advances in Technology

Adv. Technol. ***2021****, 1(2), 393-418*

Review

Literature Review on Real-time Location-Based Sentiment Analysis on Twitter

Dilmini I.G.U Rathnayaka,[a,*] Pubudu K.P.N Jayasena,[b] Iraj Ratnayake[c]

[a]Faculty of Graduate Studies, Sabaragamuwa University

[b]Department of Computing and Information Systems, Sabaragamuwa University -

[c]Department of Tourism Management, Sabaragamuwa University

Email Correspondence: D. I.G.U Rathnayaka (dilminiimbulegama@gmail.com)

Received: 23 May 2021; Revised: 13 August 2021; Accepted: 14 August 2021; Published: 31 Aigust 2021

Abstract

Sentiment analysis mainly supports sorting out the polarity and provides valuable information with the use of raw data in social media platforms. Many fields like health, business, and security require real-time data analysis for instant decision-making situations.Since Twitter is considered a popular social media platform to collect data easily, this paper is considering data analysis methods of Twitter data, real-time Twitter data analysis based on geo-location. Twitter data classification and analysis can be done with the use of diverse algorithms and deciding the most appropriate algorithm for data analysis, can be accomplished by implementing and testing these diverse algorithms.This paper is discussing the major description of sentiment analysis, data collection methods, data pre-processing, feature extraction, and sentiment analysis methods related to Twitter data. Real-time data analysis arises as a major method of analyzing the data available online and the real-time Twitter data analysis process is described throughout this paper. Several methods of classifying the polarized Twitter data are discussed within the paper while depicting a proposed method of Twitter data analyzing algorithm. Location-based Twitter data analysis is another crucial aspect of sentiment analyses, that enables data sorting according to geo-location, and this paper describes the way of analyzing Twitter data based on geo-location. Further, a comparison about several sentiment analysis algorithms used by previous researchers has been reported and finally, a conclusion has been provided.

Keywords: Twitter data, geo-location, data analysis

Introduction

The advancements of microblogging websites help the users to express and share pictures, videos, feelings, opinions, thoughts and texts, etc. According to the data in Statista, the most popular microblogging websites are Facebook, YouTube, Facebook,

https://doi.org/10.31357/ait.v1i2.4936

Instagram, Twitter,etc, which offer more information which is very helpful for doing sentiment analysis[1]. Twitter can be mentioned as a proper platform to gain user-uploaded data due to the availability of Twitter streaming API and it helps to stream the data successfully and easily to the model. Thus, there are several existing models to analyze the offline twitter data, requirement of real-time analytics is a more interesting aspect for decision-making in business, security, healthcare,etc fields[7]. The importance of analyzing the real-time Twitter data can be signified as generation of visions while data streaming rather than storing analyzing with the use of several steps. The geo-location is a vital factor of user-uploaded content in social media due to the factor of recognizing the exact location of that content. Location-based Twitter data allows to extract the location-based tweets and data analysis can be done according to the geo-location. With the use of geolocation-based Twitter data on the related topics, data analysis can be accomplished[8].Simply, people are used to uploading pictures as tweets to represent the condition and situation of a subjective object which can be considered as the simplest method to express emotions on social media platforms.

Sentiment Text Analysis

The sentiment analysis of this huge amount of visual content could be more helpful to extract the views, opinions and emotions about subjective things, events and topics. The word "Sentiment" refers to expressing feelings, views and opinions. Moreover, sentiment Analysis can describe as a basic method of mining opinions which can be introduced as a vital Neuro-Linguistic Programming (NLP) task[2]. Rather than only using the text contents, social media users are tending to share videos and images to share their experiences. The sentiment analysis from both textual contents and visual contents enables to make of various predictions according to the source extracted data[9]. According to the literature, sentiment Analysis is the exercise of natural language processing, statistics, text analysis, machine learning and computational linguistics [3,4]

to extract or mine the subjective information from the texts. These text files could be user reviews, comments, judgments, emotions etc.Further, sentiment analysis is described as the classification of main text according to the polarity that could be positive, negative or neutral [2, 5]. Commonly, the major intention of sentiment analysis can be defined as the combination of above all mentioned factors.

