Sentiment Analysis on Naija-Tweets

Taiwo Kolajo*
Covenant University, Ota, Nigeria
taiwo.kolajo@stu.cu.edu.ng
Federal University Lokoja, Kogi State, Nigeria
taiwo.kolajo@fulokoja.edu.ng

Olawande Daramola
CPUT, Cape Town, South Africa
daramolaj@cput.ac.za

Ayodele Adebiyi
Covenant University, Ota, Nigeria
ayo.adebiyi@covenantuniversity.edu.ng
Landmark University, Omu-Aran, Nigeria
ayo.adebiyi@lmu.edu.ng

Abstract

Examining sentiments in social media poses a challenge to natural language processing because of the intricacy and variability in the dialect articulation, noisy terms in form of slang, abbreviation, acronym, emoticon, and spelling error coupled with the availability of real-time content. Moreover, most of the knowledge-based approaches for resolving slang, abbreviation, and acronym do not consider the issue of ambiguity that evolves in the usage of these noisy terms. This research work 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. 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-origin tweets with an accuracy of 99.17%.

1 Introduction

Sentiment Analysis is being used to automatically detect speculations, emotions, opinions, and evaluations in social media content (Thakkar and Patel, 2015). Unlike carefully created news and other literary web contents, social media streams present various difficulties for analytics algorithms because of their extensive scale, short nature, slang, abbreviation, grammatical and spelling errors (Asghar et al., 2017). Most of the knowledge-based approaches for resolving these noisy terms do not consider the issue of ambiguity that evolves in their usage (Sabbir et al., 2017). These challenges, which inform this research work, make it necessary to seek improvement on the performance of existing solutions for pre-processing of social media streams (Carter et al., 2013; Ghosh et al., 2017; Kuflik et al., 2017).

Due to language complexity, analysing sentiments in social media presents a challenge to natural language processing (Vyas and Uma, 2018). Moreover, social media content is characterized with a short length of messages, use of dynamically evolving, irregular, informal, and abbreviated words. These make it difficult for techniques that build on them to perform effectively and efficiently (Singh and Kumari, 2016; Zhan and Dahal, 2017).

The short nature of social media streams coupled with no restriction in the choice of language has informed the usage of abbreviation, slang, and acronym (Atefeh and Khreich, 2015; Kumar, 2016). These noisy but useful terms have their implicit meanings and form part of the rich context that needs to be addressed in order to fully make sense of social media streams (Bontcheva and Rout, 2014). Just like there is ambiguity in the use of normal language there is also ambiguity in the usage of slang/abbreviation/acronym because they often have context-based meanings, which must be rightly interpreted in order to improve the

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 338–343
Florence, Italy, July 28 - August 2, 2019.

results of social media analysis. There is a dearth of social media streams preprocessing geared at resolving slang, abbreviation and acronym as well as ambiguity issues that erupt as a result of their usage (Mihanovic et al., 2014; Matsumoto et al., 2016).

2 Related Work

Many researchers have studied the effect and impact of pre-processing (which ranges from 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 unanimously agreed that when social media stream data are well interpreted and represented, it leads to significant improvement of sentiment analysis result.

Haddi et al. (2013) presented the role of text pre-processing in sentiment analysis. The pre-processing 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 abbreviation based on co-occurrence does not take care of ambiguity. The impact of pre-processing methods on Twitter sentiment classification was explored by Bao et al. (2014) 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. In the same vein, Uysal and Gunal (2014) and Singh and Kumari (2016) investigated the role of text pre-processing and found out that an appropriate combination of pre-processing tasks improves classification accuracy.

Smailovic et al. (2014) and Ansari et al. (2017) investigated sentiments analysis on twitter dataset. Their pre-processing method along with tokenization, stemming and lemmatization includes replacement of user mention, URLs, negation, exclamation, and question marks with tokens. Letter repetition was replaced with one or two occurrences of the letter. From the result of their experiments, it was concluded that pre-processing twitter data improves techniques building on them.

