Webology (ISSN: 1735-188X)
Volume 18, Number 4, 2021

Identifying Events In Social Media Streams In Real Time Using Semantic Analysis Of Irrelevant Phrases

Dr. K E Balachandrudu

Professor (CSE), MRIET, Hyderabad.

Abstract: Social media interactions have made it possible for everyone, regardless of location, to obtain instant access to information about events occurring across the world. Nevertheless, the semantic analysis of social media data is hampered by obstacles such as linguistic complexity, unstructured data, and ambiguity. In this research, the Social Media Analysis Framework for Event Detection was suggested (SMAFED). SMAFED seeks to enhance semantic analysis of noisy phrases in social media streams, content representation/embedding, and event cluster summarization in social media streams. We used essential ideas such as integrated knowledge base, addressing ambiguity, semantic representation of social media streams, and Semantic Histogram-based Incremental Clustering based on semantic relatedness to accomplish this. Two tests were done to verify the methodology. Initially, we assessed the effect of SMAFED's data enrichment layer. SMAFED beat alternative pre-processing frameworks with a loss function of 0.15 on the first dataset and 0.05 on the second. Second, we evaluated SMAFED's accuracy in detecting events in social media streams. This second experiment demonstrated that SMAFED outperformed current event detection methods by achieving higher Precision (0.922), Recall (0.793), and F-Measure (0.853) metric scores. The results of the research demonstrate that SMAFED is a more effective method for event identification in social media.

Keywords: Event detection, Event summarization, Semantic analysis, social media stream, Word sense disambiguation.

I. Introduction

With the use of computational operations, such as event detection, noteworthy happenings may be automatically identified via the examination of social media data. An event is a noteworthy occurrence at a certain location and time [1]. Detecting events in social media streams is an interesting topic of study because it gives users instant access to information about current events and the perspectives of others who are following them. The days of people being able to conceal news they do not want others to know about by "killing" it via a news agency or organisation are over. As a result of the proliferation of social media, this is no longer practicable. When news breaks via social media, it spreads like wildfire. It reaches more people who will keep spreading the word till the desired result is achieved. As such, it is crucial to conduct studies in this field in order to provide an unbiased account of unfolding events like breaking news, quick breakouts, infectious illness, and terrorist acts [2]. Today, because to the proliferation of social media platforms, there is an ocean of information at our fingertips, just waiting to be mined. Taking use of the social aspect of human interaction, social media allows users to share their thoughts and emotions, join a global community, and work together from afar [3]. The content of a social media feed is created by the users themselves. That's why ambiguity, one of the most fundamental issues with texts written in any language, is present in the content of social media streams as well. An ambiguous word is one that might have more than one meaning depending on the setting in which it is used [4, 5]. Humans (i.e., native speakers) may have no trouble with ambiguity since it may be handled via contextual knowledge and common sense. Inefficient computer applications for disambiguating text remain a challenge, though [6]. Short communications; the use of more irregular, informal, abbreviated language; grammatical and typographical errors; a blending of dialects; fuzziness; and incorrect sentence structure are hallmarks of social media streams [7]. Because of these quirks, computational approaches that depend on them tend to underperform [7-12]. Similarly, many existing methods for event identification focus primarily on the retrieval of meaningless keywords or themes. Nonetheless, the valuable semantics encoded in social media streams have not been properly examined [13, 14], which reduces the precision of event recognition. This research focuses on the difficulty in interpreting slangs, abbreviations, and acronyms (SAB) in social media streams, a challenge that arises throughout the event detection process. The semantic analysis of noisy phrases in the form of SAB and related ambiguities has not been accounted for in the design of most contemporary event detection algorithms. Reporting on major incidents or crises may be done effectively using social media. In spite of this, it has a number of drawbacks that make it hard to efficiently discover intriguing and helpful messages. In the context of this work, event summarising may be thought of as the process of selecting a subset of tweets that together best capture the essence of a larger event. Due to the chaotic nature of social media material, precise analysis of social media streams requires significant developments in semantic technology. Therefore, it is important to enhance the accuracy of event recognition methods by eliminating the distracting characteristics of social media posts.

In order to increase the reliability of event identification in social media streams, this research offers a Social Media Analysis Framework for Event Detection (SMAFED) to deal with the ambiguous and noisy phrases that often appear in these channels. To create SMAFED, we combined incremental semantic clustering with a local vocabulary comprised of slangs, acronyms, and abbreviations. The goal is to enhance event detection by facilitating the interpretation of the implicit semantics hidden in social media feeds. Existing methods such as locality sensitive hashing [15], cluster summarization [16], entity-based approach [17], and Repp framework [18] were used to compare SMAFED's performance. Accuracy of event detection was measured using Precision, Recall, and F-measure. SMAFED's pre-processing and enrichment components were compared to various pre-processing frameworks in order to extract feelings from tweets using a generic dataset, Twitter sentiment analysis training corpus, and a dataset of Nigerian provenance called Naija-tweets.

II. Literature Survey

A summary of prior work on event identification from social media feeds is provided here. Events in social media streams may be detected using many methods, including those based on unsupervised learning, semi-supervised learning, supervised learning, and semantic analysis.