There are several challenges in sentiment analysis wspecific thinghen identifying the objects, extracting the features while discovering the alignment of opinions. Sentiment analysis can be performed according to three major types as (1) Document-level (2) Sentence level (3) Feature or Aspect level [13]. The document-level sentiment analysis is a sort of classification technique to detect the overall polarity of a subjective topic neglecting the opinion holders. Document-based sentiment analysis is considering the pinion about a subjective topic which is exhibited by the documents. Mostly document-based sentiment analysis is more effective during the analysis of reviewing products, movies, etc when only one document is revealing the opinion about a movie or product. A gathering of sentences appeared as a document and sentence-based sentiment analysis assumes that every respective sentence expressing a single opinion. In fact, the sentence based sentiment analysis divides into two subcategories according to the purpose of analysis, namely subjectively detection and opinion detection[14]. The subjective detection methods are demanding to identify the personal conditions for instance, the emotions and opinions. Subjectivity detection is a crucial sub-activity of sentiment analysis since it guarantees the filtration of factual information was completed before sending to the polarity classifier.Opinion detection is another vital part of sentiment analysis which enables to identify the opinions of subjective content, which could be negative, positive or neutral with the help of textual content. For example, if thinking about visiting a hotel, a customer would tend to read the reviews about that hotel which could be negative, positive or neutral. These are highly affecting to make the correct decision about the hotel while motivating to visit there.

The sentiment text analysis of Twitter data counts on the analysis of textual tweets. The tweets of specific things can be identified using both subjective detection and opinion detection before sending them to polarity classifiers.

Sentiment Analysis of Twitter Data

The following figure illustrating the process of sentiment analysis of Twitter data.

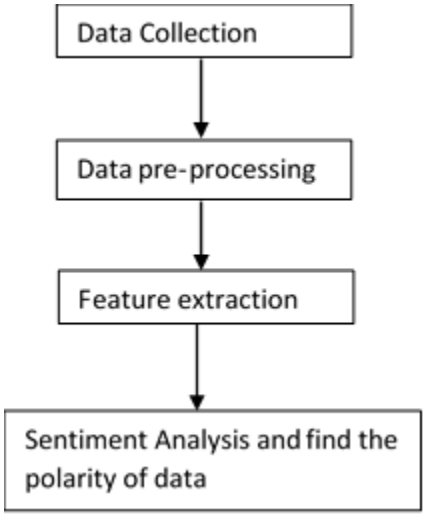

Figure 1. Sentiment Twitter data analysis process

Sentiment Analysis of Twitter data is a process of (1)data collection, (2)data pre-processing, (3)feature extraction, (4)Sentiment Analysis and find the polarity of data [15].

(1) Data Collection

Even though a model is desired to analyze the real-time Twitter data, it should test the performance and accuracy before feeding and analyzing the real-time data. Hence, after the development of the algorithm, it should be trained and tested. For that purpose, a previously collected dataset should be used and this dataset is called a training dataset. When specifying the Twitter platform, training the Twitter dataset is a collection of tweets that are accurately labelled, pre-processed and they perform as the baseline of the developing algorithm. This is directly affecting to better model performance and more quantity of training data enables to enhance the accuracy of the developing model. As

mentioned by Hua, Wang, Zheng, and Zhou[17] the minimum number of 600 tweets for each classified or group, is adequate to generate a training dataset. When it comes to feeding the real-time data to the model, Twitter API is much more supportive in extracting the Twitter available data which are called tweets. Several widespread parameters of data filtering procedures are settling during data extraction. Twitter API is very helpful to run the queries when the data filtration parameters are implemented[15,16]. When real-time Twitter data is extracting, it may consist of information like user IDs, textual content, emojis, URLs, white spaces, hashtags,the latitude and longitude coordinates if the user added the location as public, pictures etc.

