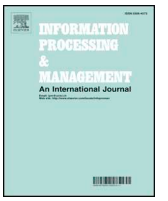
Information Processing and Management 57 (2020) 102348

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman



A framework for pre-processing of social media feeds based on integrated local knowledge base

Taiwo Kolajo[a,b,*], Olawande Daramola[c], Ayodele Adebiyi[a,d], Aaditeshwar Seth[e]

[a] Department of Computer and Information Sciences, Covenant University, Ota, Nigeria
[b] Department of Computer Science, Federal University Lokoja, Kogi State, Nigeria
[c] Department of Information Technology, Cape Peninsula University of Technology, Cape Town, South Africa
[d] Department of Computer Science, Landmark University, Omu-Aran, Kwara State, Nigeria
[e] Department of Computer Science & Engineering, Indian Institute of Technology Delhi, New Delhi, India

ARTICLE INFO

Keywords:
Social media analysis
Tweet analysis
Event detection
Machine learning
Big data analytics
Sentiments analysis

ABSTRACT

Most of the previous studies on the semantic analysis of social media feeds have not considered the issue of ambiguity that is associated with slangs, abbreviations, and acronyms that are embedded in social media posts. These noisy terms have implicit meanings and form part of the rich semantic context that must be analysed to gain complete insights from social media feeds. This paper proposes an improved framework for pre-processing of social media feeds for better performance. To do this, the use of an integrated knowledge base (*ikb*) which comprises a local knowledge source (Naijalingo), urban dictionary and internet slang was combined with the adapted Lesk algorithm to facilitate semantic analysis of social media feeds. Experimental results showed that the proposed approach performed better than existing methods when it was tested on three machine learning models, which are support vector machines, multilayer perceptron, and convolutional neural networks. The framework had an accuracy of 94.07% on a standardized dataset, and 99.78% on localised dataset when used to extract sentiments from tweets. The improved performance on the localised dataset reveals the advantage of integrating the use of local knowledge sources into the process of analysing social media feeds particularly in interpreting slangs/acronyms/abbreviations that have contextually rooted meanings.

1. Introduction

Social media has been identified as an important open channel of sourcing information (Kumar, 2016). People tend to communicate information more regularly and faster via social media due to its real-time and lightweight nature than through short message service (SMS) from the cell telephone (Hughes, Peterson & Palen, 2015; Olteanu, Vieweg, & Castillo, 2015).

Unlike painstakingly created news and different literary contents of the web, social media posts present numerous new difficulties for analytics algorithms due to their noisy, full-size scale, and social nature. Most of the contents on social media streams contain casual usage of terms, incomplete statements, noisy data, slangs, statements with spelling and grammatical mistakes, irregular sentences, abbreviated phrases/words, and improper sentence structure (Asghar, Kundi, Ahmad, Khan, & Khan, 2017; Katragadda, Benton & Raghavan, 2017; Panagiotou, Katakis, & Gunopulos, 2016). These challenges make it vital to seek improvement of the performance of existing pre-processing solutions for social media streams (Bani-Hani, Majdalaweih & Obeidat, 2017;

* Corresponding author at: Department of Computer and Information Sciences, Covenant University, Ota, Nigeria.
E-mail addresses: taiwo.kolajo@stu.cu.edu.ng, ayo.adebiyi@covenantuniversity.edu.ng, taiwo.kolajo@fulokoja.edu.ng (T. Kolajo), daramolaj@cput.ac.za (O. Daramola), ayo.adebiyi@lmu.edu.ng (A. Adebiyi), aseth@cse.iitd.ac.in (A. Seth).

https://doi.org/10.1016/j.ipm.2020.102348
Received 5 January 2020; Received in revised form 25 May 2020; Accepted 22 June 2020

Haruna et al., 2019; Nasar, Jaffry, & Malik, 2019).

The nature of social media streams makes it difficult to understand them as stand-alone statements without reference to information from external sources (Nagarajan & Gandhi, 2018). Grammatical structures which are easily understood by humans cannot be easily recognized by machines and more particularly in the case of social media feeds (Pasolini, 2015). Although efforts were made to address this issue, current knowledge-based strategies for analysing social media feeds have limitations. Most methods can only cope with shallower keyword or topic extraction, the problem with this is that such markers or keywords of interest may not exist in some informational tweets (Saleem, Al-Zamal & Ruths, 2015). Moreover, the majority of knowledge-based approaches for analysing slang/abbreviation/acronym do not take care of ambiguity issues that evolve in the utilisation of these noisy phrases (Sabbir, Jimeno-Yepes & Kavuluru, 2017). Just like ambiguity exist with the use of ordinary language, there is also ambiguity in the utilisation of slangs/abbreviations/acronyms that needs to be rightly interpreted to enhance the results of social media analysis. For instance, the word “pale” is ambiguous, when used on social media. In the Urban Dictionary,[1] it means “having a pale skin”, “to be a platonic soulmate”, and “a really bad hangover” whereas within the Nigerian social media circles, the word "pale" is a pidgin word (pronounced as "per le") that is synonymous in meaning to biological father. Thus, the ambiguity problem does not only apply to regular Standard English words but also the content of social media feeds. Most of the existing approaches have not addressed this problem of ambiguity in social media feeds. As a result of this, knowledge-based approaches must have the capability to semantically analyse the noisy and syntactically irregular language of social media (Bohra, Vijay, Singh, Akhtar & Shrivastava, 2018; Kuflik et al., 2017).

This paper extends a previous paper (Kolajo, Daramola & Adebiyi, 2019) where an abridged report was presented. This study is based on the use of more elaborate data to derive a robust conclusion on the importance of resolving ambiguity in the usage of slangs/acronyms/abbreviations in social media feeds. Specifically, the following extensions were undertaken relative to (Kolajo et al., 2019):

1) An open and standardised Twitter dataset, and a localised dataset of tweets of Nigerian origin (*Naija-tweets*) were used for the evaluation of the proposed framework instead of just the localised dataset that was used in Kolajo et al. (2019).
2) A bigger and improved version of localised dataset (*Naija-tweets*) was used. The new *Naija-tweets* dataset contains additional 2920 tweets, and is a balanced dataset containing an even distribution of tweets with positive and negative sentiments compared to the one in Kolajo et al. (2019) that is an imbalanced dataset. The improvement of the size and content of the localised dataset is to enable a more accurate basis for the evaluation of the proposed framework.
3) A proposed framework has been enhanced with the addition of an automatic spell-checking module to improve the quality of output of the framework.
4) Further elucidation of the adapted Lesk algorithm that was used for disambiguation of social media slangs/abbreviations/acronyms is presented, with an illustration example to foster a better understanding of the disambiguation process.

The adopted approach incorporates the use of local knowledge-base to analyse the contextual meanings that are hidden in the slangs, abbreviations, and acronyms that are contained in social media streams. This is because a better understanding and interpretation of the rich semantics and contextual meanings that are hidden in social media feeds will offer critical information for algorithms or techniques that builds on social media data for useful operations like event detection or sentiments analysis (Zhang, Li, Lv & Chen, 2015). The objectives of this paper are to:

1) Investigate the effect of using integrated knowledge sources for the pre-processing of social media feeds and its performance on computational algorithms that rely on such for their results.
2) To determine whether resolving ambiguity in the usage of slangs/acronyms/abbreviations will lead to improved accuracy of social media analytics.
3) To investigate the impact of local knowledge source in analysing a relevant localised dataset.

As a contribution, this paper proposes an improved approach to pre-processing of social media streams by (1) capturing the full meaning, and literal essence of slangs, abbreviations and acronyms that are used in social media feeds and (2) introducing the use of localised knowledge sources to decipher contextual knowledge in a way that ensures better and more accurate interpretation of social media streams.

The rest of this paper is structured as follows. Section 2 discusses the background and related work, Section 3 presents the methodology, while the evaluation experiment is presented in Section 4. The results of the study are presented and discussed in Section 5, while Section 6 concludes the paper with an overview of further work.

2. Background and related work

Two of the major applications of social media data are sentiments analysis/opinion mining and event detection (Aggarwal, 2018; Al-garadi et al., 2018). Opinion mining is used to extract the opinion of people on issues and topics of interest (Bohra et al., 2018), while event detection is used to report what is happening (Symeonidis, Effrosynidis & Arampatzis, 2018). These two major application areas of social media data have several other subcategories such as product/service improvement, brand awareness, advertising/marketing strategies, identification of trends, network structure analysis, news propagation, disaster and emergency

[1] https://www.urbandictionary.com/

management, political campaigns, security, press clipping, entity extraction, and more. This section presents an overview of sentiment analysis and event detection. It also includes a review of related work.

2.1. Sentiment analysis

Performing sentiments analysis in social media is quite challenging for natural language processing because of the intricacy and variability within the dialect articulation coupled with the availability of real-time content (Taimoor et al., 2016). The first applications of sentiments analysis were on longer texts such as emails, letters, and longer web contents. Applying sentiment analysis to the microblogging fraternity is a challenging job due to the inherent noisy characteristics of social media feeds such as frequent usage of slangs, abbreviation, acronyms, incomplete statements, spelling and grammatical errors, emoticons, and hashtags amongst others (Zhan & Dahal, 2017).

Sentiment analysis has gained much attention because of the explosion of internet applications such as forums, microblogging websites, and social networks. The approaches to sentiment analysis fall into 3 categories (Vyas & Uma, 2018), which are knowledge-based/lexical approaches (Silva & Fernando, 2016), statistical/machine learning approaches (Elouardighi, Maghfour, Hammia & Aazi, 2017), and the combination of both (hybrid) approaches (Alfrjani, Osman, & Cosma, 2019; Giatsoglou et al., 2017; Raut, Kukkarni & Barve, 2017). In this paper, a combination of machine learning and knowledge-based approaches (viz. hybrid approach) to remedy ambiguity in slangs, abbreviations, acronyms and to extract sentiment from tweets was employed. Also, the use of local knowledge source has been introduced to existing knowledge sources to specifically analyse and interpret the localised slangs of Nigerian origin.