Learning with supervision for spotting events in a social network feed When given labelled examples to study, machine learning algorithms belonging to the supervised learning model class are able to generalise a prediction or classification function. Each input (vector) and each output (sample) in the training examples are included (supervisory signal). Each instance is represented by the pair (x, y), where x is a vector and y are the class or target attribute (or scalar). In supervised learning, a mapping model from x to y is constructed by discovering a mapping m(.) such that m(x)=y. The result of an unlabelled instance may be calculated using m(x) and m(.) learnt from training data. After that, we'll look at some examples of supervised learning in action with respect to event detection. In [12], the authors suggested a method for detecting geo-spatial events in a Twitter feed. Naive Bayes, Multilayer perceptron, and Prune C4.5 are three of the machine learning techniques used to determine whether the geo-spatial clusters include actual events. The locations of the observed events (candidate clusters) were shown in real time on a map, ordered by the tweet ranking score. To better collect time, content, and location information from social media, the authors of developed a graphical-based model, location-time-constrain-topic, and LTT (an improvement above LDA). You've got a kill back Uncertain media content similarity was quantified using Lieber, KL-divergence. Variable-dimensional extendible hash indexing was used to identify social gatherings (VDEH). To account for potential subject drift, the LTT model was updated after each batch of incoming tweets in each time slot. In [46], the authors introduced a system called Transaction-based Rule Change Mining (TRCM) that uses Association Rule Mining to derive association rules from a hashtag in a tweet. Changes to the rules of consequence and conditionality were graded and compared across all available time periods. Hashtags were then matched with the ground truth from BBC Sport commentary throughout the same time period to see whether they were similar. Top-R topic identification in real-time on Twitter, including topic hijack filtering, was the subject of research by the authors of [19]. Using the Streaming Non-negative Matrix, we were able to incorporate the extraction of relevant themes and the filtering of noisy messages from the Twitter stream (NMF). Due to a misspecification of the model, false positives (false negatives) of hijacked themes occurred. Using TF-IDF for the similarity score and SFPM for the classification, [18] introduced the Twitter Life Detection Framework. When dealing with huge vocabularies, TF-IDF might be sluggish since it directly calculates document similarity on the word-count space. No methods were mentioned that specifically addressed the issue of how to deal with SAB keywords in tweets. According to [19], TF-IDF and SVM may be used together to classify social media events in a multimodal manner. Stop words, special characters, digits, emojis, HTML elements, and words with less than four letters were all eliminated during the pre-processing phase. In [20], the authors suggested a method for detecting multimedia events using recurrent neural networks that relies on audio. In order to establish whether or not an event may be linked to a video, the authors used a recurrent neural network for feature representation and classification, which takes into account a wider range of temporal information. Along these lines, [21] introduced the concept of multimodal event detection.

By using a two-stage convolutional neural network, the author suggested techniques for recognising complicated events in online movies. In this work, we leverage the chaotic nature of user-generated material seen in social media as our inspiration (social media content). There was no mention of any attempts to identify SAB words or grammatical mistakes in video data, despite the prevalence of both in the media. Foodborne illness detection using Weibo data was also the subject of [22], which developed a method based on Text Rank and SVM. An SVM was used to weed out irrelevant tweets. The suggested framework, however, was proven to be subpar when confronted with sparsity and idea drift. The work introduces a deep learning strategy for identifying road accidents from social media. More than three million tweets were parsed for tokens and paired tokens. The tokens and pairs were fed into deep learning models such as Deep Belief Network and Long Short-Term Memory to identify traffic accidents. In order to identify hate speech directed towards marginalised groups on Facebook using Amharic text data, the authors suggested a hate speech recognition algorithm. Data cleansing and feature extraction were carried out using the distributed Apache Spark platform. Word2Vec was used as the embedding model for feature extraction. The classification phase made use of Gated Recurrent Units. To review the various supervised learning methods used for event detection, see Table 1.

Table 1: Different supervised learning approaches for event detection

Author	Data sources/ Features	Algorithms
Zhou X, Chen L	Twitter/Textual, Spatial	Variable Dimensional Extendible Hash
Adedoyin-Olowe M	Twitter/Textual	Association Rule Mining
4. Walther M, Kaisser M	Twitter/Textual, Geo-spatial	Naïve Bayes, Multilayer perceptron, and Prune C4.5
Hayashi K, Maehara T	Twitter/Textual	Streaming Non-negative Matrix
Gaglio S, Rea GL	Twitter/Textual	Soft Frequent Pattern Mining
Zeppelzauer M	Flickr, Instagram/ Multimodal	TF-IDF, SVM
Wang Y, Neves L	Video/Multimedia	RNN, LSTM
Lan Z	Video/Multimedia	CNN, Smoothing technique
Cui W, Wang P	Weibo/Textual	Text Rank, SVM
Zhang Z, He Q, Gao J	Twitter/Textual	Deep Belief Network, LSTM
Mossie Z, Wang JH	Facebook/Textual	Word2Vec, Random Forest, Gated Recurrent Unit, LSTM

Event detection in social media streams using semantic techniques Scalable distributed event detection utilising Twitter streams was the focus of the authors. Using a lexical key partitioning technique to distribute the detection process over several workstations, the study proposes scalable automated distributed real-time event detection. The suggested framework has been deployed using a Storm architecture. Despite the fact that Twitter streams are chaotic, ephemeral, and rife with slang, it was discovered that no pre-processing was performed. Locality Sensitive Hashing was suggested for social media event detection by authors. In order to acquire events from both Twitter and Facebook, LSH was employed twice throughout the event identification procedure. It was then used to find instances when the two social media streams intersected. In [17], we suggested LITMUS, a system that would mine social media for information on landslides using keywords. The system then used a semantic interpreter and an enhanced Explicit Semantic Analysis algorithm to identify material as either relevant or irrelevant by extracting characteristics from a section of Wikipedia. The location was estimated using a semantic clustering technique based on semantic distance. Instead of analysing the complete data set, researchers focused just on the portions that had geographical references. These semantically-based methods did not evaluate how SAB words were handled. Arm Tweet, described by the authors, is a system that uses NLP to parse tweets for structured data, which is then combined with RDF from DBpedia and WordNet. The system employed semantic searches to find relevant tweets, and then sent them to an anomaly detection algorithm to see whether they matched to real-world occurrences. This enhances keyword search and may be used for event detection that is niche-specific. Acronyms, slang, abbreviations, and passive words are common in social media data, but the accuracy of the pre-processing component was not studied.