(2) Data Pre-processing

The extracted Twitter data may consist of irrelevant data and which will be useless and consists of various kinds of characters. During this step, the Tweets are filtered by removing the irrelevant content like URL, white spaces, hashtags, "@" symbol of usernames, characters, symbols, emojis etc. Natural Language Processing (NLP) tool is an appropriate tool to filter this irrelevant data[18,19]. NLP tools can be used to check grammar, language translations and to convert the speech into textual contents. There are about 50 predefined dependencies are available with NLP and nsubj, amod, dobj are the three most practicing dependencies among them[15]. These are also called relations and the more relevant and meaningful tweets are identified by these dependencies. But these dependencies are not applicable in data filtration procedures. The nsubj dependency is helping to detect the cognation between nouns, verbs and adjectives. Stemming is a NLP method of data pre-processing by removing the inflectional words for example the "travelling" becomes "travel" and "visiting" becomes "visit" by removing "ing". Lemmatization is another method of NLP that enables Twitter data cleaning by properly minimizing the inflected words.The pre-processed Twitter data can be stored in

HDFS[7](Hadoop Distributed File System) and it enables to storage huge amount of data before feeding to the machine learning algorithm.

(3) Feature Extraction

At the beginning of feature extraction, the more prominent features are collected. In this step, the collected textual content is converting to a feature vector using the data-driven approach[13]. The common features used in sentiment analysis are Term Presence Vs Term Frequency, N-gram Features, Parts of Speech, Term Position, Negation[20].

3.1 Negation

By this feature, the negative words, which are associated with positive words in a sentence, the polarity is reversing positive to negative. For example, "not a good place" consists of "good" but the overall opinion is negative. Even though the sentence consists of "good", this feature inverts the polarity into negative.

3.2 Term Presence Vs Term Frequency

This method is also called Spare Vector Representation. “Term Frequency” detects the count of terms that emerged in the collected text content. The “Term Presence” helps to confirm whether the term is contained within the sentence or not, which is considered as a binary-valued vector. 1 and 0 are using to distinguish the availability where 1 is denoting presence and 0 denotes the absence.

3.3 N-gram Features

This feature is mostly applying in NLP. N- grams are known as the contiguous sequence of a term within a text sentence[14]. N- gram is considered as the count of terms that appear in a text sentence. Unigram is considering only one term as feature extraction

while bigram is considering two terms to extract as features. According to the literature [21], Bigrams and Trigrams are effectively performing feature extraction.

3.4 Parts of Speech (POS) Tagging

When considering the English language, both spoken and writing consist of verbs, adjectives and adverbs. These are helping to express peoples' opinions. POS tagging feature is assisting to detect the tagged words in the textual content. The irrelevant words are neglected by this feature while the adjectives, adverbs and verbs are collecting. Due to the removal of irrelevant words, the vocab size is minimizing.

(4) Sentiment Analysis

Sentiment analysis methods can be classified as machine learning-based, lexicon-based and hybrid methods, basically[22]. When developing a model, a collected twitter data set which can be named as a training dataset should train with the classification algorithm to check the performance of the model. This dataset is used to extract the features as mentioned in the above step and these features are used to categorize the polarity of the extracted data and after feature extraction, this dataset is considered as the inputs of the model [23]. The extracted textual twitter data or the inputs are analyzed during the sentiment analysis step and there are several methods for this purpose. Selecting one or more appropriate methods depends on the final analysis result requirements and nature.