2.2. Event detection

Recently, event detection has received significant attention because of the prevalence of social media and its wide application in decision making and crisis management (Chen, Zhou, Sellis, & Li, 2018). Event detection is aimed at organising a social media stream into sets of documents in such a way that each set is coherently discussing the same event. Understanding and analysing social media posts is challenging due to the presence of unrelated, noisy, and different format data (Sreenivaslu & Sridevi, 2018). Event types range from a planned event with a predefined location and time, for instance, a concert; sudden or unplanned event (e.g., terror attack or earthquake); breaking news discussed on social media platforms; a local event that is confined to a specific geographical area (e.g. automobile accident); and entity related event (e.g. a popular singer with a new video clip) (Gei, Cui, Chang, Sui & Zhou, 2016; Laylavi, Rajabifard & Kalantari, 2017; Panagiotou, Katakis, & Gunopulos, 2016).

Event detection can also be categorised into specified and unspecified events. A specified event depends on known specific characteristics and information about the event which may include time, venue, type, and description. This information can be gathered from an event content material or through social media users. An unspecified event is based totally on the temporal alerts or signals of the social media streams in detecting real-world event occurrence (Atefeh & Khreich, 2015).

2.3. Related work

This section discusses some of the previous research efforts that are associated with pre-processing/ambiguity handling in sentiment analysis and event detection. These are discussed in the following subsections.

2.3.1. Ambiguity handling in social media streams for sentiment analysis

The impact and effect of pre-processing which involves activities such as tokenization, removal of stop-words, lemmatization, slang analysis, and redundancy elimination on the accuracy of the result of techniques that rely on the analysis of social media data for sentiment analysis have been studied by many researchers. There is a general belief that once social media stream contents are well interpreted and represented, it leads to improvement of sentiment analysis results.

(Nigam & Yadav, 2018) presented the position of text pre-processing in sentiment analysis using a lexicon-based approach. The pre-processing stages entail stop-words removal, removal of HTML tags, URL, mentions, emoticons. A manually created dictionary that contains positive and negative words was used for abbreviation expansion. Emoticon and slang terms were not handled. It should be pointed out that analysing emoticons and slangs found in social media can greatly influence the result of sentiment analysis. Also, representing abbreviations based on words found in the dictionary without considering the context in which such abbreviations occur may not take care of ambiguity issues completely.

The impact of pre-processing methods on sentiment classification accuracy was explored by Singh and Kumari (2016) with the usage of Twitter data. The pre-processing method targeted removal of URLs, mentions, hashtag, retweets. An n-gram was used to obtain binding of slang words to other co-existing words and conditional random fields to find significance and sentiment translation of such slang words. The result of the study showed an improvement in sentiment classification accuracy.

In the same vein, (Bijari, Zare, Kebriaei & Veisi, 2020; Romero & Becker, 2018) investigated the position of text pre-processing and found that a suitable combination of different pre-processing tasks can improve the accuracy of social media streams classification.

(Ansari, Seenivasan & Anandan, 2017) investigated sentiments analysis on a Twitter dataset. The pre-processing method alongside tokenization, stemming and lemmatization that was used consists of the replacement of URLs, user mention, negation, exclamation, and question marks with tokens. Letter repetition with one or two occurrences of a letter was replaced with two. The

result shows that the pre-processing of Twitter data improves the performance of sentiment analysis techniques.

The pre-processing method adopted by Ouyang, Guo, Zhang, Yu and Zhou (2017); Ramadhan, Novianty and Setianingsih (2017) includes deletion of stop-words, punctuation, URLs, mentions, 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 & Xiaolin, 2017) studied the effect of pre-processing and discovered that expanding acronyms and changing negation enhanced classification while the elimination of stop-words, numbers or URLs did not result in enormous improvement. On the contrary, (Symeonidis et al., 2018) evaluated classification accuracy based on pre-processing techniques and proved that getting rid of numbers, performing lemmatization, and replacing negation improve accuracy. From the assessment of the literature, most of the previous studies have not focused specifically on how to handle noisy terms such as slangs, abbreviations, and acronyms as well as resolving ambiguity when these noisy terms are used in social media feeds.

In addition to traditional text pre-processing, some authors also used external sources such as internet slang, dictionary, and UT-Dallas corpus and rule-based approach for improving text processing stages. (Sharma, Srinivas & Balabantaray, 2015) conducted research to normalise text presented in two languages, specifically in Roman and Indian and then tried to predict sentiment from the combination of the two languages. Nevertheless, the ambiguity issue that erupted from the use of these two languages were not handled.

An approach for acronym discovery, enrichment, and disambiguation in short texts was proposed by Barua and Patel (2016). Expansion of acronym was done using two measures, which are popularity expansion in news articles, and the use of contextual information surrounding the acronym with metadata information in the acronym dictionary data. The result shows that combining popularity expansion with context performs better than when used individually. However, the authors did not handle slang terms.

(Asghar, Kundi, Ahmad, Khan, & Khan, 2017) proposed a rule induction framework for Twitter sentiment evaluation. In the preprocessing stage, slangs or abbreviations were filtered and taken as misspelt words. Aspell library in Python was used for treating these noisy terms and whenever the spelling checker was unable to figure out the correct meaning or word to replace a slang or an abbreviation, such a noisy term was discarded. Aspell is not sufficient to take care of these noisy terms and resolve ambiguity issues that are associated with the usage of slangs and abbreviations.

Slang sentiment dictionary was proposed by Wu, Morstatter and Liu (2018). Slang words were extracted from the urban dictionary. Urban dictionary and related words were exploited to estimate sentiment polarity. This is tailored specifically to sentiment analysis. Our proposed preprocessing framework is generic in the sense that it can be used not only for sentiment analysis but also for event detection and any related application domain that rely on the analysis of social media feeds. Moreover, using only the urban dictionary for slang analysis may not properly take care of ambiguity in the usage of slangs.

Sentiment classification of tweets with the use of hybrid techniques (i.e. combination of Particle Swarm Optimisation (PSO), Decision Tree (DT), and Genetic Algorithm (GA)) was proposed by Pasolini (2015). Although the author claimed that they did abbreviation expansion and correction of misspelt words, how this was done was not clearly explained in the paper. Also (Gandhe, Varde & Du, 2018) proposed a hybrid technique to perform sentiment analysis on Twitter data that can be used for future recommendation. Naïve Bayes classifier was trained on labelled data to find the probability of the unlabelled data based on similar predefined data. Nevertheless, the authors did not mention how the noisy data found in the tweets was handled.

The work of (Ray & Chakrabatti, 2017) focused on acronym removal by mapping an acronym with meaning from a lexicon. However, direct matching with available meaning from a lexicon without considering the context cannot adequately address ambiguity issues.

There are approaches where commonsense knowledge (CSK) have been used to obtain implicit knowledge of specific entities for natural language processing. In Tandon, Varde and Melo (2017), common knowledge was acquired from open web knowledge sources such as DBpedia, ConceptNet5 and WordNet to process natural language texts. The application of the framework reported in Tandon et al. (2017) is found in Puri, Varde, Du and Melo (2018) where concepts, properties, and relationship in the CSK known as WebChild were deployed. CSK was used to filter tweets and perform ordinance tweet mapping based on Smart City Characteristics (SCC) for sentiments analysis. To assess public opinions on local laws governing a particular urban region, SentiWordNet was used for polarity-based classification. Due to redundant information in these open web knowledge sources, inference from such sources can be extremely slow (Suliman et al., 2016). Also, this framework did not handle the noisy terms commonly found in tweets.

Cross-lingual propagation for deep sentiment analysis was presented by Dong and Melo (2018) to yield sentiment embedding vectors for numerous languages. While this is an appreciable effort in mitigating the scarcity of data in multiple languages, the paper did not focused on enriching the pre-processing stage of social media streams for a faster, understandable and interpretable social media data. Besides, the authors did not elaborate on how the noisy data commonly associated with social media data was handled.

2.3.2. Ambiguity handling in social media streams for event detection

Many of the previous efforts that focused on event detection have attended to aspects of natural language pre-processing such as tokenization, stop-word removal and stemming but not much attention has been given to noisy and abnormal phrases consisting of slangs, abbreviations, and acronyms that are rampant in social media streams. (Zeppelzauer & Schopfhauser, 2016) presented a multimodal classification of events in social media. The pre-processing stage focused on the removal of stop words, special characters, numbers, emoticons, HTML tags, and words with less than 4 characters. (Zhang, He, Gao & Ni, 2018) pre-processed tweets using tokenization and stop-word removal.

Event detection algorithm based on social media was proposed by Cui et al. (2017). The work focused on detecting foodborne disease event from Weibo (a social media network like Twitter commonly used in China). The authors adopted keyword extraction for filtering Weibo data that contained foodborne disease-related words. Support Vector Machine was also trained using ten features

which include the number of followings, number of followers, user tweets, average retweets, average recommendation, average comments, average link in tweets, the length of personal description, the average length of tweets, and time of posting to filter Weibo data. The problem here is that the same keywords can refer to different things based on the context in which they are used. Also, the authors did not handle slangs/acronyms/abbreviations.

Real-time event detection from social media streams was proposed by Hasan, Orgun, & Schwitter, 2019. The pre-processing stage involves removal of tweets with spam using a manually curated list of about 350 spam phrases, removal of mentions, URLs, stop words, and username. This was then followed by tokenization. Tweets were transformed into a vector form using TF-IDF. The authors did not handle slangs/acronyms/abbreviations. They also used TF-IDF for vector generation which only does a pairwise comparison but does not consider the semantics of text and the order of words.