Author	Data sources/ Features	Algorithms
Kaleel SB, Almeshary M	Twitter, Facebook/ Textual	LSH
McCreadie R, Macdonald C	Twitter/ Textual	Hash key grouping
8. Tonon A, Cudré-Mauroux P	Twitter/ Textual	RDF
Romero S, Becker K	Twitter/ Textual	TF-IDF, Named entity Recognition, Page Rank, CfsSubsetEval
Sun X, Liu L, Ayorinde A	Twitter/Textual	IPLSA, EM, RS Scoring algorithm, word2vec

In [19], the author proposes a system for event classification in tweets that makes use of a combination of TF-IDF, Named Entity Recognition, and Page Rank for semantic enrichment. Tweets were sorted using a combination of named entity extraction, external document enrichment, and semantic enrichment. An event detection technique based on scoring and word embedding was presented by the authors of [20] to identify significant events in large data streams. Stop words, auxiliary verbs, URLs, and emoticons were eliminated during the pre-processing phase. As a means of embedding, we employed Word2vec, which resulted in a significant increase in the process of detecting events was conducted using expected maximisation. Word2Vec can only do one thing: find comparable words. Nonetheless, not one of these methods thought about clarifying SAB words in online conversations. In this article, we provide a brief overview of the ways in which semantic-based approaches have been put to use in event detection techniques. Based on our research, we know that prior approaches to event recognition in social media either ignored or filtered out SAB during the preliminary processing of social media streams. Unfortunately, they did not engage in semantic analysis of potentially misleading terminology like "SAB" to ascertain how their use affected the precision of their findings. Short messages, slang, acronyms, mixed languages, grammatical and spelling errors, dynamically evolving, irregular, informal, abbreviated words, and improper sentence structure are examples of the types of "noisy terms" that pose a challenge to the efficient performance of learning algorithms [7, 8]. This void in knowledge was the impetus for our investigation. According to [24], the semantics of social media material must be maintained in any depiction of a social media stream. Therefore, it is essential for relevant and accurate findings to make advantage of the contextual cues around a social media feed. To this end, improving the precision of event identification in social media streams necessitates the creation of a framework that places special emphasis on the semantic analysis of slangs, acronyms, and abbreviations terminology and the ambiguity associated with their use. SMAFED aims to make a difference in this area when most prior research initiatives have failed to do so. Table 3 provides a synopsis of the advantages and disadvantages of several current event detection methods and their characteristics.

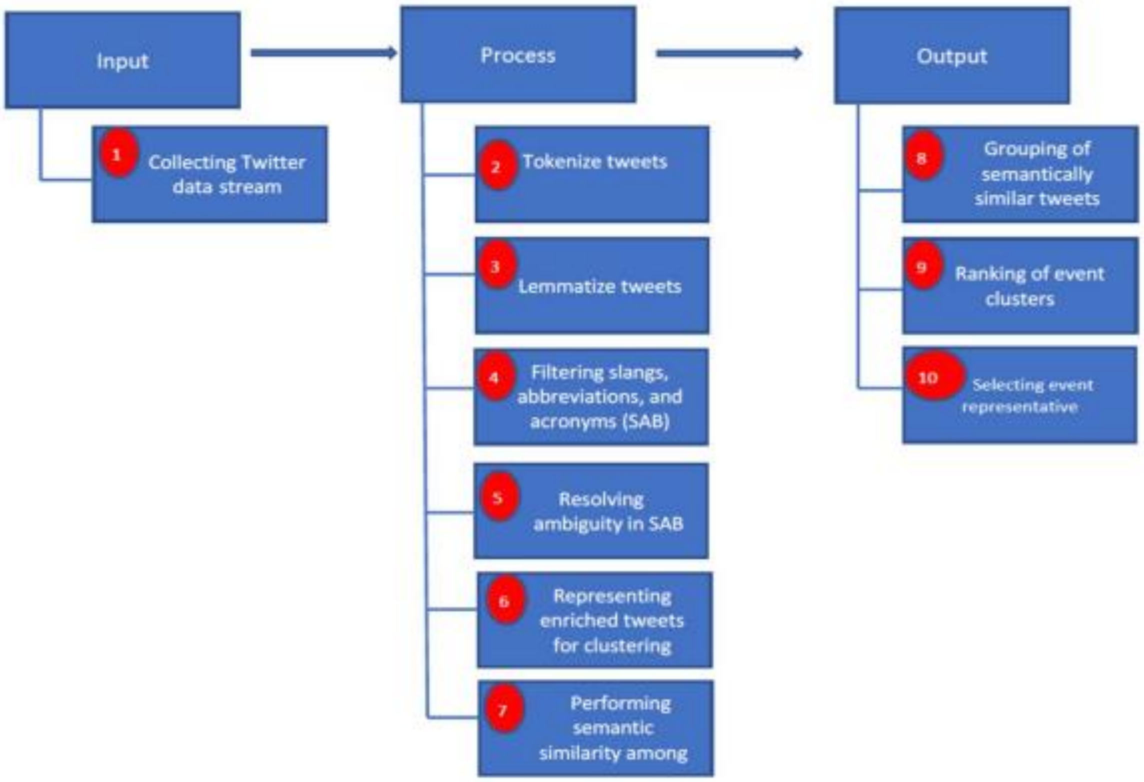
Input
Process
Output
1 Collecting Twitter data stream
2 Tokenize tweets
3 Lemmatize tweets
4 Filtering slangs, abbreviations, and acronyms (SAB)
5 Resolving ambiguity in SAB
6 Representing enriched tweets for clustering
7 Performing semantic similarity among
8 Grouping of semantically similar tweets
9 Ranking of event clusters
10 Selecting event representative