The pre-processing method adopted by Ouyang et al. (2017) and Ramadhan et al. (2017) includes deletion of URLs, mentions, stop-words, punctuation, and stemming. Ramadhan et al. (2017) added the handling of slang conversion in their work although the authors did not state how the slang conversion was done. Jianqiang and Xiaolin (2017) discussed the effect of pre-processing and found that expanding acronyms and replacing negation improve classification while removal of stop-words, numbers or URLs do not yield any significant improvement. On the contrary, Symeonidis et al. (2018) evaluated classification accuracy based on pre-processing techniques and found out that removing numbers, lemmatization, and replacing negation improve accuracy. Zhang et al. (2017) presented Arc2Vec framework for learning acronyms in twitter using three embedding models. However, the authors did not take care of contextual information. From the review, most 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.

3 Methodology

3.1 Data Collection

The dataset (referred to as Naija-tweets in this paper) was extracted from tweets of Nigeria origin. The dataset focused on politics in Nigeria. A user interface was built around an underlying API provided by Twitter to collect tweets based on politics-related keywords such as “politics”, “governments”, “policy”, “policymaking”, and “legislation”. The total tweets extracted was 10,000. These were manually classified into positive (1) or negative (0) by three experts in sentiment analysis. 80% was used as training data while 10% was used as test data and 10% for dev set. The general preprocessing method (GTPM), Arc2Vec Framework and the proposed preprocessing method (PTPM) are depicted in figure 1 (a), (b) and (c) respectively.

3.2 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 words in natural language toolkit (NLTK). The filtered

slangs/abbreviation/acronyms are then passed to the Integrated Knowledge Base (IKB) for further processing.

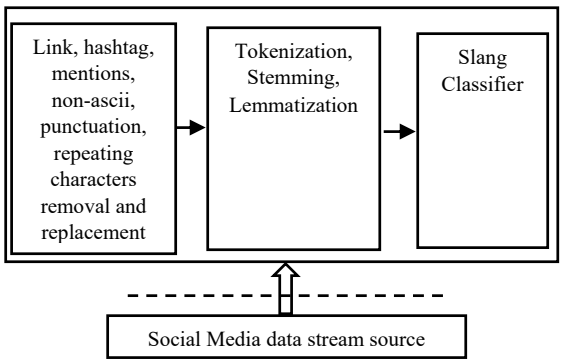

Figure 1a: General Textual Preprocessing Method

3.3 Data Enrichment

The IKB is an API centric resource that communicates with three (3) internet sources which are Naijalingo, Urban dictionary, and Internetslang.com. Naijalingo is included in order to take care of adulterated English commonly found in social media feeds in Nigeria and some parts of Anglophone West Africa. Moreover, the presence of Naijalingo is very important in order to resolve ambiguity in the usage of slang/abbreviation/acronym in tweets originated from Nigeria.

The PTPM framework will allow the integrations of any other local knowledge base that may suit some other contexts in order to capture slang/abbreviation/acronym that has locally defined meaning. The IKB API is also responsible for slang/abbreviation/acronym disambiguation, spelling correction and emoticon replacement. The IKB is to cater for slangs, abbreviation or acronyms, and emoticons found in tweets and to provide a single platform where all these knowledge sources can be easily referenced. About two million slang/abbreviation/acronym and emoticons terms were crawled from these knowledge sources and stored on MongoDB. All lexicons that were used for the enrichment of the collected tweets in the IKB were derived from Naijalingo, Urban dictionary, and Internet slang knowledge sources. A lexicon of noisy terms (slang/acronym/abbreviation) in the IKB has four elements which are (1) slang/acronym/abbreviation term, (2) a descriptive phrase, (3) example (i.e. how it is being used) and (4) related terms from the three knowledge sources. Each term can have multiple entries which imply that each term can be associated with any number of descriptive phrases. Each term is seen as a key-value pair where each term is the key and a network of associated descriptive phrases represent the value.