(Mossie & Wang, 2020) worked on the identification of vulnerable communities using hate speech detection on social media. The authors performed social media data pre-processing by removing punctuation, URLs, null values, different symbols, and retweets. Tokenization and part of speech tagging were also performed. The authors did not handle slangs/abbreviations/acronyms.

Real-time health classification from social media data with SENTINEL using deep learning was proposed by Șerban, Thapen, Maginnis, Hankin and Foot (2019). The pre-processing involves removal of email address, links, mentions, punctuations, excess spaces, translation of HTML, emoji, emoticons using emoji dictionary, removing quotes (“ ”) from quoted words and prefixing with a quote, splitting hashtags into individual words. However, slang/abbreviation/acronym terms were not handled.

(Nazura & Muralidhara, 2015) proposed corpora building automation to facilitate social media analytics. The proposed technique was tested with tweets about Indian politics and was able to achieve 85.55% in identifying the correct domain terms and entities. However, the algorithm could not recognise slangs from regional languages which underscores the significance of integrating local knowledge sources.

Social media textual content normalisation using a hybrid, and modular pipelines to normalise slangs referred to as out-of-vocabulary (OOV) to in-vocabulary (IV) was proposed by Sarker (2017). The author focused on one-to-one normalisation and one-to-many normalisation. However, the author did not focus on many-to-one normalisation. An example of many-to-one normalisation is “S.T.O.P” to “stop”. This puts a limitation on the overall performance of the proposed system on the intrinsic evaluation data, besides the fact that ambiguous OOV were not addressed. (Zhang, Luo & Ma, 2017) proposed a framework for learning a continuous representation for acronym referred to as Arc2vec using three embedding models: Average Pooling Definition Embedding (APDE), Max Pooling Definition Embedding (MPDE) and Paragraph-Like Acronym Embedding (PLAE) to resolve the ambiguity problem. The result shows that both APDE and MPDE perform better than PLAE. However, contextual information was not taken into consideration for acronym embedding.

(Moseley, Alm, & Rege, 2015) used abbreviation and lexicon to deduce demographic attributes such as gender, age, and education level. Abbreviation (or OOV) was handled with the use of a regex and recursive logic. However, their framework was unable to identify a large number of abbreviation due to the generation of contextually incorrect unrelated replacement.

(Wankhede, Patil, Sonawane & Save, 2018) proposed a pre-processing framework that utilises N-gram and Hidden Markov Model for spell-checking and tweet correction. The authors tested on only 10 tweets and got an accuracy of 80%. The result cannot be generalised due to the very small number of tweets used, and the approach was not compared with different techniques. Analysis of Twitter hashtag for users’ popularity detection was proposed by Zadeh, Abbasov and Shahbazova (2015) using Fuzzy clustering approach. While user-defined metalinguistic labels such as hashtags, comments, free-text metadata, social tagging, and geo-tagging can be useful in social media analysis, some of these labels may be ambiguous or shift in meaning over time. For instance, popular hashtags have a high tendency of being spammed. Table 1 provides an outline of the previous research efforts (that focused on handling slang or abbreviation or acronym) and the various aspects that have so far received the attention of researchers in the literature.

To the best of our knowledge, none of the previous approaches has treated all of slang, abbreviation, and acronym issues or introduced the use of localised knowledge sources for resolving ambiguity in the usage of slangs/abbreviations/acronyms originating from a specific location. Moreover, out of 15 papers (as shown in Table 1) that focused on handling slangs or abbreviations or acronyms, only 3 (20%) addressed the issue of ambiguity which is one of the main concerns in natural language processing. This limits the accuracy of the interpretation of social media feeds. For instance, a research performed by Asghar, Khan, Khan and Kundi (2016), used a hybrid approach with a slang classifier, where the slangs were categorised into negative or positive sentiments without contemplating the context in which the slangs have been used. The method had an overall accuracy of 90%.

Therefore, in contrast to previous efforts, this paper attempts to improve on existing approaches by finding a representation for handling noisy terms and resolving the ambiguity based on the contextual analysis of terms and their use. The proposed framework is designed to analyse the rich semantics embedded in social media streams for better understanding and interpretation to enhance the accuracy of computational techniques that rely on such analysis.

3. Methodology

This section describes the approach used to realise the Social Media Feed Pre-processing (SMFP) framework. This includes data collection, data pre-processing, and data enrichment. The architectural framework of the proposed SMFP is depicted in Fig. 1.

3.1. Description of the SMFP framework

The proposed SMFP architectural framework consists of 3 layers, which are briefly described in the following sections.

Table 1
Comparison of existing social media pre-processing research work.

S/N	Sources Slang	Abbreviation	Acronym	Disambiguation	Spell-checking	Handles Localised knowledge source	Includes	Research Focus
1	(Nigam & Yadav, 2018)		√					Sentiment analysis
2	(Nazura & Muralidhara, 2015)	√						Sentiment analysis/event detection
3	(Moseley, Alm, & Rege, 2015)		√					Sentiment analysis/event detection
4	(Sharma et al., 2015)	√	√					Sentiment analysis
5	(Barua & Patel, 2016)			√	√			Sentiment analysis/event detection
6	(Zhang et al., 2017)			√	√			Sentiment analysis/event detection
7	(Asghar, Kundi, Ahmad, Khan, & Khan, 2017)	√				√		Sentiment analysis
8	(Ramadhan et al., 2017)	√						Sentiment analysis
9	(Sarker, 2017)	√						Sentiment analysis/event detection
10	(Ray & Chakrabatti, 2017)			√				Sentiment analysis
11	(Khan, Zubair, Hussain, Mazud, & Sadia, 2017)	√						Sentiment analysis
12	(Gupta et al., 2019)	√			√	√		Sentiment analysis/event detection
13	(Nagarajan & Gandhi, 2018)					√		Sentiment analysis
14	(Wankhede et al., 2018)					√		Sentiment analysis
15	(Wu et al., 2018)	√						Sentiment analysis

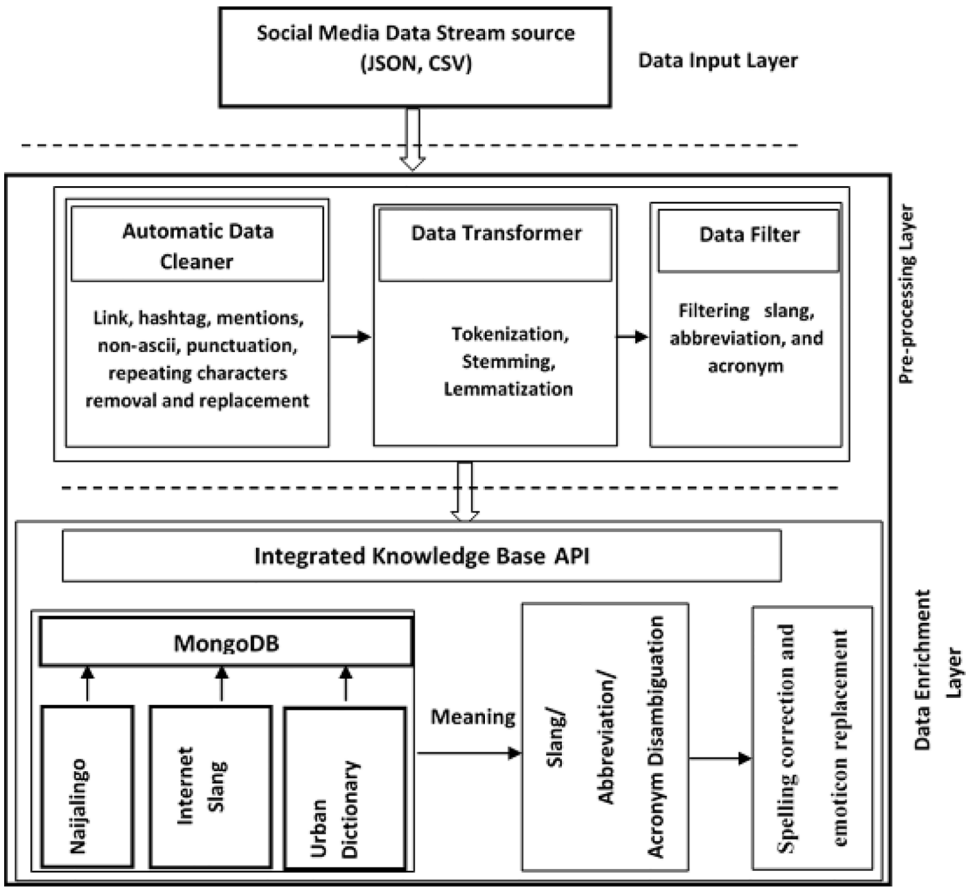


Fig. 1. Proposed Social Media Feeds Pre-processing (SMFP) Framework.

3.1.1. Data layer

This layer enables the input of social media data that are scheduled for pre-processing by the SMFP framework. This is suitable for both data at rest and data in motion. Data at rest can be stored as Comma Separated Values (CSV) or JavaScript Object Notation (JSON). For data in motion, a user interface that was built around an underlying API provided by Twitter is used to collect tweets in JSON format. With JSON format, each line can be parsed as an object.

3.1.2. Data pre-processing layer

This layer includes three subcomponents which are data cleaner, data transformer, and data filtering. The data cleaner performs hashtag, link, non-ASCII mentions, repeating characters, punctuation removal and replacement. The data transformer is responsible for token extraction, stemming, and lemmatization. The data filter takes care of slang, abbreviation, and acronym extraction for onward transfer to the data enrichment layer for further processing.

3.1.3. Data enrichment layer

This layer houses the Integrated Knowledge Base (*ikb*) which comprises urban dictionary, Naijalingo, and internet slang. The *ikb* API allows integration of other suitable local knowledge sources for addressing slangs, abbreviations, and acronyms. The *ikb* API also caters for slang/abbreviation/acronym disambiguation, spelling correction, and emoticon replacement. The structure of a typical lexicon of the *ikb* is depicted in Fig. 2.