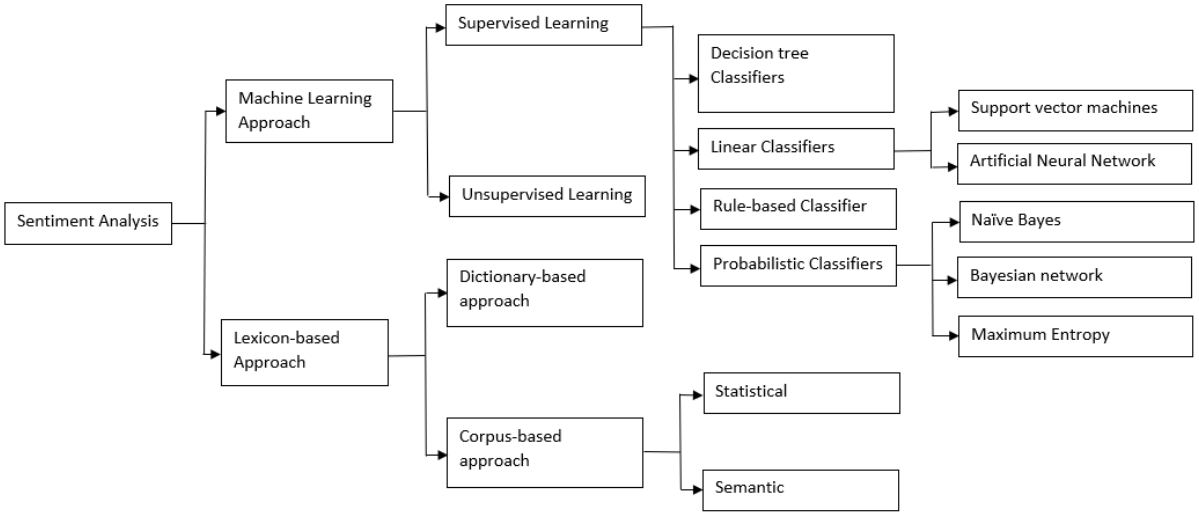

Figure 2. Sentiment Analysis methods

4.1 Machine Learning Methods

Machine learning is a crucial session of artificial intelligence that enables the investigation of novel algorithms to develop models. It can be signified as the procedure that stimulates or inspires the human brain functions to finalize the intelligent computer output. Artificial Neural Networks (ANN) entangled with Support Vector Machine (SVM) is a more cooperated approach in machine learning technology[9]. Deep learning algorithms made several specific advances like high-level semantic logic for instance machine translation, image analysis and classification and the most wisely the visual content analysis. The cross-modality consistent regression (CCR) schemes are recalling the difficulties in text and visual content analysis[54]. This approach can be described as a regression model which enables deep level text and visual analysis. Machine Learning methods are dividing into two major sections as supervised learning and unsupervised learning and there is another section called Lexicon based method which is a method of sentiment analysis but not falling under machine learning techniques[3,6].

4.1.1 Supervised Learning

Mainly, supervised learning is following two stages, the first one is "train the model" and the second one is "prediction". The "classification" and "regression" are the two types of supervised learning. During the "classification" conclusions are making to label and define the dataset according to the recognized entities. In "Regression" the correlation between the dependent and independent variables will be determined.

As the first step, the training data set with the labels are feeding to the classifying algorithm. As the next step, the testing data that didn't stream to the model previously, are feeding to the model to predict the relevant category. Testing data is a sample of data that utilizing to generate unbiased evaluation about the last model to validate its functioning and performance. The model can be upgraded and adjust until it gains better accuracy andperformance by implementing and testing various methods of sentiment analysis [13,15].

4.1.1.1 Decision Tree Classifiers

During this classifier, the data is broken down according to a definite parameter and it is similar to flowchart structure. It is utilized for the classification of data as well as for regression tasks. The training data is split hierarchically and this is more helpful for data segregation. Every node of the tree is capable of testing the feature of the dataset. The branches of the tree denote the conjunctions of the features. Each node consists of a question or function which is required to classify the data into relevant groups. The leaf node denotes the end classes of information[6]. Decision trees is can be utilized as a predictive model [19]. The prediction depends on the presence and absence of single or many terms and continuously running the process of searching for the term until finding the pure terms.

4.1.1.2 Linear Classifier

Linear classification is consisting of two classifiers a Support vector machine (SVM) and Artificial Neural Network (ANN)[24]. This method is concentrating on classifying the labelled data into separate classes based on their extracted features. The basic benefits of these linear classifiers are simplicity and user supportive computational ability.