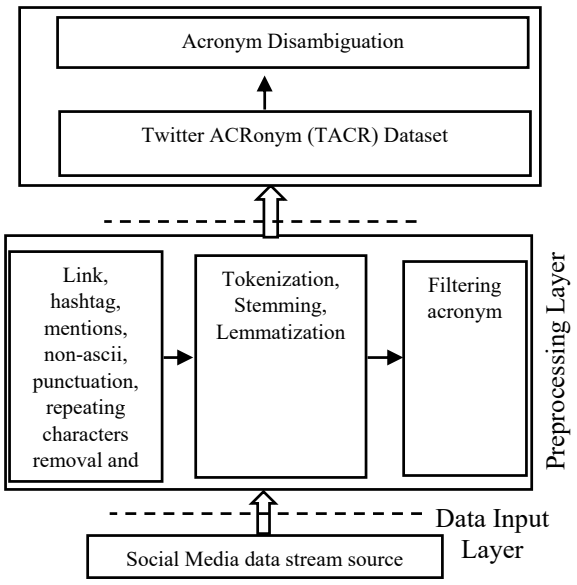

Figure 1b: Arc2Vec Preprocessing Framework

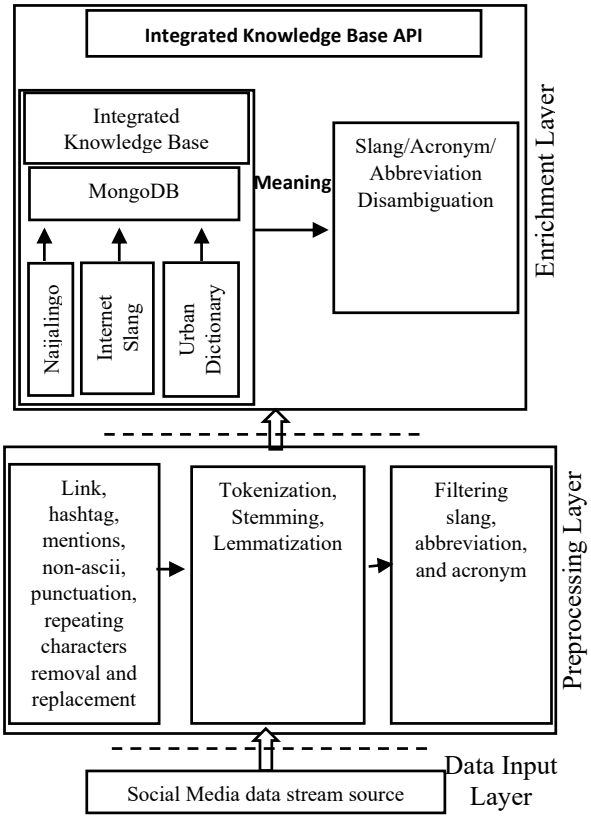

Figure 1c: Proposed Textual Preprocessing Method

3.4 Resolving Ambiguity in Slang/Abbreviation/Acronym

The next stage is to extract meanings 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 pool of various definitions in the IKB based on how it is used in the tweet (see Listing 1).

Usage examples (*st*) with a total number, *n*, mapped to various definitions of the slang/abbreviation/acronym term (*sabt*) that is to be interpreted are extracted from the IKB, where there are no usage examples mapped to definitions for a particular *sabt* in the ikb, the definitions are used instead. The tweet (*sjk*) in which this slang/abbreviation/acronym term appears and the extracted usage examples *(st*) is represented as a set data structure. After this, intersection operation between the tweet and each of the usage examples (*relatedness*(*st,sjk*)) is performed. Then the usage example (*sti*) with the highest intersection value (*best_score*) is selected. A lookup of the meaning attached to the selected usage example from the IKB is performed (*map sense_i with definition*), the definition is then used to replace the slang in the tweet as the best possible semantically meaningful elaboration of each slang/abbreviation/acronym based on how it is being used in the tweet.

3.5 Feature Extraction

For feature extraction, a total of 3,000 unigrams and 8,000 bigrams were used for vector representation. Each tweet was represented as a feature vector of these unigrams and bigrams. For convolutional neural network, Glove twitter 27B 200d was used for the dataset vector representation.

4 Result

The proposed PTPM framework was benchmarked with the General Textual Pre-processing Method (GTPM) and Arc2Vec Framework by running them on three classifiers. The GTPM (i.e. general pre-processing method) does not take care of slang/abbreviation/acronym ambiguity issue while that of Arc2Vec framework only took care of