3.1.4. Process workflow of the smfp framework

The operational workflow of the SMFP framework as depicted in Fig. 3 is presented in this section. From the collected data stream, URLs, Tags, non-ASCII characters, and mentions are automatically removed using a regular expression. Then, the subsequent stage of the preparation will be to perform tokenization and normalization followed by slangs, abbreviations, acronyms, and emoticons filtering.

This is done with the use of corpora of English words in the NLTK library. The filtered slangs/abbreviations/acronyms are then handled by the *ikb*. Due to several meanings that may be attached to each of the slang/abbreviation/acronym terms, there is a need to disambiguate the ambiguous terms and select the best sense from the several meanings provided. This is done by adapting the Lesk

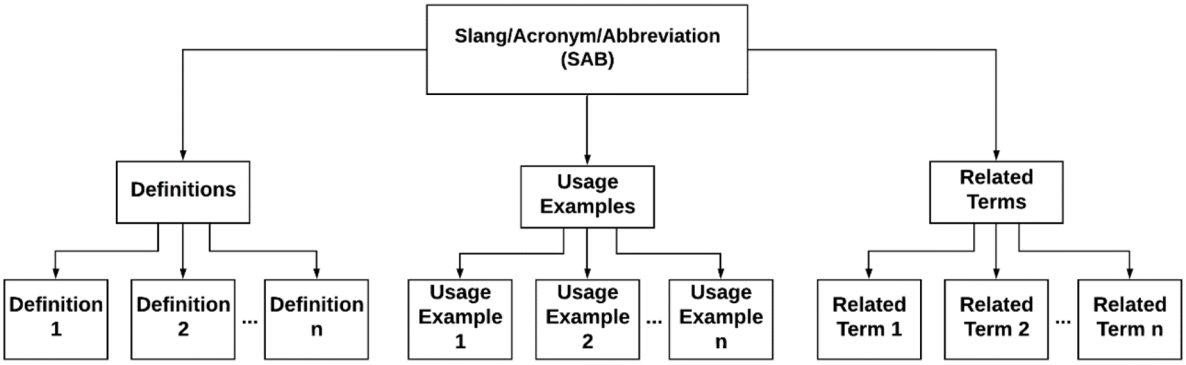


Fig. 2. The structure of a typical lexicon of the integrated knowledge base.

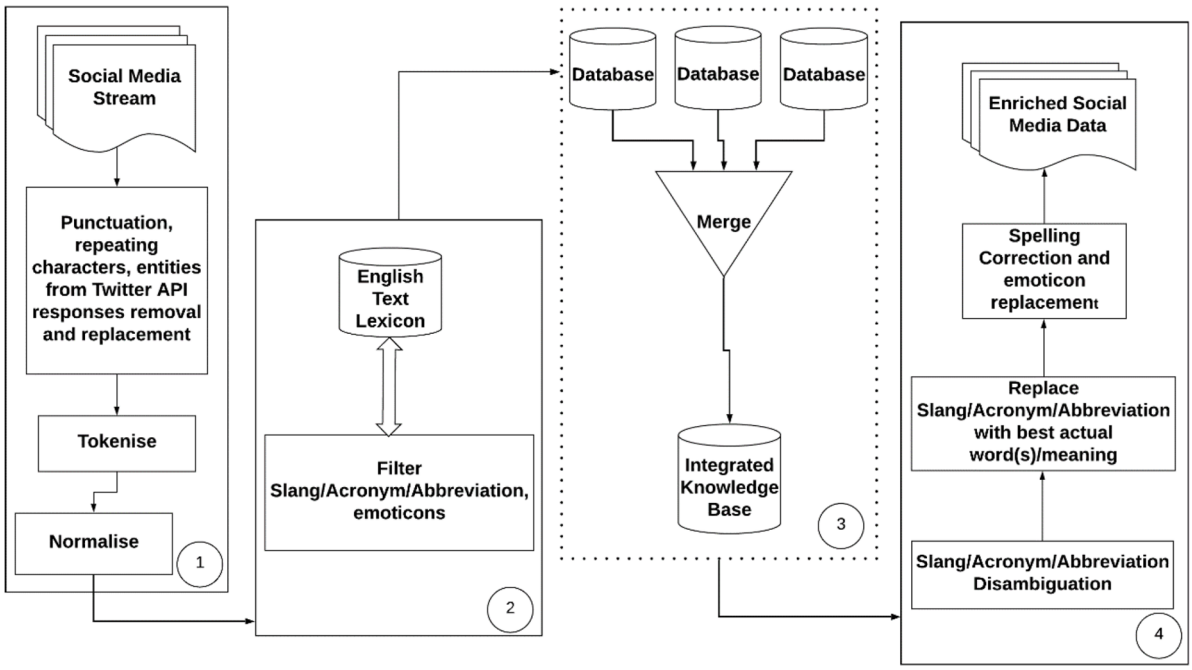


Fig. 3. The workflow of the social media feed pre-processing (SMFP) framework.

algorithm to *ikb* to extract the best sense of ambiguous slangs/abbreviations/acronyms based on their context. The description of the slang/abbreviation/acronym disambiguation process is presented in Section 3.1.5. The data enrichment stage is then concluded with spellings correction and emoticon replacement. The automatic spell-checking process involves correcting typographical errors on the output from the disambiguation stage that might have escaped the *ikb* enrichment stage using the automatic spell checker library of Python. Also, the emoticon texts in the tweets are replaced with their meanings from the *ikb*.

3.1.5. Resolving ambiguity in slang/acronym/abbreviation

Since it is viable for each slang/acronym/abbreviation term to have more than one meaning or definition, this results in ambiguity. To resolve ambiguity in the definitions, the Lesk algorithm (Abed, Tiun & Omar, 2016) was adapted. Lesk algorithm is a famous set of rules for word sense disambiguation. It performs disambiguation by doing a comparison between the target document and the glosses of every sense as described in WordNet. The sense with the highest cosine similarity is then treated as the best sense of the word in the target document. WordNet is a lexical database of only regular English words. WordNet does not contain slangs/acronyms/abbreviations and that necessitates the use of *ikb*. In this paper, the Lesk algorithm (see Fig. 4) was adapted to determine the best sense of each slang/abbreviation/acronym observed in tweets by performing a comparison between the target slang/abbreviation/acronym with the glosses of every sense as described in *ikb*. This is illustrated with the example below.

Example: Consider the case of disambiguating “pale” in the tweet: **Sam, your pale has come back from work**

a. Given the following ikb senses

```
Input: tweet text, sabt
Output: enriched tweet text
//Procedure to disambiguate ambiguous slang/acronyms/abbreviation in tweets by
//adapting the Lesk algorithm over the usage examples of
//slangs/acronym/abbreviation found in the integrated knowledge base (ikb)
Notations:
sjk: the current tweet being processed
slngs: slangs; acrs: acronyms; abbrs: abbreviation
sabt: a collection of slang/acronym/abbreviation terms
wi: an individual slang/abbreviation/acronym in sabt
sti: ith usage example of wi in sabt found in ikb

procedure disambiguate_all_slngs/acrs/abbrs
    for each wi in sabt do
        best_sense=disambiguate_slng/acr/abbr (wi, sjk)
        display best_sense
    end for
end procedure

function disambiguate_slng/acr/abbr(wi, sjk)
    usage_senses = extract_usage_examples(wi, ikb)
    //usage_senses: a collection of usage examples of wi in sabt found in the ikb
    //usage_senses → {st 1, st2, …stn | m ≧ 1}
    int[] score //  an array of semantic relatedness scores

    for each sti ∈ usage_senses of wi  do
        for i= 1 to n do
            // n is the total number of usage examples for wi
            score [i] = relatedness(sti,sjk)
        end for
        best_score = max(score[i])
    end for
    sti ← best_score
    return sti
    map sti with defi  //(where defi ∈ definition)
    replace wi in sabt in the tweet with defi
end function
```

Fig. 4. Adapted lesk pseudocode.

pale[1]	Definition: The male parent that gave birth to you Usage example: Ugonna your father don come from work o Related term: old man, papa, patriarch, daddy, pa
pale[2]	Definition: Having a white skin Usage example: Having a pale skin isn't bad neither is having a tan skin Related term: white, tan, beauty, skin, change, dull, boring
Pale[3]	Definition: A really bad hangover Usage example: I feel so pale with what happened last night Related term: whitie, pale, white, white boy
Pale[4]	Definition: A platonic soulmate Usage example: I am really pale for Gamzee. Help Related term: homestuck, quadrants

b. Choose the sense with the most word overlap between usage examples (usage_senses) and context (that is, tweet in question, sjk)

The tweet in question, *sjk*, "Sam, your pale don come back from work" is compared with all the available usage examples (*usage_senses*) in the *ikb* relating to "pale". The score for each comparison (relatedness(st_i,sjk)) is stored in an array (*score)*. The comparison with the highest score is taken as the best score. sam, your [pale] has come back from work

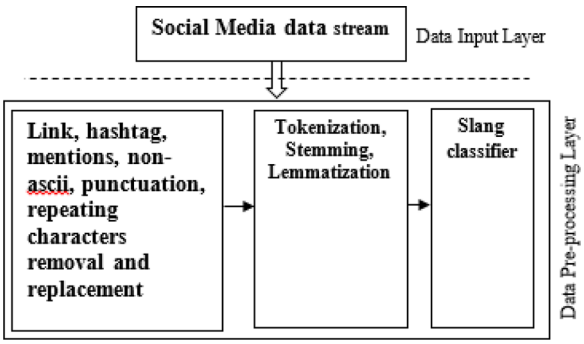


Fig. 5. An overview of the general pre-processing method (GSMFPM).

pale[1]	Definition: The male parent that gave birth to you Usage example: Ugonna your father don come from work o Related term: old man, papa, patriarch, daddy, pa	4 overlaps
pale[2]	Definition: Having a white skin Usage example: Having a pale skin isn't bad neither is having a tan skin Related term: white, tan, beauty, skin, change, dull, boring	1 overlap
pale[3]	Definition: A really bad hangover Usage example: I feel so pale with what happened last night Related term: whitie, pale, white, white boy	1 overlap
pale[4]	Definition: A platonic soulmate Usage example: I am really pale for Gamzee. Help Related term: homestuck, quadrants	1 overlap

Usage example 1 with 4 overlaps is chosen as the most appropriate st_i. c. Map the best usage example to the corresponding definition

The st_i with the highest score is then mapped with the corresponding definition. The usage example 1 is mapped with the definition of pale[1], therefore the best sense for "pale" in this context is "the male parent that gave back to you". The pseudocode of the Adapted Lesk Algorithm is presented in Fig. 4.

