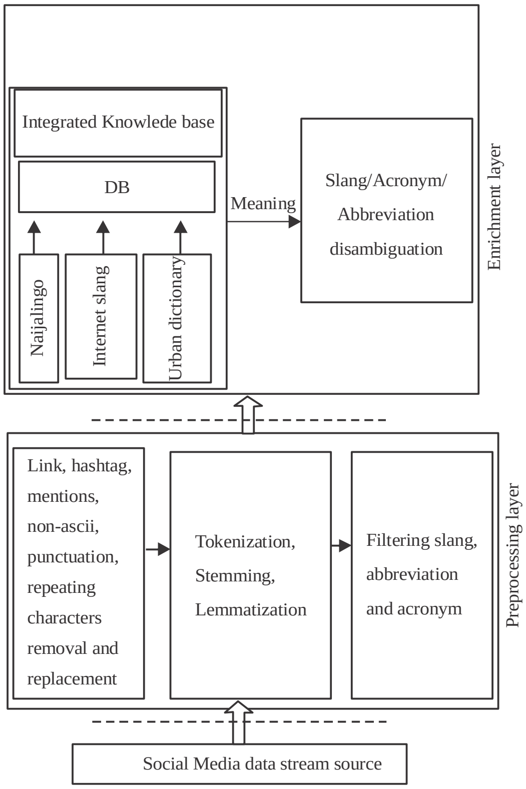

Fig. 2: The Improved Preprocessing Method (ITPM)

Data preprocessing: From the data stream collected, Tags, URLs, mentions and non-ASCII characters were automatically removed using a regular expression. This was followed by tokenization and normalization. Thereafter, slangs, abbreviation, acronyms and emoticons were filtered from the tweets using corpora of English and Igbo words in Natural Language Toolkit (NLTK). The filtered slangs/abbreviation/acronyms are then passed to the Integrated Knowledge Base (IKB) for further processing.

Data enrichment and resolving ambiguity: The IKB is an API centric resource that links with three internet sources which are Naijalingo, Urban dictionary, as well as Internetslang.com. Naijalingo is included in order to take care of tainted English regularly found in social media feeds in Nigeria. Naijalingo is vital in order to resolve ambiguity in the usage of slang/abbreviation/acronym in tweets made from Nigeria users. The ITPM structure will permit the assimilations of any other local knowledge base that will fit the contexts in order to capture slang/abbreviation/acronym that has locally defined meaning.

The next stage is to extract connotations and meaning of slang/abbreviation/acronym terms from IKB. Ambiguous slang/abbreviation/acronym terms were resolved by leveraging adapted Lesk algorithm based on the context in which they appear in the tweet. For ambiguous slang/abbreviation/acronym, there is a need to obtain the best sense from the group of various definitions in the IKB based on how it is used in the tweet (Listing 1). The tweet (twt) in which this slang/abbreviation/acronym term appears and the extracted usage examples (st) is represented as a setdata structure. After this, intersection operation between the tweet and each of the usage examples(relatedness(st,sjk)) is performed.

Listing 1:

```
Input: tweet text
Output: enriched tweet text
// Process to disambiguate ambiguous
// slang/acronyms/abbreviation in tweets
// by adapting Lesk algorithm over usage
// in the integrated knowledge base (ikb)
Notations:
slngs: slangs; acrs: acronyms; abbrs: abrreviations
sab: slang/acronym/abbreviation;
sabt: slang/acronym/abbreviation term
st: ith usage example of target word sabt found in the ikb
procedure disambiguate_all_slngs/acrs/abbrs
for all sab(word) in input do //the input is the
//extracted slang/abbreviation/acronym
//from tweet
best_sense=disambiguate_each_
            slng/acr/abbr(sabt)
            display best_sense
end for
end procedure
function disambiguate_each_ slng/acr/abbr(sabt)
// target word represent
//slang/acronym/abbreviation in the tweet
st→ith usage example of target word sabt
found in the ikb
twt→ the current tweet being processed
sense→ { s1, s2, …sn | m 1} // sense
is the set of senses of st found in the ikb
for all st of the target word sabtdo
// stiis the ith usage example of target
//word sabtfound in the ikb
        score i = 0
for i= 1 to ndo
// n is the total number of
//usage examples for each
//slang/acronym/abbreviation
//in tweet
for twtof word sabt
temp_score k =
relatedness(st,twt)
end for
best_score =
max(temp_score)
score i += best_score
end for
end for
return si∈Sense
// siis the ithusage example from the ikbthat
best matches
// slang/acronym/abbreviation in the tweet
map siwith defi(where defi∈definition)
replace sabtin tweet with defi
end function
```