4.1.1.2.1 Support Vector Machine (SVM)

SVM is more productive when classifying the textual contents, during research necessities, application domains and it is more applicable for supervised classifications and resolving the regression complications[25]. It classifies the classes after recognizing the best hyperplane among the classes. Hyperplanes can be described as the decision boundaries which enables to the categorization of the data points. Further, SVM is more supportive for non-linear regression. SVM consists of various kernel functions that are more supportive to generate a better model by recognizing hyperplane. These kernel functions can be mentioned as rbf, Poly, Sigmoid, and Linear. Linear Support Vector machines are more effective and supportive in penalties and losses. Further, a Linear Support Vector machine is capable of processing a huge amount of data[26]. SVM is used to analyze Twitter data about the weather[27] and during research about sentiment analysis of Twitter data related to global warming [28]. Both pieces of research showed the highest accuracy and performance than other methods when analyzing data. The classification of data using SVM required higher computer memory for instance, during the research of analyzing 24,335 tweets was not capable with the use of 8GB RAM[27].

4.1.1.2.2 Artificial Neural Networks(ANN)

An artificial Neural Network (ANN) is similar to the neural structure of the human brain. ANN is a combination of three layers namely input, hidden and output[24]. We can use ANN for regression attributes as well as for classification purposes. During the input

layer, the raw data is fed to the network. The input layer is obtains the values of explanatory attributes from each extracted observation. Typically, the quantity of input nodes is equivalent to the explanatory variables. The input layer is connects with one or more hidden layers by recognizing the patterns of the data feeds. The input layer is not capable of making changes in data, but this layer is able to duplicate the single value of the input to more outputs and direct them to the hidden layer. In the hidden layer, the values (incoming arcs) from the input layer are transforming using the interconnected nodes in the network. The incoming arcs are directing to hidden nodes and multiplying by weights, and a cluster of predetermined numbers are loading. Then the weighted inputs are totaling to generate a single number[29]. During the output layer, links from the hidden layer are collected and giving the output values which are capable of providing predictions. When the requirement is a classification, there is a possibility of consisting only one output layer. Appropriate selection of weights is the secret behind successful data manipulation. An advantage of ANN is better performance with both linear and non-linear data[29]. We can use ANN for image analysis and classification as well as for video content analysis.

Sentiment Image Classification by Convolutional Neural Networks

There are several specific methods to classify the images. Convolutional Neural Networks(CNN) are mostly used to analyze visual imagery. The visual sentiment analysis is grabbing the images and videos, extracting the features, classification of the embedded emotion patterns with the use of the appropriate convolutional neural network(CNN)[10]. CNN's are utilizing different multi-layer perception designs to preprocess the data by minimizing the workload. These CNN are also named as shift invariant or space invariant artificial neural networks (SIANN) which enable the recognition of image characters[11]. CNN is a class of artificial neural networks and it is a mathematical construct which is consisting of three blocks namely convolution,

pooling, and fully connected[30]. The convolution and pooling layers are designed to feature extraction while the fully connected layer is assigned to provide the output after classifying the extracted data[10]. The pooling layer is consisting another four categories like max pooling, average pooling, global max pooling and global average pooling[31]. In the max-pooling phase, the maximum occurring value and the average pooling stores the mostly storing average value. The fully connected layer is transforming the high level featured data from the pooling layer to more meaningful low dimensional vectors which is useful to provide probability[30]. There are several commonly used convolutional neural networks like ResNet, LeNeT-5, VGG, AlexNet. When using the highly productive and predictive CNN based classifying models, the training phase of data sets required less time. Even though CNNs are more productive, one main problem of using CNN for image classification is recognizing the correct learning parameters while specifying the appropriate network topology to gain successive performance. When choosing the best hyper-parameter, it is required to have deep knowledge about machine learning algorithms. The Hyper-parameter optimization technique is using for the automation of selecting the most appropriate parameter[12].

4.1.1.3 Rule-Based Classifier:

This is a classifier that uses "if..else" rules to decide the class in classification. The space of the data is modeled with the use of the rules[4, 24]. The status or the condition of the feature is exhibited in the disjunctive normal form in the left sideways and the right sideways is exhibiting the class label. The availability of terms is considered in the condition of formulation. The presence of relevant terms in extracted data is counted here. The formulation of rules is based on diverse criteria and rules depends on the selected criteria during training the data. Confidence and support are the usual criteria. The confidence is representing the right sideways of the rule which comprises the conditional probability while satisfactory is denotes by the left side[4].