Figure 2: Main Tasks During Event Detection using SMAFED

III. Methods and Metrics

In order to better recognise events in social media, this study introduces the Social Media Analysis Framework for Event Detection (SMAFED), an efficient and integrative method to social media stream analysis that makes use of pre-processing and enrichment of social media streams. Event detection using SMAFED: a summary of the main steps as shown in Fig. 1, our suggested method for event identification in social media streams consists of ten core tasks (1-10), each of which is progressively organised and can be abstracted using the Input-Process-Output paradigm. Task 1 is part of the input phase, tasks 2-7 make up the process phase, and tasks 8-10 are part of the output phase. The steps are listed below. Processes include: (1) obtaining a stream of tweets; (2) tokenizing those tweets; (3) lemmatizing those tokens; (4) filtering out slangs, abbreviations, and acronyms (SAB); (5) resolving ambiguity issues with SAB usage via disambiguation; (6) representing enriched tweets in a form amenable to clustering; and (7) performing semantic similarity among tweets. Aligning tweets with shared meanings into groups 9 clusters of events ranked by importance choosing an indicator of an occurrence.

High-level overview of SMAFED

Here, we show SMAFED in its more abstract, process-level form. The SMAFED procedure (shown in Fig. 2) may be broken down into the following four stages. First, using the Python programming language, a user interface is created around the underlying API offered by Twitter in order to gather tweets written in either Standard English or Pidgin English by people from Nigeria. Python is favoured because of its effectiveness and adaptability for developing data- and traffic-intensive processes. Tweets from each time interval are gathered and queued up. Once tweets have been gathered, they are sent to the pre-processing phase. Secondly, a regular expression was used to strip out any URLs, Tags, mentions, and non-ASCII characters from the gathered data stream. Next, tokenization and normalisation of the data needed to be completed. By cutting down on the number of characteristics, this simple pre-processing helps prevent overfitting. After that, the tweets are cleaned of slang, acronyms, and abbreviations by comparing them to dictionaries of standard English terms included in the natural language toolkit (NLTK). SAB words are filtered and sent to the IKB for additional analysis.

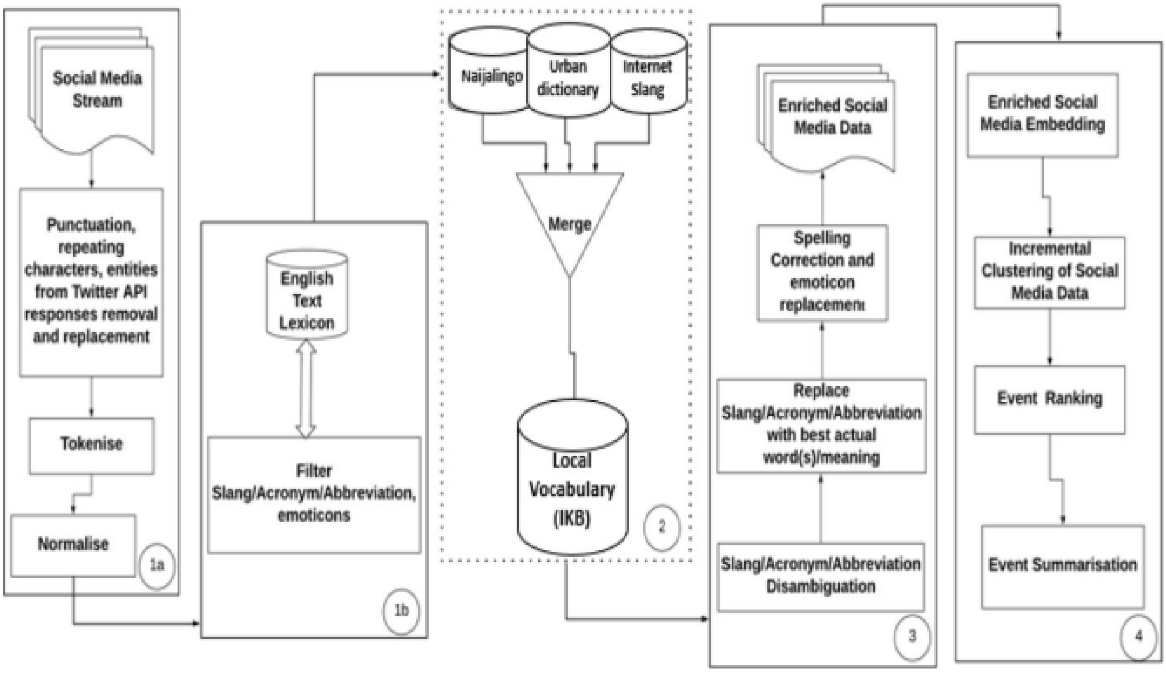
Social Media Stream
Punctuation, repeating characters, entities from Twitter API responses removal and replacement
Tokenise
Normalise
1a
English Text Lexicon
Filter Slang/Acronym/Abbreviation, emoticons
1b
Naijalingo
Urban dictionary
Internet Slang
Merge
Local Vocabulary (IKB)
2
Enriched Social Media Data
Spelling Correction and emoticon replacement
Replace Slang/Acronym/Abbreviation with best actual word(s)/meaning
Slang/Acronym/Abbreviation Disambiguation
3
Enriched Social Media Embedding
Incremental Clustering of Social Media Data
Event Ranking
Event Summarisation
4