RESULTS AND DISCUSSION

The improved ITPM framework was benchmarked with the General Textual Pre-Processing Method (GTPM) and Arc2Vec Framework by running them on one classifiers.

The GTPM (i.e., general pre-processing method) does not take care of slang/abbreviation/acronym ambiguity. The classifiers used for the benchmarking were Support Vector Machine (SVM) and Naïve Bayes to extract sentiments from tweets. The goal is to ascertain the impact of this on the algorithms building on them. The result of the sentiment classification of Nigerian tweets dataset is shown in Table 1.

In Table 1, the result of the experiment reveal that the ITPM outperformed the GTPM and also improve the accuracy of the result obtained. This highlights the significance of using a localized knowledge base in pre-processing social media feeds to fully capture the loud terms that are domiciled in the social media feeds coined from a particular location.

Table 1: Sentiment classification results by SVM and Naïve Bayes

Methods	Algorithm	Accuracy (%) unigram	Accuracy (%) bigram
GTPM	SVM	76	75
ITPM		89	90
GTPM	Naïve Bayes	76	78
ITPM		90	90

CONCLUSION

This study affirms the implementation of the use of an improved approach to preprocessing of social media streams as proposed by Kolajo *et al*.[6]:

- Assimilating localized knowledge sources as extension to knowledge-based approaches
- Capturing the rich local Nigerian semantics embedded in slangs, abbreviation and acronym
- Resolving ambiguity in the usage of slangs, abbreviation and acronym to better interpret and understand social media feeds

REFERENCES

01. Manjula, A. and R.M. Rama, 2019. Sentiment analysis on social media. Int. J. Comput. Eng. Res. Trends, Vol. 6. 10.22362/ijcert
02. Thakkar, H. and D. Patel, 2015. Approaches for sentiment analysis on Twitter: A state-of-art study. Department of Computer Engineering, Montreal, Quebec, Canada.
03. Sabbir, A., A. Jimeno-Yepes and R. Kavuluru, 2018. Knowledge-based biomedical word sense disambiguation with neural concept embeddings. IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE), October 23-25, 2017, IEEE, pp: 163-170.
04. Singh, T. and M. Kumari, 2016. Role of text preprocessing in twitter sentiment analysis. Procedia Comput. Sci., 89: 549-554.
05. Zhan, J. and B. Dahal, 2017. Using deep learning for short text understanding. J. Big Data, Vol. 4. 10.1186/s40537-017-0095-2
06. Kolajo, T., O. Daramola and A. Adebiyi, 2019. Sentiment analysis on naija-tweets. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, July 28-August 2, 2019, Student Research Workshop, pp: 338-343.
07. Jianqiang, Z. and G. Xiaolin, 2017. Comparison research on text pre-processing methods on twitter sentiment analysis. IEEE Access, 5: 2870-2879.
08. Haddi, E., X. Liu and Y. Shi, 2013. The role of text pre-processing in sentiment analysis. Procedia Comput. Sci., 17: 26-32.
09. Krouska, A., C. Troussas and M. Virvou, 2016. The effect of preprocessing techniques on Twitter sentiment analysis. 7th International Conference on Information, Intelligence, Systems & Applications (IISA), July 13-15, 2016, IEEE, pp: 1-5.
10. Uysal, A.K. and S. Gunal, 2014. The impact of preprocessing on text classification. Inf. Process. Manage., 50: 104-112.
11. Bao, Y., C. Quan, L. Wang and F. Ren, 2014. The role of pre-processing in Twitter sentiment analysis. Proceedings of the International Conference on Intelligent Computing (ICIC'14), August 3-6, 2014, Springer, Taiyuan, China, ISBN:978-3-319-09338-3, pp: 615-624.