4.1.1.4 Probabilistic Classifier:

Probabilistic classifiers can be denoted as a combination of classification models. This combination comprises three main classifiers namely, Naive Bayes (NB), Bayesian Network (BY) and Maximum Entropy (ME). This classifier is capable of providing a probability distribution regarding a cluster of classes as an output.

4.1.1.4.1 Naïve Bayes

This is utilizes textual content classification and when it comes to Twitter, the textual Twitter data can be classified using this method. This method is formulated around the Bayes theorem through an assumption of independence between predictors. This method assumes the availability of a specific feature is not related to the availability of any other feature. This is considered the smoothest and regularly processing classifier. This method estimating the probability of input text files which are considered as cooperative significant terms and classifying them into the classes [4]. This can be applicable in huge data amounts. This method is processing by three steps as the conversion of data set to a frequency table, generate a likelihood table by recognizing probabilities, use the Naïve Bayes equation to calculate the probability[32].

$$P(C = c|D = d) = \frac{P(D = d|C = c)P(C = c)}{P(D = d)} \quad [33]$$

In this equation[33], D represents the document, C represents the category(label), d and c are instances of D and C.

4.1.1.4.2 Bayesian Network

Bayesian Network cab be denoted as a probabilistic graphical model where the nodes signify the random variables while edges signify the conditional dependencies through a

Directed Acyclic Graph (DAG) [24]. Bayesian Networks are more appropriate for predicting the likelihood of numerous unknown causes affecting contributing factors. BN are responsible to discover the factors and the connections among them. As a result, the overall probability distribution of each element can be discovered. Efficient algorithms are capable of better performance with BN and the models that include a sequence of variables are known as dynamic Bayesian networks. But these networks are more expensive during text mining due to the computational complications.

4.1.1.4.3 Maximum Entropy

The basic impression of ME is the selection of the most static probabilistic model which consists of maximum entropy. This method is not assuming that the features are independent of each other and the bigram feature can be added not considering the feature overlapping[34]. This is capable of classifying the labelled feature collections into vectors. ME is utilizing to discover the appropriate label for feature collections after calculating the weights of features respectively, and the joined results. The Twitter data can be classified using this method as well as with the use of unigrams, bigrams and joining them together. The model can be denoted by the equation as

$$P_{ME}(c|d,\lambda) = P\frac{\exp[\sum_i \lambda_i f_i(c,d)]}{\sum_{c'} \exp[\sum_i \lambda_i f_i(c,d)]} \quad [34]$$

In this equation[34], c represents the class, d represents the tweet while λ denotes the weight vector.

4.1.2 Unsupervised Learning

During this method, clustering is a significant factor. This method is utilizing during the reliability of labelled data is less[13]. The gathering of unlabeled data is much easier than gathering labelled data. The sentences are classified according to the keywords of the

category. Comparison is used to categorize the data. When considering the Twitter data, the tweets can be compared according to the components (word lexicon) and clustered into positive and negative categories using this unsupervised method[4]. During the social media image analysis, the unsupervised sentiment analysis is comparatively much challenging than supervised methods [35].

4.2 Lexicon-Based Approach

This approach is assigned to determine the polarity of opinion words. The opinion words are categorizing into two parts as positive and negative where the positive opinion words are utilizing to assert the essential things while the negative opinion words are utilizing to assert unessential things [36]. It needs a sentiment lexicon to create an approach or it is possible to create it by partial-automation or manually. The manual approaches like general inquirer, opinion lexicon are consuming more time to practice and they are required to combine with automated approaches when the final results are needed to be more accurate by reducing the mistakes[24]. There are two sub-classifications as dictionary-based approach and the corpus-based approach.

4.2.1 Dictionary-Based Approach

This approach is suitable for a minor set of pinion words that are gained from the identified positionings. Bootstrapping the small set of opinion words is the basic concept of this method (e.g.WordNet)[36]. This set is expanding by exploring the synonyms and antonyms within the identified thesaurus or lexicon. The process of searching will iterate until no novel word is detected. After this process, the manual inspection is processing to eliminate the errors or correct the errors[24].