Figure 2: A View of SMAFED Process Workflow

The next step involves pulling definitions of SAB from the IKB. There are several interpretations of each SAB, so it's important to go through them all and choose the one that makes the most sense. The Slang, Acronym, and Abbreviation, Disambiguation Algorithm is used to determine the intended meaning of a SAB in a tweet when there is more than one possible interpretation. In order to give a wealth of information and enhance overall disambiguation accuracy, SABDA of the IKB was selected on the basis of its unique characteristics. The last steps of the data enrichment process include replacing missing emoticons and fixing spelling errors (with the help of Python's built-in spell-checking module). The fourth step involves adapting the enhanced tweets from the previous step so that the clustering algorithm may use them as input. Through the use of the sent2vec paradigm, enhanced tweets are converted to a vector representation. Compared to other supervised and unsupervised algorithms for sentence or paragraph embedding, Sent2Vec is a major step forward, as shown by the literature [19]. Using semantic histogram-based incremental clustering, the embedded tweets are grouped in real time. The goal is to establish groups that are as similar as possible in terms of the events they reflect. Within clusters, SHC keeps cohesion strong, suggesting that similarities are widely dispersed. This makes SHC the only viable option. Each group of tweets is then evaluated according to the information density of its individual words. After identifying the top n candidates for event clusters, we next choose a single tweet to serve as a sample example for each cluster.

The impact of the data enrichment layer of SMAFED SMAFED was evaluated by benchmarking it with the General Social Media Feed pre-processing Method (GSMFPM) to determine the impact of the enrichment layer of SMAFED.

S/N	Dataset	Source(s)	Total	Selected	Training/Testing

1	Twitter sentiment analysis training corpus	1. University of	1,578,627	104,857	83,886/20,971
		Michigan Sentiment	1,048,575	(10%)	
		Analysis on Kaggle	(After		
		2. Twitter sentiment	download)		
		corpus by Niek Sanders			
2	Naija-Tweets	Extracted from	12,920	12,920	10,336/2,584
		Nigeria origin		(100%)	

S/N	Dataset	Unigram	Bigram
1	Twitter sentiment analysis training	76,522	501,026
	corpus	Top-K word (50,000)	Top-K (150,000)
2	Naija-Tweets	3,296	10,187
		Top-K (3,000)	Top-K (8,000)

Dataset description

Experiment one used the Twitter sentiment analysis training corpus and the Naija-tweets datasets.

The Process of Feature Extraction and Representation

For each dataset, we took out both unigram and bigram characteristics. The feature extraction process made use of the algorithm developed by the Global Vector for Word Representation (GloVe). To extract word-word co-occurrence data from a corpus, GloVe uses an unsupervised learning approach. As a consequence, the linear structures of the word vector space are highlighted in the resulting representations. Combining elements of both the local context window and the global matrix factorization approaches, Glove is a log-bilinear model with weighted least squares. According to the model's fundamental concept, there is meaning encoding potential in the ratios of word co-occurrence.

The supervised learning methods multilayer perceptron and convolutional neural networks were used for text categorization. The MLP method models the interdependencies between inputs and outputs by training on input-output pairs. The Convolutional Neural Network (CNN) is a kind of deep learning architecture that use convolutions to learn higher-order features in data.

IV. Result And Discussion

Two different classifiers were used to compare the performance of the proposed SMAFED to that of the General Social Media Feed Pre-processing Method (GSMFPM). Analysis of SAB words and resolution of ambiguity in SAB in social media streams were at the heart of comparing the effects of the conventional pre-processing approach - GSMFPM and the suggested SMAFED on the classifiers. As part of this process, we examined the classifiers' (MLP and CNN) cross-entropy loss function while using GSMFPM and SMAFED. An evaluation of a classifier's predictive ability is made via a loss function comparison. Tables following exhibit the cross-entropy results for sentiment classification using the Twitter sentiment analysis training corpus, using Multilayer Perceptron and Convolutional Neural Networks with five and eight epochs, respectively. Tables 1 and 2 exhibit the Naija-tweets dataset simultaneously. The table below shows the cross-entropy loss function for the multilayer perceptron used by GSMFPM and SMAFED on the Twitter Sentiment Analysis Training Corpus for Unigram, Bigram, and Unigram +Bigram words. When comparing how well the anticipated and real sentiments matched, SMAFED fared better than GSMFPM, as seen in the table. It's also worth noting that epoch 5 yielded the lowest loss function for both methods across the board for both unigrams and bigrams and unigrams and bigrams, indicating that the higher the number of epochs, the better the classifier's performance.

Table 1 Multi-layer perceptron cross-entropy loss function for GSMFPM and SMAFED on twitter sentiment analysis training corpus

	FEATURES					
	UNIGRAM		**BIGRAM**		**UNIGRAM + BIGRAM**	
EPOCH	**SMAFED**	**GSMFPM**	**SMAFED**	**GSMFPM**	**SMAFED**	**GSMFPM**
1	0.5478	0.5499	0.5612	0.5729	0.5535	0.5405
2	0.4911	0.5132	0.4405	0.4872	0.5004	0.5192
3	0.4537	0.4658	0.3909	0.4427	0.471	0.5087
4	0.4045	0.4147	0.3015	0.3528	0.4047	0.4584
5	0.3741	0.3762	0.2181	0.3177	0.3484	0.3698