Listing 1. Adapted Lesk Pseudocode

```
Input: tweet text
Output: enriched tweet text
// Procedure to disambiguate ambiguous
// slang/acronyms/abbreviation in tweets
// by adapting Lesk algorithm over usage
// examples of slangs/acronym/abbreviation
// found in the integrated knowledge base (ikb)
Notations:
 slngs: slangs; acrs: acronyms; abbrs: abrreviations
 sab: slang/acronym/abbreviation;
 sabt: slang/acronym/abbreviation term
 st: i^th 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 → i^th usage example of target word sabt
        found in the ikb
      sjk → 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 sabt do
                // sti is the i^th usage example of target
                //word sabt found in the ikb
                score i = 0
                for i= 1 to n do
                        // n is the total number of
                        //usage examples for each
                    //slang/acronym/abbreviation
                        //in tweet
                        for sjk of word sabt
                            temp_score k =
                            relatedness(st,sjk)
                        end for
                        best_score =
                        max(temp_score)
                        score i += best_score
                end for
        end for
        return si ∈ Sense
        // si is the i^th usage example from the ikb that
        best matches
        // slang/acronym/abbreviation in the tweet
        map si with defi (where defi ∈ definition)
        replace sabt in tweet with defi
end function
```

Acronym. The classifiers used for the benchmarking were support vector machine (SVM), multi-layer perceptron (MLP), and convolutional neural networks (CNN) to extract sentiments from tweets. The essence of running both the general pre-processing method – GTPM, Arc2Vec framework, and the proposed SMFP framework on the classifiers was to compare the results of capturing of slangs/acronym/abbreviation and resolving ambiguity in social media streams slang/acronym/abbreviation have been undertaken, and whether it has not been undertaken. The goal is to ascertain the impact of this on the algorithms building on them. The result of the sentiment classification of naija_tweets dataset is shown in Tables 1 and 2

Method	Algorithm	Accuracy (%) Unigram	Accuracy (%) Bigram	Accuracy (%) Unigram + Bigram
GTPM	SVM	77.50	67.50	72.50
Arc2Vec		66.97	66.58	66.32
PTPM		**80.00**	**70.00**	**87.50**
GTPM	MLP	74.80	93.00	**99.00**
Arc2Vec		62.78	90.36	75.04
PTPM		**75.00**	**95.00**	**99.00**

Table 1. Sentiment Classification Results by SVM, Arc2Vec, and MLP

In Table 1, the result of the experiment did not only reveal that the PTPM outperformed the GTPM and Arc2Vec but there is also an improvement in the accuracy of the result obtained. This underscores the importance of using a localized knowledge base in pre-processing social media feeds to fully capture the noisy terms that are domiciled in the social media feeds originating from a particular location.

Method	Algorithm (Kernel size = 3)	Accuracy (%) 1-Con-NN	Accuracy (%) 2-Con-NN	Accuracy (%) 3-Con-NN	Accuracy (%) 3-Con-NN
GTPM	CNN	97.78	97.78	94.72	**93.61**
Arc2Vec		83.00	93.00	73.4	70.74
PTPM		**99.17**	**98.61**	**96.94**	93.33

Table 2. Sentiment Classification Results by CNN

The result presented in Table 2 also supports our argument that there should be the inclusion of localized knowledge source in pre-processing social media feeds originating from a specific location in order to better interpret slang/abbreviation/acronym emanating from such social media feeds content. It is also worthy to note that convolutional neural networks performed better than support vector machines and multilayer perceptron algorithms in tweet sentiment analysis with the accuracy of 99.17%.

5 Conclusion

This paper provides an improved approach to pre-processing of social media streams by (1) integrating localized knowledge sources as extension to knowledge-based approaches, (2) capturing the rich semantics embedded in slangs, abbreviation and acronym, and (3) resolving ambiguity in the usage of slangs, abbreviation and acronym to better interpret and understand social media streams content. The result shows that in addition to normal preprocessing techniques of the social media stream, understanding, interpreting and resolving ambiguity in the usage of slangs/abbreviation/acronyms lead to improved accuracy of algorithms building on them as evident in the experimental result.

Acknowledgements

The research was supported by Covenant University Centre for Research, Innovation, and Discovery (CUCRID); Landmark University, Omu-Aran, Osun State, Nigeria; The World Academy of Sciences for Research and Advanced Training Fellowship, FR Number: 3240301383; and the Cape Peninsula University of Technology, South Africa.

References

Ansari, A. F., Seenivasan, A., and Anandan, A. (2017). Twitter Sentiment Analysis. https://github.com/abdulfatir/twitter-sentiment-analysis

Asghar, M. Z., Kundi, F. M., Ahmad, S., Khan A., and Khan, F. (2017). T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme. Expert System, 1-19.