4.2.2 Corpus-Based Approach:

This approach is useful to detect the opinion words which consists of definite orientation complications. The solutions for these problems depend on the pattern that arises with a list of origin of opinion words[24]. This method is not very successful and operative as a dictionary-based approach due to the reason of difficulty in creating a lexicon that covers all vocabulary in the English language. Although this is not very effective, it may assist to discover the domain and content specific opinion words which can be determined as a benefit of this approach. This approach can be accomplished by two main sub approaches namely statistical and semantic.

4.2.2.1 Statistical Approach

This approach is assigned to detect the co-occurrence of the pattern. The polarities of existing adjectives within the lexicon is considered by this method. The set of indexed documents on the web is using as the lexicon in order to complete this process[24]. This approach is capable of managing the unavailability of some words due to the lack of space within the lexicon. The statistical approaches required a vast amount of training data.

4.2.2.2 Semantic Approach

This approach promptly providing the semantic values and is based on the diverse principles for computing the equalization among the words. WordNet is an example of this approach[24]. Providing the alike sentiment values to nearly semantic words while computing the polarity is one major duty in this approach.

4.3 Hybrid Approach

The hybrid approach is a phase that uses several methods for data classification. The hybrid approach of analyzing sentiments consisting the statistical and knowledge-based methods to recognize the polarity[37]. Most novel researchers are using a hybrid

approach to sentiment analysis due to the reason that this method can enhance the accuracy of the sentiment analyzing model[38,39,40].

4.4 Recursive Neural Network (RNN)

Recurrent Neural Network is using sequence data as the input and it is generated by applying a similar set of weights continuously or recursively on that sequence of input data[38]. This is helps to analyze the deep structured information and it is integrally a complex network. These are consist of different architecture layers which enable the operation of input data and it can be denoted as a tree-like hierarchical structure consisting of linked nodes through a chain. Twitter sentiment analysis by RNN models enabling the analysis of complex compositions and classify them into positive, negative and neutral groups.

3.3 Long-Short-Term-Memory (LSTM)

LSTM is sort of Recursive Neural Network and capable of classifying, processing and analyzing only one data point such as picture as well as a series of data, for instance video clips. This method consists of a higher capability of detecting the long-term dependencies over RNN model[40].The default behavior of LSTM is the ability to memorize information for a long period. LSTM is capable of selectively recall or forget the data.

Tools and Techniques in Real-time Twitter Data Analysis Algorithm

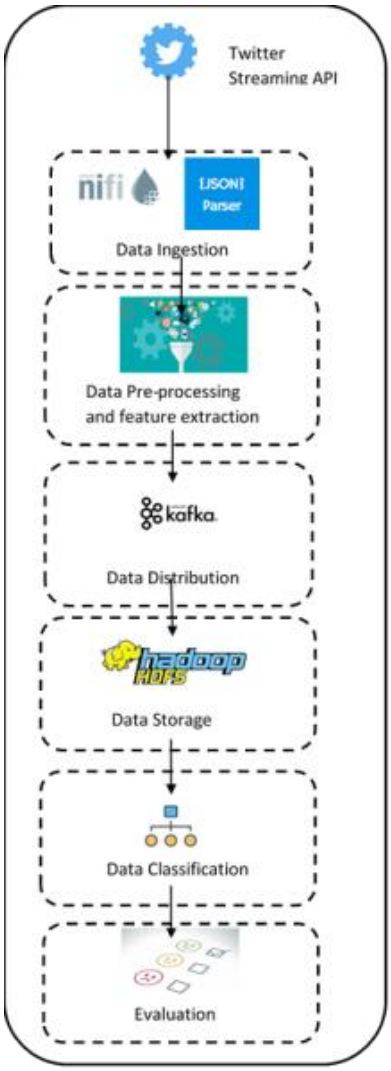

Figure 3. Tools and Techniques in Real-time Twitter Data AnalysisAlgorithm

Figure 3 demonstrates a proposed algorithm to analyze Twitter real-time data. Though the main steps of the sentiment analysis are depicted in figure 1, the model required specific tools and techniques to accomplish the workload. This figure depicts the tools and techniques that can be used for each step. Real-time Twitter data feeding, data ingestion, data pre-processing, data distribution, data storage, and data classification are the main approaches of this algorithm.