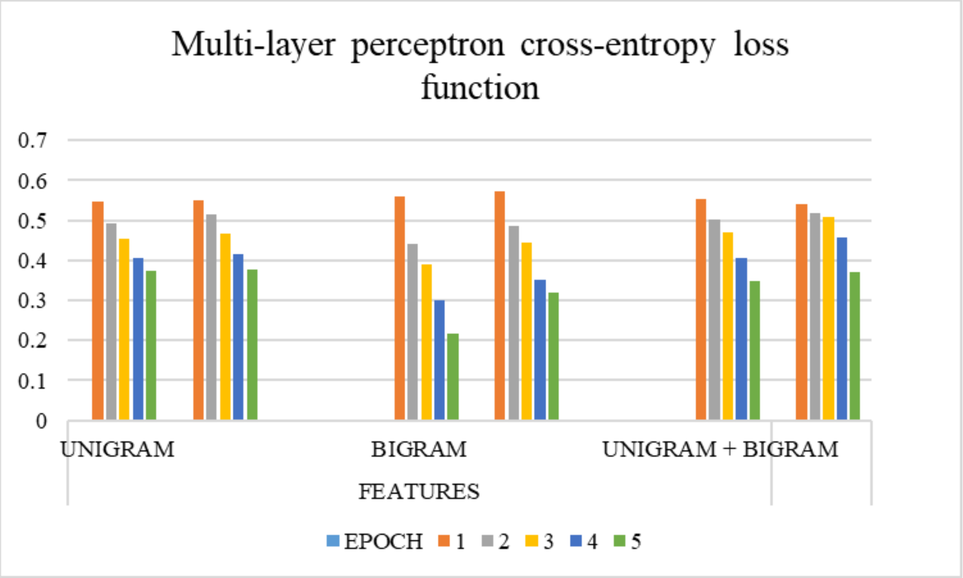

Figure 3: Multi-layer perceptron cross-entropy loss function

Table 2 Four-Layer convolutional neural network cross-entropy loss function for GSMFPM and SMAFED on twitter sentiment analysis training corpus

	1-LAYER		2-LAYER		3-LAYER		4-LAYER	
EPOCH	**SMAFED**	**GSMFPM**	**SMAFED**	**GSMFPM**	**SMAFED**	**GSMFPM**	**SMAFED**	**GSMFPM**
1	0.5852	0.4369	0.636	0.7843	0.6279	0.7762	0.5584	0.7067
2	0.4761	0.3928	0.5399	0.6232	0.4927	0.576	0.4713	0.5546
3	0.4213	0.3725	0.4674	0.5162	0.4434	0.4922	0.4263	0.4751
4	0.3622	0.3556	0.4195	0.4261	0.3943	0.4009	0.3836	0.3902
5	0.3026	0.3397	0.3797	0.4168	0.3515	0.3886	0.332	0.3691
6	0.2241	0.3255	0.3403	0.4417	0.299	0.4004	0.277	0.3784
7	0.1864	0.3002	0.3045	0.4183	0.2736	0.3874	0.2501	0.3639
8	0.1496	0.2892	0.2576	0.3972	0.2425	0.3821	0.2169	0.3565

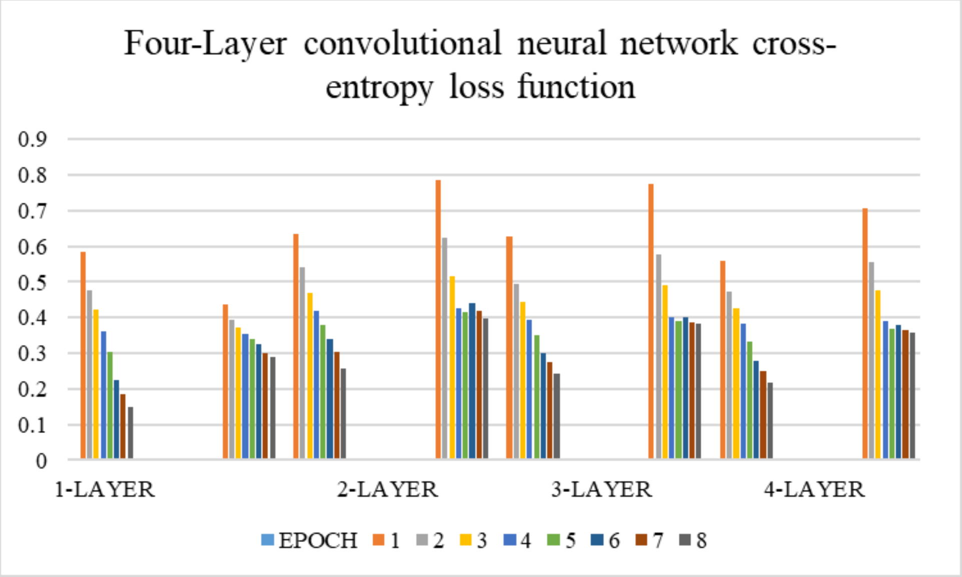

Figure 4: Four-Layer convolutional neural network cross-entropy loss function

Table2 shows the results of a comparison between SMAFED and GSMFPM on the Twitter Sentiment Analysis Training Corpus using the Cross-Entropy Loss Function of a CNN with kernel size=3 and one-four convolution layers and eight epochs. Based on the data in the table, it is clear that SMAFED's pre-processing and data enrichment components performed better than GSMFPM when it came to comparing projected and real sentiment. It's also worth noting that, for both methods, the loss function of the first layer of CNN cross entropy is less than the loss functions of the subsequent layers. Comparison of SMAFED and GSMFPM on the Naija-Tweets dataset based on the Cross-Entropy Loss Function of Multilayer Perceptron with Unigram, Bigram, and Unigram +Bigram features across five iterations is shown in Table 1. SMAFED's performance in the table above reveals that it is superior than GSMFPM's when it comes to predicting how people would feel. Using the Twitter Sentiment Analysis Training Corpus and the Naija-Tweets datasets, it was shown that epoch 5 yielded the lowest loss function for the unigram, bigram, and unigram +bigram techniques. The experiment results showed that the suggested SMAFED outperformed the standard pre-preparing method. Also, its precision has increased. This highlights the necessity of employing a local vocabulary in pre-processing social media feeds to disambiguate the noisy phrases included within the social media feeds from a particular origin. For the Naija-Tweets dataset, Table 2 shows the Cross-Entropy Loss Function of a CNN with kernel size=3 and one-four convolution layers and eight epochs for SMAFED and GSMFPM. According to the statistics, SMAFED's pre-processing and data enrichment components perform better than GSMFPM when it comes to comparing projected and real sentiment. It's also worth noting that the loss function of CNN cross-entropy in the first layer is less than in later layers for both methods.