4. Evaluation experiment

The proposed SMFP framework was evaluated by benchmarking it with the General Social Media Feed Pre-processing Method (GSMFPM). The details of the proposed SMFP framework has been presented in Fig. 1. Fig. 5 presents the general pre-processing framework. It entails data collection, data cleaning, data transformation, and slang classification. The first three modules of the GSMFPM work like the proposed SMFP framework. The slang classifier module uses a slang dictionary to classify slangs into negative and positive sentiments without considering the context in which those slangs are found. The proposed SMFP framework provides improvement on the general pre-processing framework by leveraging *ikb* and the adapted Lesk algorithm to better interpret, analyse, and take care of ambiguity in slangs/abbreviations/acronyms found in tweets. The difference between GSMFPM and SMFP is highlighted in Table 2.

4.1. Dataset description

Two datasets were used for evaluation of the proposed SMFP. The first dataset was extracted from Twitter sentiment analysis training corpus dataset.[2] The dataset originated from both Twitter sentiment corpus by Niek Sanders and the University of Michigan Sentiment Analysis on Kaggle. The dataset has 1578,627 classified tweets with each row being marked as 1 for positive and 0 for negative. When the dataset was downloaded there was a total of 1048,575 out of which 10% was selected by selecting every tenth record to have a total of 104,857 tweets. The selected dataset was further subdivided into training and testing datasets. Every fifth record of the 104,857 was selected as test data (i.e. 20% of the selected dataset, so that 20,971 tweets was used as test data) while the remaining was for training resulting in a total of 83,886 tweets.

The second dataset, which is called *Naija-tweets* in the previous paper Kolajo et al., 2019) was extracted from tweets of Nigerian origin. The dataset focused on politics in Nigeria and contains a total of 10,000 tweets. More information on the dataset is provided in Kolajo et al. (2019). All tweets extracted had been manually classified into positive sentiment ((1) or negative sentiment (0) by three computer scientists who are knowledgeable in sentiment evaluation. The sentiment results from the three experts were placed side by side and where at least two of the three result matches for a single row of a tweet, such result is selected as the sentiment for that

[2] http://thinknook.com/Twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/

Table 2
Difference between GSMFPM and SMFP.

Feature	GSMFPM	SMFP
Abbreviation handling	No	Yes
Acronym handling	No	Yes
Disambiguation	No	Yes
Inclusion of Localised Knowledge Source	No	Yes
Spell-checking module	No	Yes

particular tweet. 80% of the *Naija-tweets* dataset was used as training data while 20% was used as test data.

The amount of tweets extracted from Twitter sentiment analysis training corpus is much more than (approximately 10 times) that of localised dataset (*Naija-tweets*). While the Twitter sentiment analysis training corpus can be considered as a generalised/standardised dataset, the *Naija-tweets* dataset is from a specific location, Nigeria. The Twitter sentiment analysis training corpus dataset is relatively balanced because out of the 104,856 tweets from Twitter sentiment analysis training corpus dataset, 52.7% was classified as positive and 47.3% as negative. The *Naija-tweets* dataset with 10,000 tweets has 64.6% classified as negative sentiments and 35.4% classified as positive sentiments, which makes it somewhat imbalanced. *Naija-tweets* dataset is about politics in Nigeria, the sentiment tilted towards more negative sentiments than positive sentiments. To correct the imbalance in *Naija-tweets* dataset in the previous paper in Kolajo et al. (2019), Random Over Sampling (ROS) technique, which has better classification accuracy than both Random Under Sampling (RUS) and Synthetic Minority Oversampling Technique (SMOTE) was adopted (Leevy, Khoshgoftaar, Bauder & Seliya, 2018). ROS implies randomly adding instances from the minority class, in the case of *Naija-tweets* dataset, instances from positive sentiments (minority class) were randomly selected and added to make up for the imbalance. Specifically, 82.5% of the instances (totalling 2920 tweets) from the minority class was added to make the *Naija-tweets* dataset become balanced in the ratio of 50% positive (6460 tweets) and 50% negative (6460 tweets).

4.2. Feature extraction and representation

Feature extraction has to do with the transformation of input into a set of features. For text classification, the simplest and the most commonly used features are n-grams (Boukkouri, 2018; FurnKranz, 2018; Popovic, 2018, June). In this work, two types of features namely, unigram and bi-gram were extracted from the dataset. For the unigram, a total of 76,522 unique words were extracted from the sentiment analysis training corpus dataset. After the extraction of unigrams, top K words were selected to ward off words that occur a few times so as not to influence the classification result. A total of 50,000 unigrams were used for vector representation. There is a total of 501,026 unique bigrams in the dataset. The unique bigrams were ranked and top K amounting to 150,000 bigrams were used for vector representation.

For the *Naija-tweets* dataset, there is a total of 3296 and 10,187 unique words and bigrams respectively. A total of 3000 unigrams and 8000 bigrams were selected for vector representation.

4.3. Classifiers

For sentiment analysis, supervised or unsupervised techniques can be used to extract sentiments from tweets. With unsupervised sentiment analysis techniques, features of a given text are compared with a polarity based on discriminatory word-lexicon to classify tweets. In a supervised technique, a classifier is trained with labelled examples. In this paper, supervised techniques were used. For text classification, the three classifiers used were support vector machines (SVM), multilayer perceptron (MLP), and convolutional neural networks (CNN).

4.3.1. Support vector machines

Support Vector Machines (SVM) are discriminative classifiers that make use of statistical learning theory (Gholami & Fakhari, 2017; Muscolino & Pagano, 2018) to find distinct optimal separating hyperplane between two classes (Yusof, Lin, & He, 2018). This is achieved by locating the maximum margin between the classes' closest points. Intuitively, an excellent separation may be accomplished by finding the hyperplane that has the biggest distance to the closest training data points of any class (so-referred to as functional margin), since in general, the higher the margin, the lower the classifier generalization error (RapidMiner, 2019). Given a set of training points (x_i, y_i), where x is referred to as the feature vector and y is the class, the objective is to locate the maximum margin hyperplane that separates the points with $y = 1$ and $y = -1$ (Faul, 2019). A discriminating hyperplane will satisfy the following inequalities:

$$w'x_i + w_0 \geq 0 \, if \; t_i = +1; \tag{1}$$

$$w'x_i + w_0 < 0 \, if \; t_i = -1; \tag{2}$$

where $(w', \; w_0)$ and (x_i, t_i) represent the primal variables and training set item respectively. The distance between any point x and a hyperplane is calculated by $| w'x_i + w_0 | \; / \; \| w \|$ while the distance to the origin is $|w| / ||w_0||$. SVM is (1) suitable for high dimensional spaces; (2) effective when the dimensions are greater than the samples; (3) memory efficient (uses support vectors, a set of

training points in the decision function); and (4) versatile, which implies that exclusive kernel features can be used for the decision function (Boussouar, Meziane, & Walton, 2019). The disadvantage of the support vector machines includes (1) it results to over-fitting if the features are much larger than the samples kernel functions, to avoid this, the regularization term must be carefully chosen; (2) probability estimates are calculated using costly five-fold cross-validation since SVM do not provide probability estimates directly.

For this work, the SVM model of the Scikit-learn Machine learning library of Python was used. The regularization parameter (C), also known as the penalty parameter of the error term was set to 0.1. The SVM experiment involved the use of Unigram, Bigram, and Unigram + Bigram.

4.3.2. Multilayer perceptron

A multilayer perceptron is a deep, feed-forward artificial neural network consisting of at least three layers of neurons. The multilayer perceptron is often applied to supervised learning problems. It trains on input-output pairs and learns to model the dependencies between the inputs and outputs. It learns a function $f(.)$: $R^m \rightarrow R^o$, where m represents input dimensions and o represents output dimensions. Given a set of features $X = x_1, x_2, ...x_m$ and a target y, it can learn a non-linear characteristic or function approximation for both regression and classification. The input layer consists of a set of neurons $\{x_i|x_1, x_2, ...x_m\}$ that represent the input features. Each neuron within the hidden layer is responsible for the transformation of the values from the previous layer along with a weighted linear summation $w_1x_1 + w_2x_2 + ...w_mx_m$, this is then followed by a non-linear activation function $g(.)$: $R \rightarrow R$ – just like the hyperbolic tangent function. The output layer receives input from the last hidden layer and transforms them into output values. The mathematical formula is as follows: Given a set of training samples $(x_1, y_1), (x_2, y_2), ...(x_ny_n)$, where $x_i \in R^n$ and $y_i \in \{0, 1\}$, one hidden layer with one hidden neuron MLP learns the function $f(x) = W_2g(W_1^T x + b_1) + b_2$ in which $W_1 \in R^m$ and $W_2, b_1, b_2 \in R$ are model parameters. W_1; W_2 constitute the weights of the input layer and hidden layer respectively; and b_1; b_2 represents the bias added to the hidden layer and the output layer respectively. $g(.)$: $R \rightarrow R$ is the activation function set. This is given as:

$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{3}$$

To carry out binary classification, $f(x)$ is passed through a logistic function $g(z) = 1/(1 + e^{-z})$ (where z is the element of input) to get output values between {0, 1}. A threshold that is set to 0.5, implies that outputs' samples greater than or equal to 0.5 would be assigned to the positive class, and the rest to the negative class. For multiple classifications, $f(x)$ itself is a vector of (n_classes). Instead of passing via the logistic function, it passes via a softmax function (Tray, Cicilio, Brekken & Cotilla-Sanchez, 2017). Softmax is often utilised in neural networks to map un-normalised output to a probability distribution over predicted output classes. When vector elements are greater than 1 or negative and might not sum up to 1, softmax function is used for normalisation in such a way that each element x_l lies within the interval {0,1} and $\sum x_i = 1$. Softmax function (Lo & Marculescu, 2019) is written as:

$$softmax(z)_i = \frac{exp(z_i)}{\sum_{l=1}^{k} exp(z_l)} \tag{4}$$

where z_l represents the ith input element to softmax corresponding to class i, and K represents the number of classes. This results in a vector containing the probabilities that sample x belongs to each class. The output with the highest probability is then taken as the class. MLP is efficient for complex classification by learning non-linear models. It can also learn models at real-time using partial_fit. The disadvantages include (1) MLP with hidden layers have a non-convex loss function where more than one local minimum exists. This implies that different random weight initializations lead to different validation accuracy; (2) MLP requires tuning some of the hyperparameters such as the number of hidden neurons layers, and iterations; and (3) MLP is sensitive to feature scaling.

The Keras deep learning library, running on top of TensorFlow as the back engine was used to implement the multilayer perceptron model trained with back-propagation rule and a 10-fold cross-validation procedure in this paper. The choice of Keras was informed by its ability to abstract away much of the complexity of building a neural network (Rosebrock, 2016). One-hidden layer with 500 hidden units was used. The output gives a $Pr(positive|tweet)$, which implies the probability of the tweet sentiment being positive (i.e. threshold $\geq$ 0.5). This was rounded off at the prediction step to convert to class labels 1 (positive) and 0 (negative). The model was trained using binary cross-entropy loss. The MLP experiment involved the use of Unigram, Bigram, and Unigram + Bigram

4.3.3. Convolutional neural networks

Convolutional Neural Networks (CNN) is a model of deep learning architecture that aims at learning higher-order features present in data through convolutions. In CNN, the input data is transformed through all connected layers into a set of class scores given by the output layer (Gibson & Patterson, 2017). CNN can be used to classify a sentence into predetermined categories by considering n-grams. Given a sequence of words $w_{1:n} = w_1, w_2, ..., w_n$, where each word is associated with the embedding vector of dimension d. One-dimensional convolution with width-k is the result of moving sliding-window of size k over the sentence, and making use of the same convolution filter or kernel to every window in the sequence, i.e. a dot-product between the concatenation of the embedding vectors in a given window and a weight vector u, which is then followed by a non-linear activation function g.

Considering a window of words $w_i, ..., w_{i+k}$ the concatenated vector of the i^{th} window is then:

$$x_i = [w_i, w_{i+1}, w_{i+k}] \in R^{k\ X\ d} \tag{5}$$

The convolution filter is applied to each window, ensuing in scalar values r_i, each of the i^{th} window:

$$r_i = g(x_i.\ u) \in R \tag{6}$$

In practice, one typically applies more filters, u_1, ..., u_l, which can then be represented as a vector multiplied by a matrix U and with an addition of bias term b:

$$r_i = g(x_i.\ U + b) \tag{7}$$

with $r_i \in R^l, x_i \in R^{k\,X\,d}, U \in R^{k\,.\,d\,X\,l}$, *and* $b \in R^l$

The Keras deep learning library, running on top of TensorFlow as the backend was used also to implement the CNN model. GloVe Twitter 27B with 200d was used for both datasets vector representation. GloVe is an unsupervised learning algorithm for obtaining word embedding or vector representation of words (Liu, Niu, Yang, Wang & Zhang, 2016). The idea is based on obtaining the statistics of word-word co-occurrence from a corpus, which results in representations that showcase interesting linear structures of the word vector space (Yang, Lee, Lee, Cho & Koo, 2015).

A vocabulary of top 180,000 and 9000 words from the training dataset was used for Twitter sentiment analysis training corpus dataset and *Naija-tweets* dataset respectively. Each of the 180,000 and 9000 words was represented by a 200- dimensional vector. The Embedding layer, which is the first layer is a matrix with (ν + 1) x d shape, where ν is the vocabulary size, 180,000 or 9000 as the case may be, and d is the dimension of each word vector. A total of 600 filters with kernel size 3 was used for the convolution. The embedding layer was initialized with random weights from N (0, 0.01), with each row representing 200-dimensional word vector for each of the words in the vocabulary. For words in the vocabulary which match the GloVe word vectors provided by the Stanford NLP group, the corresponding row of the embedded matrix was seeded from the GloVe vectors. The model was trained with binary cross-entropy loss, with 8 epochs used in running the model. Four different CNN architectures were used in the experiment, which was as follows: 1-Con-NN with a convolutional layer of 600 filters, 2-Con-NN with convolutional layers of 600 and 300 filters respectively, 3-Con-NN with convolutional layers of 600, 300, and 150 filters respectively, and 4-Con-NN was used with convolutional layers of 600, 300, 150 and 75 filters respectively.

5. Result and discussion

The proposed SMFP framework was evaluated by benchmarking with the General Social Media Feed Pre-processing Method (GSMFPM) with the use of three classifiers mentioned in the previous section to extract sentiments from tweets. It has been established from the previous paper (Kolajo et al., 2019) that the general social media pre-processing technique does not take care of slang/abbreviation/acronym ambiguity issues. The evaluation conducted in this paper with the use of additional standardized dataset was used to complement the result in the previous paper. The result of the sentiment classification for Twitter sentiment analysis training corpus dataset is shown in Tables 3 and 4 while the analysis of an improved version of the *Naija-tweets* dataset is presented in Tables 5 and 6. The results of our previous paper (Kolajo et al., 2019) were based on an imbalanced *Naija-tweets* dataset but by using a balanced dataset in this paper, a better result was obtained.

From the analysis of the result, as shown in Table 3, SVM and MLP were run on the two pre-processing methods using unigram, bigram, and the combination of unigram and bigram. The proposed SMFP framework outperformed the existing generic pre-processing method except for SVM with unigram. It should be noted that using the combination of unigram and bigram performs better than using either unigram or bigram except in the case of SVM where using the unigram with existing pre-processing method performed better. Also, using unigram alone generally performs better than using bigram alone.

From Table 4, 4 different CNN architectures; 1-Con-NN, 2-Con-NN, 3-Con-NN, and 4-Con-NN, each with kernel size 3 was run on the two pre-processing methods. The proposed SMFP framework performed better than the general pre-processing method. This shows that capturing slangs/abbreviations/acronyms found in tweets and resolving ambiguity in their usage lead to improvement of subsequent analysis algorithms. Surprisingly, convolutional neural networks with one layer performed better than the two, three, and four layers for both pre-processing methods.

In Table 5, an experiment from Table 3 was repeated using an improved localised dataset, tagged as *Naija-tweets*, which originated from Nigeria as described in Section 4.1, the result of the experiment revealed that the proposed SMFP framework outperformed the GSMFPM with improvement within the accuracy of the result obtained. This underscores the significance of the use of a localised knowledge base in the pre-processing of social media feeds to fully capture the noisy terms that are contained within the social media feeds originating from a specific location.

The result presented in Table 6 is also a proof of the inclusion of localised knowledge source in the pre-processing of social media feeds from a specific origin to understand and better interpret slangs/abbreviations/acronyms stemming from such social media feeds.

Table 3
Sentiment classification results by SVM and MLP based on Twitter sentiment analysis training corpus.

Method	Algorithm	Accuracy (%) Unigram	Accuracy (%) Bigram	Accuracy (%) Unigram + Bigram
GSMFPM	SVM	**78.84**	74.16	76.58
SMFP		77.61	**74.29**	**78.90**
GSMFPM	MLP	82.60	79.40	83.20
SMFP		**84.80**	**85.30**	**86.20**

Table 4
Sentiment classification results by CNN based on Twitter sentiment analysis training corpus.

Method	Algorithm (Kernel size = 3)	Accuracy (%) 1-Con-NN	Accuracy (%) 2-Con-NN	Accuracy (%) 3-Con-NN	Accuracy (%) 4-Con-NN
GSMFPM	CNN	92.17	87.79	87.77	87.95
SMFP		**94.07**	**89.31**	**89.79**	**91.01**

Table 5
Sentiment classification results by SVM and MLP based on *Naija-tweets* dataset.

Method	Algorithm	Accuracy (%) Unigram	Accuracy (%) Bigram	Accuracy (%) Unigram + Bigram
GSMFPM	SVM	**88.80**	88.80	94.60
SMFP		88.24	**94.12**	**98.04**
GSMFPM	MLP	73.64	**88.41**	97.46
SMFP		**73.80**	88.20	**99.00**

Table 6
Sentiment classification results by CNN based on *Naija-tweets* dataset.

Method	Algorithm (Kernel size = 3)	Accuracy (%) 1-Con-NN	Accuracy (%) 2-Con-NN	Accuracy (%) 3-Con-NN	Accuracy (%) 4-Con-NN
GSMFPM	CNN	97.81	97.59	96.30	95.66
SMFP		**99.78**	**99.34**	**98.47**	**96.28**

6. Conclusion and further work

This paper presents an approach that enables better understanding, interpretation, and handling of ambiguity that is associated with the usage of slangs/acronyms/abbreviations in social media feeds. This is critical for subsequent learning algorithms to perform effectively and efficiently. It is an extension of the concept reported in Kolajo et al. (2019) but strengthened with the use of more elaborate data in the form of a bigger, an open, and standard dataset - Twitter sentiment analysis corpus and improved version of *Naija-tweets* dataset to generate a more robust conclusion.