Atefeh, F. and Khreich, W. (2015). A survey of techniques for event detection in twitter. Computational Intelligence, 31(1), 132-164.

Bao, Y., Quan, C., Wang, L., and Ren, F. (2014). The role of text pre-processing in twitter sentiment analysis. In: D. S. Huang, K. H. Jo, and L. Wang (Eds.), Intelligent Computing Methodologies. ICIC

2014. Lecture Notes in Computer Science 8589, 615-629. Taiyuan, China: Springer.

Bontcheva, K. and Rout, D. (2014). Making sense of social media streams through semantics: A survey. Semantic Web, 5(5), 373-403. Available from: semantic-web-journal.org>swj303_0.pdf

Carter, S., Weerkamp, W., and Tsagkias, E. (2013). Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. Language Resources and Evaluation Journal, 47(1), 195-215.

Ghosh, S., Ghosh, S., and Das, D. (2017). Sentiment identification in code-mixed social media text. https://arxiv.org/pdf/1707.01184.pdf

Haddi, E., Liu, X., and Shi, Y. (2013). The role of text pre-processing in sentiment analysis. Procedia Computer Science, 17, 26-32.

Jianqiang, Z., and Xiaolin, G. (2017). Comparison research on text pre-processing methods on twitter sentiment analysis. IEEE Access, 5, 2870-2879.

Kuflik, T., Minkov, E., Nocera, S., Grant-Muller, S., Gal-Tzur, A., and Shoor I. (2017). Automating a framework to extract and analyse transport related social media content: the potential and challenges. Transport Research Part C, 275-291.

Kumar, M.G.M. (2016). Review on event detection techniques in social multimedia. Online Information Review, 40(3), 347-361.

Matsumoto, K., Yoshida, M., Tsuchiya, S., Kita, K., and Ren, F. (2016). Slang Analysis Based on Variant Information Extraction Focusing on the Time Series Topics. International Journal of Advanced Intelligence, 8(1), 84-98.

Mihanovic, A., Gabelica, H., and Kristic, Z. (2014). Big Data and Sentiment Analysis using KNIME: Online reviews vs. social media. MIPRO 2014, 26-30 May, Opatija, Croatia, (pp. 1464-1468).

Ouyang, Y., Guo, B., Zhang, J., Yu, Z., and Zhou, X. (2017). Senti-story: Multigrained sentiment analysis and event summarization with crowdsourced social media data. Personal and Ubiquitous Computing, 21(1), 97-111.

Ramadhan, W. P., Novianty, A., and Setianingsih, C. (2017, September). Sentiment analysis using multinomial logistic regression. Proceedings of the 2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC) (pp.46-49). Yogyakarta, Indonesia: IEEE.

Sabbir, A. K. M., Jimeno-Yepes, A., and Kavuluru, R. (2017, October). Knowledge-based biomedical word sense disambiguation with neural concept embeddings. 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE). Washington, DC, USA: IEEE.

Singh, T., and Kumari, M. (2016). Role of text pre-processing in twitter sentiment analysis. Procedia Computer Science 89, 549-554. Available from: https://doi.org/10.1016/j.procs.2016.06.095

Smailovic, J., Grcar, M., Lavrac, N. Znidarsic, M. (2014). Stream-based active learning for sentiment analysis. Information Sciences, 285, 181-203.

Symeonidis, S., Effrosynidis, D., and Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. Expert Systems with Applications, 110, 298-310.

Thakkar, H., and Patel, D. (2015). Approaches for sentiment analysis on twitter: A state-of-art study. arXuv preprint arXiv:1512.01043, 1-8.

Uysal, A. K., and Gunal, S. (2014). The impact of pre-processing on text classification. Information Processing and Management, 50(1), 104-112.

Vyas, V., and Uma, V. (2018). An extensive study of sentiment analysis tools and binary classification of tweets using rapid miner. Procedia Computer Science, 125, 329-335.

Zhan, J., and Dahal, B. (2017). Using deep learning for short text understanding. Journal of Big Data, 4:34. doi: 10.1186/s40537-017-0095-2

Zhang, Z., Luo, S., and Ma, S. (2017). Arc2Vec: Learning acronym representations in twitter. In: L. Polkowski et al. (Eds.) IJCRS 2017, Part I, LNAI 10313, 280-288