SMAFED efficiency

Run-time performance measures were used to evaluate the SMAFED framework's efficacy and viability. Python was used to implement the suggested approach of event detection (v 3.7). For our tests, we used a 64-bit version of Windows 10 on a machine equipped with an Intel(R) Core (TM) i5-6200U CPU at 2.30 GHz, along with 12 GB of RAM. In order to get the framework up and running in the cloud, we used a Docker Droplet with 4 GB of space provided by Digital Ocean. Twitter posts from tit took 5 second she Nigeria were utilised in the prototype's development. It was discovered that without any kind of filter, the average number of tweets coming from Nigeria is 45 every minute. When compared to the global average of 350,000 tweets per minute, this demonstrates how little the scope of tweeting in Nigeria really is. Spell check was the slowest component of the operation. Getting 40 tweets ready for clustering took under 5 seconds, including pre-processing and enrichment. Each tweet takes roughly 0.125 seconds on average to process. We're well inside the parameters required to handle Nigeria's typical daily tweet volume of 1. In the very unlikely scenario that every tweet is treated as an event, the system will be able to process eight times as many tweets originating in Nigeria. The maximum retention period for an event cluster in SMAFED is 4 days. This presumption was founded on the idea that newer data had less potential value. The decision to give each cluster of events just 4 days to live was chosen to save memory use, keep only a fixed number of clusters in RAM at any one time, and cut down on the sheer volume of necessary comparisons. Figure 4 shows the effectiveness of SMAFED with regards to pre-processing and tweet streaming from a Nigerian origin.

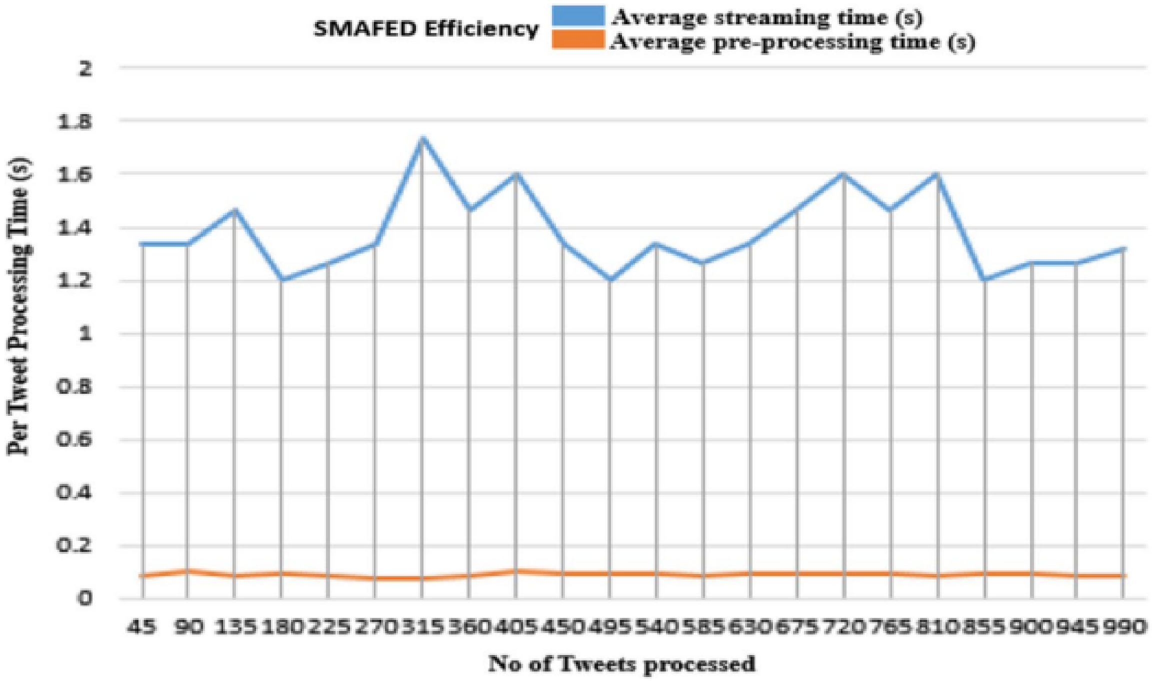

Figure 4: SMAFED Evaluation. The average processing time for each tweet is 0.125 s

V. Conclusion and further work

In this research, we introduce SMAFED, a social media analysis framework that analyses the rich but hidden information in social media streams to enhance the precision with which events may be detected. Improved metric scores in Precision (0.922), Recall (0.793), and F-Measure demonstrate that the suggested SMAFED method is superior to previously proposed methods for event detection (0.853). As an additional step, SMAFED was compared to GSMFPM in order to gauge the data enrichment layer's efficacy. As measured by cross-entropy, SMAFED beat GSMFPPM on the sentiment classification of the Twitter sentiment analysis training corpus and the Naija-Tweets dataset while using Multilayer Perceptron and Convolutional Neural Networks with five and eight epochs, respectively. In the field of big data analytics, this work makes a contribution to the study of event detection in social media streams. To be more specific, it (1) does semantic analysis of SAB phrases together with ambiguity in their use, hence addressing the constraints identified in prior event detection systems. This improves our ability to (1) recognise and make sense of noisy phrases in social media streams, (2) distinguish across similar but distinct SAB terms, and (3) create a unified knowledge base to aid in the semantic analysis of such terms. For the purpose of this work, SMAFED relied only on text extracted from social media streams to identify events. If photographs and their accompanying text from social media streams are used, the event detection result is strengthened even more. While Twitter is often used for research, more events may be noticed and findings can be more easily standardised by investigating and/or combining it with data from other social media sources. Since just a few methods have made use of this media, much remains to be learned about it.