Also, the use of local knowledge base coupled with other knowledge sources for pre-processing of data gives a better interpretation of noisy terms such as slangs, abbreviations, and acronyms in social media feed. Thus, leading to an improvement in the accuracy of algorithms building on them. This suggests that the context from which social media feeds emanate have impact on the interpretation of terms that are contained in social media feeds, which determines the overall accuracy of operations such as sentiment analysis and event detection. The specific contributions of the paper are the following:

a) *An approach for improved semantic analysis and interpretation of noisy terms in social media streams*. Currently, resolving slang/abbreviation/acronym terms found in social media streams is done by getting a lexicon and plugging in the meaning without considering the context of usage of slang/abbreviation/acronym terms. The SMFP proposed in this paper leverages an integrated knowledge base (*ikb*) coupled with localised knowledge sources to capture the rich but hidden knowledge in slangs/abbreviations/acronyms, to resolve ambiguity in the usage of these noisy terms. This will foster a better understanding and interpretation of these noisy terms, which lead to an improvement in the performance of learning algorithms that rely on such analysis for their results.
b) *Reusable knowledge infrastructure for analysis of social media terms with contextually rooted meanings.* While the focus of this paper is on improving sentiment analysis in social media streams, the enrichment component of the SMFP framework can also be used for detecting events from social media streams. In addition, Integrated Knowledge Base (*ikb*) API that was created can be used as a plug-in tool by other applications for interpreting and disambiguating slang/acronym/abbreviation terms. Furthermore, the *ikb* API in the SMFP framework allows the integration of any other local knowledge base or more knowledge sources that may suit other contexts to capture slangs/abbreviations/acronyms that have locally defined meaning.

To address the limitation of this paper, further work must focus on a better way of social media stream representation or embedding. The use of word vector for representing tweet in text classification has a problem of word order preservation and may not accurately capture the semantic context. As a result, more research effort should be directed to phrase or sentence representation for an improved result. Moreover, while Twitter is very prominent as a research data source, exploring other sources or harmonizing Twitter with other social media sources can lead to a more significant result. The concept of harmonizing multiple social media sources for sentiments analysis and event detection is still an open area of research that is worthy of further exploration since only a few approaches have considered this so far.

Declaration of Competing Interest

The authors hereby declare that there is no conflict of interest.

Acknowledgments

The research was supported by Covenant University Centre for Research, Innovation, and Discovery; Cape Peninsula University of Technology, South Africa; The World Academy of Sciences for Research and Advanced Training Fellowship, FR Number: 3240301383; Federal University Lokoja, Nigeria; Indian Institute of Technology, Delhi.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ipm.2020.102348.

References

Abed, S. A., Tiun, S., & Omar, N. (2016). Word sense disambiguation in evolutionary manner. *Connection Science, 28*(3), 226–241. https://dx.doi.org/10.1080/09540091.2016.1141874.

Aggarwal, C. C. (2018). *Machine learning for textMachine learning for text* (1st ed.). New York, NY, USA: Springer.

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

Alfrjani, R., Osman, T., & Cosma, G. (2019). A hybrid semantic knowledgebase-machine learning approach for opinion mining. *Data & Knowledge Engineering, 121*, 88–108. https://doi.org/10.1016/j.datak.2019.05.002.

Al-garadi, M. A., Mujtaba, G., Khan, M. S., Friday, N. H., Waqas, A., & Murtaza, G. (2018). Applications of big social media data analysis: An overview. *2018 International Conference on Computing* (pp. 1–5). Mathematics and Engineering Technologies (iCoMET). https://doi.org/10.1109/ICOMET.2018.8346351.

Asghar, M. Z., Khan, A., Khan, F., & Kundi, F. M. (2016). RIFT: A rule induction framework for Twitter sentiment analysis. *Arabian Journal for Science and Engineering, 43*(2), 857–877. https://doi.org/10.1007/s13369-017-2770-1.

Asghar, M. Z., Kundi, F. M., Ahmad, S., Khan, A., & Khan, F. (2017). T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme. *Expert System, 35*(1), 1–19. https://doi.org/10.1111/exsy.12233.

Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in Twitter. *Computational Intelligence, 31*(1), 132–164.

Bani-Hani, A., Majdalaweih, M., & Obeidat, F. (2017). The creation of an Arabic emotion ontology based on e-motive. *Procedia Computer Science, 109*, 1053–1059.

Barua, J., & Patel, D. (2016). Discovery, enrichment and disambiguation of acronyms. In S. Madria, & T. Hara (Eds.). *DaWaK 2016, LNCS 9829* (pp. 345–360). . https://doi.org/10.1007/978-3-319-43946-423.

Bijari, K., Zare, H., Kebriaei, E., & Veisi, H. (2020). Leveraging deep graph-based text representation for sentiment polarity applications. *Expert Systems with Applications, 144*, Article 113090. https://doi.org/10.1016/j.eswa.2019.113090.

Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018). A dataset of Hindi-English code-mixed social media text for hate speech detection. *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media* (pp. 36–41). ACL.

Boukkouri, H. E. (2018). *Text classification: The first step towards NLP mastery. data from the trenches.* A Medium Corporationhttps://medium.com/data-from-the-trenches/text-classification-the-first-step-toward-nlp-mastery-f5f95d525d73.

Boussouar, A., Meziane, F., & Walton, A. L. (2019). Plantar fascia ultrasound images characterization and classification using support vector machine. In A. Hassanien, K. Shaalan, & M. Tolba (Eds.). *Proceedings of the international conference on advanced intelligent systems and informatics* (pp. 102–111). Cham: Springer. https://doi.org/10.1007/978-3-030-31129-2.

Chen, X., Zhou, X., Sellis, T., & Li, X. (2018). Social event detection with retweeting behavior correlation. *Expert Systems with Applications, 114*, 516–523. https://doi.org/10.1016/j.eswa.2018.08.022.

Cui, W., Wang, P., Du, Y., Chen, X., Guo, D., & Li, J. (2017). An algorithm for event detection based on social media data. *Neurocomputing, 254*, 53–58.

Dong, X., & De Melo, G. (2018). Cross-lingual propagation for deep sentiment analysis. *32nd Association for Advancement of Artificial Intelligence Conference, AAAI 2018* (pp. 5771–5778).

Elouardighi, A., Maghfour, M., Hammia, H., & Aazi, F. (2017). A machine Learning approach for sentiment analysis in the standard or dialectal Arabic Facebook comments. *2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech)* (pp. 1–8). . https://doi.org/10.1109/CloudTech.2017.8284706 2017.

Faul, A. C. (2019). *A concise introduction to machine learning* (1st ed.). New York, NY, USA: Chapman and Hall/CRC.

FurnKranz, J. (2018). A study using n-gram features for text categorization (Research Report No. OEFAI-TR-98-30). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.49.133&rep=rep1&type=pdf.

Gandhe, K., Varde, A. S., & Du, X. (2018). Sentiment Analysis of Twitter Data with Hybrid Learning for Recommender Applications. *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 57–63). . https://doi.org/10.1109/UEMCON.2018.8796661 2018.

Gei, T., Cui, L., Chang, B., Sui, Z., & Zhou, M. (2016). Event Detection with Burst Information Networks. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3276–3286).

Gholami, R., & Fakhari, N. (2017). Support vector machine: Principle, parameters and applications. *Handbook of Neural Computation,* 515–535. https://doi.org/10.1016/B978-0-12811318-9.00027-2.

Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., & Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications, 69*, 214–224.

Gibson, A., & Patterson, J. (2017). *Deep learning. Deep learning.* Sebastopol, CA, USA: O'Reilly Media.

Gupta, A., Taneja, S. B., Malik, G., Vij, S., Tayal, D. K., & Jain, A. (2019). SLANGZY: A fuzzy logic-based algorithm for English slang meaning Selection. *Progress in Artificial Intelligence, 8*(1), 111–121. https://doi.org/10.1007/s13748-018-0159-3.

Haruna, I., Abughofa, T., Mahfuz, S., Ajerla, D., Zulkernine, F., & Khan, S. (2019). A survey of distributed data stream processing frameworks. *IEEE Access, 7*, 154300–154316. https://doi.org/10.1109/ACCESS.2019.2946884.

Hasan, M., Orgun, M. A., & Schwitter, R. (2019). Real-time event detection from the Twitter data stream using the TwitterNews + Framework. *Information Processing & Management, 56*(3), 1146–1165. https://doi.org/10.1016/j.ipm.2018.03.001.

Hughes, A. L., Peterson, S., & Palen, L. (2015). Social media in emergency management. In J. E. Trainor, & T. Subbio (Eds.). *Issues in disaster science and management: A critical dialogue between scientists and emergency managers* (pp. 349–381). FEMA in Higher Education Program.

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

Katragadda, S., Benton, R., & Raghavan, V. (2017). Framework for real-time event detection using multiple social media sources. *Hawaii International Conference on System Sciences* (pp. 1716–1725).

Khan, A., Zubair, A. M., Hussain, A., Mazud, F., & Sadia, I. (2017). A rule-based sentiment classification framework for health reviews on mobile social media. *Journal of Medical Imaging and Health Informatics, 7*(6), 1445–1453. https://doi.org/10.1166/jmihi.2017.2208.

Kolajo, T., Daramola, O., & Adebiyi, A. (2019). Sentiment analysis on Naija-tweets. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* (pp. 338–343). . https://doi.org/10.18653/v1/P19-2047.