VI. References

1. Panagiotou N, Katakis I, Gunopulos D. Detecting events in online social networks: defnitions, trends and challenges. In: Michaelis S, editor. Solving large scale learning tasks: challenges and algorithms. Cham: Springer; 2016. p. 42–84
2. Win SSM, Aung TN. Automated text annotation for social media data during natural disasters. Adv Sci Technol Eng J. 2018;3(2):119–27.
3. Olsson T, Jarusriboonchai P, Wozniak P, Paasovaara S, Vaananen K, Lucero A. Technologies for enhancing collocated social interaction: review of design solutions and approaches. Comput Supported Coop Work (CSCW). 2020;29:29– 83. https://doi.org/10.1007/s10606-019-09345-0.
4. Carbezudo MAS, Pardo TAS. Exploring classical and linguistically enriched knowledge-based methods for sense disambiguation of verbs in Brazilian Portuguese news texts. Nat Lang Process. 2017;59:83–90.
5. Gutierrez-Vazquez Y, Vazquez S, Montoyo A. A semantic framework for textual data enrichment. Expert Syst Appl. 2016;57:248–69.
6. Alkhatlan A, Kalita J, Alhaddad A. Word sense disambiguation for Arabic exploiting WordNet and word embedding. Procedia Comput Sci. 2018;142:50–60.
7. Kolajo T, Daramola O, Adebiyi A, Seth A. A framework for pre-processing of social media feeds based on integrated local knowledge base. Inf Process Manag. 2020;57(6):102348.
8. Atefeh F, Khreich W. A survey of techniques for event detection in Twitter. Comput Intell. 2015;31(1):132–64.
9. Jain VK, Kumar S, Fernandes SL. Extraction of emotions from multilingual text using intelligent text processing and computational linguistics. J Comput Sci. 2017;21:316–26.
10. Rao D, McNamee P, Dredze M. Entity linking: fnding extracted entities in a knowledge base. In: Poibeau T, Saggion H, Piskorski J, Yangarber R, editors. Multi-source, Multilingual information extraction and summarization. Theory and Applications of Natural Language Processing. Heidelberg: Springer; 2013. p. 93–115.

11. Singh T, Kumari M. Role of text pre-processing in Twitter sentiment analysis. Procedia Comput Sci. 2016;89:549–54.
12. Zhan J, Dahal B. Using deep learning for short text understanding. Journal of Big Data. 2017;4:34. https://doi.org/10. 1186/s40537-017-0095-2.
13. Katragadda S, Benton R, Raghavan V. Framework for real-time event detection using multiple social media sources. Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS). Waikoloa, Hawaii, 2017. p. 1716–1725
14. Xia C, Schwartz R, Xie K, Krebs A, Langdon A, Ting J, Naaman, M. CityBeat: Real-time social media visualisation of hyperlocal city data. Proceedings of the 23rd International World Wide Web Conference Committee (IW3C2). Seoul, South Korea. 2014. p. 167–170.
15. Petrovic S, Osborne M, Lavrenko V, Streaming frst story detection with application to Twitter. Proceedings of Human Language Technologies: The Annual Conference of American Chapter of the Association for Computational Linguistics Los Angeles. CA, USA. 2010;2010:181–9.
16. Aggarwal CC, Subbian K. Event detection in social streams. Proceedings of the SIAM International Conference on Data Mining. California, USA, 2012. p. 624–635.
17. McMinn AJ, Jose AM. Real-time entity-based event detection for Twitter. In: Mothe J, editor. Experimental IR Meets Multilinguality, Multimodality, and Interaction. Cham: Springer; 2015. p. 65–77.
18. Repp QK. Event detection in social media: Detecting news event from the Twitter stream in real-time (Master’s thesis). Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway, 2016.
19. Boushaki SI, Kamel N, Bendjeghaba O. High-dimensional text datasets clustering algorithm based on cuckoo search and latent semantic indexing. J Inf Knowl Manag. 2018;17(3):1–24.
20. Weng J, Lee BS. Event detection in Twitter. ICWSM. 2011;11:401–8.
21. Zubiaga A, Spina D, Amigó E, Gonzalo J. Towards real-time summarization of scheduled events from Twitter streams. Proceedings of the 23rd ACM Conference on Hypertext and Social Media. Milwaukee, WI, USA. 2012. p. 319–320.
22. Lee C. Mining Spatio-temporal information on microblogging streams using a density-based online clustering method. Expert Syst Appl. 2012;39(10):9623–41.
23. Abdelhaq H, Sengstock C, Gertz M. EvenTweet: Online localized event detection from Twitter. Proc VLDB Endow. 2013;6(12):1326–9.
24. Abhik D, Toshniwal F. Sub-event detection during natural hazards using features of social media data. Proceedings of 22nd International Conference on World Wide Web New York, NY: ACM. 2013. https://doi.org/10.1145/2487788.24880 46.
25. Fuchs G, Andrienko N, Andrienko G, Bothe S, Stange H. Tracing the German centennial food in the stream of tweets: First lessons learned. Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, ACM. GEOCROWD ’13. Orlando, FL, USA, 2013. p. 31–38.