Kuflik, T., Minkov, E., Nocera, S., Grant-Muller, S., Gal-Tzur, A., & Shoor, I. (2017). Automating a framework to extract and analyse transport-related social media content: The potential and challenges. *Transport Research Part C: Emerging Technologies, 77*, 275–291. https://doi.org/10.1016/j.trc.2017.02.003.

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

Laylavi, F., Rajabifard, A., & Kalantari, M. (2017). Event relatedness assessment of Twitter messages for emergency response. *Information Processing & Management, 53*(1), 266–280.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data, 5*, 42. https://doi.org/10.1186/s40537-018-0151-6.

Liu, K., Niu, Y., Yang, J., Wang, J., & Zhang, D. (2016). Product related information sentiment-content analysis based on convolutional neural networks for the Chinese micro-blog. *2016 international conference on network and information systems for computers (ICNISC)* (pp. 357–361). . https://doi.org/10.1109/ICNISC.2016.083.

Lo, C., & Marculescu, R. (2019). MetaNN: Accurate classification of host phenotypes from metagenomic data using neural networks. *BMC Bioinformatics, 20*, 314. https://doi.org/10.1186/s12859-019-2833-2.

Moseley, N., Alm, C. O., & Rege, M. (2015). On utilizing nonstandard abbreviations and lexicon to infer demographic attributes of Twitter users. In T. Bouabana-Tebibel, & S. H. Rubin (Vol. Eds.), *Formalisms for reuse and systems integration. 346. Formalisms for reuse and systems integration* (pp. 346–). Cham: Springer. https://doi.org/10.1007/978-3-319-16577-6_11.

Mossie, Z., & Wang, J. H. (2020). Vulnerable community identification using hate speech detection on social media. *Information Processing & Management, 57*(3), Article 102087.

Muscolino, A. M., & Pagano, S. (2018). Sentiment analysis, a support vector machine model based on social network data. *International Journal of Research in Engineering & Technology, 7*(7), 154–157. https://doi.org/10.15623/ijret.2018.070720.

Nagarajan, S. M., & Gandhi, U. D. (2018). Classifying streaming of Twitter data based on sentiment analysis using hybridization. *Neural Computing and Applications, 31*(5), 1425–1433. https://doi.org/10.1007/s00521-018-3476-3.

Nasar, Z., Jaffry, S. W., & Malik, M. K. (2019). Textual keyword extraction and summarization: State of the art. *Information Processing & Management, 56*(6), Article 102088. https://doi.org/10.1016/j.ipm.2019.102088.

Nazura, J., & Muralidhara, B. L. (2015). Automating corpora generation with semantic cleaning and tagging of tweets for multi-dimensional social media analytics. *International Journal of Computer Applications, 127*(12), 11–16.

Nigam, N., & Yadav, D. (2018). Lexicon-based approach to sentiment analysis of tweets using R Language. *Advances in Computing and Data Sciences,* 154–164. https://doi.org/10.1007/978-981-13-1810-8_16.

Olteanu, A., Vieweg, S., & Castillo, C. (2015). What to expect when the unexpected happens: Social media communications across crises. *CSCW '15 proceedings of the 18th ACM conference on computer supported cooperative work & social computing* (pp. 994–1009). ACM.

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

Pasolini, R. (2015*). Learning methods and algorithms for semantic text classification across multiple domains* (Doctoral Dissertation). Alma Mater Studiorum Universita di Bologna. doi: 10.6092/unibo/amsdottorato/7058.

Panagiotou, N., Katakis, I., Gunopulos, D., et al. (2016). Detecting events in online social networks: Definitions, trends and challenges. In S. Michaelis, (Ed.). *Solving large scale learning tasks* (pp. 42–84). Switzerland: Springer.

Popovic, M. (2018). Complex word identification using character n-grams. *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 341–348).

Puri, M., Varde, A., Du, X., & de Melo, G. (2018). Smart governance through opinion mining of public reactions on ordinances. *IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI). Volos. 2018. IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI). Volos* (pp. 838–845). . https://doi.org/10.1109/ICTAI.2018.00131.

Ramadhan, W. P., Novianty, A., & Setianingsih, C. (2017). Sentiment analysis using multinomial logistic regression. *Proceedings of the 2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC)* (pp. 46–49). . https://doi.org/10.1109/ICCEREC.2017.8226700.

RapidMiner (2019). Documentation. https://docs.rapidminer.com/.

Raut, M., Kukkarni, M., & Barve, S. (2017). A survey of approaches for sentiment analysis and applications of OMSA beyond product evaluation. *International Journal of Engineering Trends and Technology (IJETT), 46*(7), 396–400.

Ray, P., & Chakrabatti, A. (2017). Twitter sentiment analysis for product review using lexicon method. *International Conference on Data Management, Analytics and Innovation (ICDMA)* (pp. 211–216).

Romero, S., & Becker, K. (2018). A framework for event classification in tweets based on hybrid semantic enrichment. *Expert Systems with Applications, 118*, 522–538. https://doi.org/10.1016/j.eswa.2018.10.028.

Rosebrock, A. (2016). Installing Keras with TensorFlow backend. https://www.pyimagesearch.com/2016/11/14/installing-keras-with-tensorflow-backend/.

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

Saleem, H. M., Al-Zamal, F., & Ruths, D. (2015). Tackling the challenges of situational awareness extraction in Twitter with an adaptive approach. *Procedia Engineering, 107*, 301–311.

Sarker, A. (2017). A customizable pipeline for social media text normalization. *Social Network Analysis Mining, 7*, 45. https://doi.org/10.1007/s13278-017-0464-z.

Șerban, O., Thapen, N., Maginnis, B., Hankin, C., & Foot, V. (2019). Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. *Information Processing & Management, 56*(3), 1166–1184.

Sharma, S., Srinivas, P. Y. K. L., & Balabantaray, R. C. (2015). Text normalization of code mix and sentiment analysis. *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1468–1473).

Silva, T., & Fernando, M. (2016). Knowledge-based approach for concept-level sentiment analysis for online reviews. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 5*(1), 16–23.

Singh, T., & Kumari, M. (2016). Role of text pre-processing in Twitter sentiment analysis. *Procedia Computer Science, 89*, 549–554.

Sreenivaslu, M., & Sridevi, M. (2018). A survey on event detection methods on various social media. In P. Sa, S. Bakshi, I. Hatzilgeroudis, & M. Sahoo (Vol. Eds.), *Recent findings in intelligent computing techniques. advances in intelligent computing. 709. Recent findings in intelligent computing techniques. advances in intelligent computing* (pp. 87–93). Singapore: Springer.

Suliman, A. T., Al Kaabi, K., Wang, D., Al-Rubaie, A., Al Dhanhani, A., Davies, D. R. J., & Clarke, S. S. (2016). Event identification and assertion from social media using auto-extendable knowledge base. *2016 International Joint Conference on Neural Networks (IJCNN)* (pp. 4443–4445). . https://doi.org/10.1109/IJCNN.2016.7727781.

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

Taimoor, K., Mehr, D., Armughan, A., Irum, I., Shehzad, K., & Kamran, H. K. (2016). Sentiment analysis and complex natural language. *Complex Adaptive Systems Modeling, 4*, 2. https://doi.org/10.1186/s40294-016-0016-9.

Tandon, N., Varde, A. S., & Melo, G. (2017). Commonsense knowledge in machine intelligence. *SIGMOD Record, 46*(4), 49–52.

Tray, K., Cicilio, P., Brekken, T., & Cotilla-Sanchez, E. (2017). Dynamic composite load signature detection and classification using supervised learning over disturbance data. *2017 IEEE Energy Conversion Congress and Exposition (ECCE)* (pp. 1560–1566).

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

Wankhede, S., Patil, R., Sonawane, S., & Save, S. (2018). Data preprocessing for efficient sentimental analysis. *2018 Second International Conference on Inventive*

Communication and Computational Technologies (ICICCT) (pp. 723–726).
Wu, L., Morstatter, F., & Liu, H. (2018). SlangSD: Building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Lang Resources & Evaluation, 52*(3), 839–852. https://doi.org/10.1007/s10579-018-9416-0.
Yang, H., Lee, Y., Lee, H., Cho, S. W., & Koo, M. (2015). A study on word vector models for representing Korean semantic information. *Phonetics and Speech Sciences, 7*, 41–47. https://doi.org/10.13064/KSSS.2015.7.4.041.
Yusof, N. F. A., Lin, C., & He, Y. (2018). Sentiment analysis in social media. In R. Alhajj, & J. Rokne (Eds.). *Encyclopedia of social network analysis and mining*New York, NY: Springer. https://doi.org/10.1007/978-1-4939-7131-2_120.
Zadeh, L. A., Abbasov, A. M., & Shahbazova, S. N. (2015). Analysis of Twitter hashtags: Fuzzy clustering approach. *2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing* (pp. 1–6). . https://doi.org/10.1109/NAFIPS-WConSC.2015.7284196.
Zeppelzauer, M., & Schopfhauser, D. (2016). Multimodal classification of events in social media. *Image and Vision Computing, 53*, 45–56.
Zhan, J., & Dahal, B. (2017). Using deep learning for short text understanding. *Journal of Big Data, 4*, 34. https://doi.org/10.1186/s40537-017-0095-2.
Zhang, X., Li, Z., Lv, X., & Chen, X. (2015). Integrating multiple types of features for event identification in social images. *Multimedia Tools and Applications, 75*(6), 3301–3322.
Zhang, Z., He, Q., Gao, J., & Ni, M. (2018). A deep learning approach for detecting traffic accidents from social media data. *Transportation Research Part C, 86*, 580–596.
Zhang, Z., Luo, S., Ma, S., et al. (2017). Acr2Vec: Learning acronym representations in Twitter. In L. Polkowski, (Ed.). *IJCRS 2017, part I, LNAI 10313* (pp. 280–288). .