Data Ingestion

Data ingestion is a process of streaming the data to the model for pre-processing. There are several open-source ingestion tools to accomplish this process. The default behavior of these tools is extracting the data and in addition to that, manipulating the data and organizing the data are costly proceedings of them. These tools can be used to stream real-time data from sources. As the open-source ingestion tools Apache Nifi, NSQ, Gobblin, Amazon Kineses, MOTT, RabbitMQ, ZeroMQ[7] etc can be mentioned. The real-time Twitter data ingestion process can be accomplished using Apache NiFi[7] which is more supportive in routing, transformation and automate the flow of data to the model. JSON(JavaScript Object Notation) parser connected with Apache NiFi may more supportive to extract the Twitter data including the geo-location (coordinate) details. Unless using a new streaming method, the open-source ingestion tools can be used for model developments. When streaming Twitter real-time data, the Twitter Streaming API is more supportive.

Location-based Twitter Data Extraction

JSON parser enables to extract the real-time Twitter data with geo-location details, more specifically the coordinates (Longitude and latitude). There are two classes of location-based data as tweet tagged location or tweet location data and account location data[42]. Both two classes of location data can be extracted by JSON parser. The streamed data by Twitter Streaming API is JSON encoded.

Data Pre-Processing and Feature Extraction

The Term Presence Vs Term Frequency, N-gram Features, Parts of Speech(POS), Term Position, Negation[12], SentiWordNe[43] are several methods that can be used for Twitter data pre-processing and feature extraction.

Data Distribution

Data Distribution for the data storage can be accomplished using Apache Kafka platform. Kafka can be denoted as a framework implementation or streaming platform which assist

in stream monitoring, sequence distribution of data etc and written using Java and Scala [41]. It is an open-source platform which can be used to pipeline, manage and queue the extracted data.

Data Storage

The processed Twitter data is stored for application of the proposed model and open-source databases like Apache Cassandra, Hadoop Distributed File System(HDFS) can be used for this purpose. Apache Cassandra is a distributed NoSQL database and it enables to store the of data types like maps, user-defined data types, functions, aggregations and allows the storage large volume of data[44]. HDFS is a data storing file system by breaking the data into minor chunks[45].

Classification and Evaluation

The Machine Learning based classifications are using numerous methods for data classification like SVM, NB, ME[46] etc. and as the novel methods RNN, CNN, LSTM, GRU[39,47,48]can be mentioned. The analyzed results can be evaluated and test the accuracy. After testing the accuracy and performance of one or combined sentiment analysis methods, the most accurate and well-performing algorithm can be finalized as the best algorithm for analyzing real-time location-based Twitter data.

Table 1. Comparison of data analysis algorithms used in recent researches

Research Title and Year of Publication	**Dataset**	**Used Techniques**	**Advancements**	**Accuracy**
Sentiment Analysis Using Deep Learning Approach[38]	IMDB movie reviews	Recurrent Neural Network (RNN), Convolutional Neural Networks	Checking for more accurate analysis results for IMDB text out of RNN, CNN methods and compare with the former published SVM[49] and RNTN[48]	RNN - 68.64% CNN - 88.22%

		(CNN)	methods	
Building a Twitter Sentiment Analysis System with Recurrent Neural Networks[39]	Integrum annotated 1,578,627 tweets	Experimental RNN Architectures with LSTM and GRU layers	Designing the RNN Model with many advanced layers to analyse big amount of data	LSTM- 80.39% GRU - 80.74%
CNN for situations understanding based on sentiment analysis of Twitter data[47]	Movie reviews with one sentence per review(MR) and essential collection of the real Twitter dataset (STS Gold)	CNN Model	A Convolutional neural network model to examine the accuracy and performance compared with the traditional methods	MR - 74.5% STS Gold - 75.39%
Deep-Sentiment: Sentiment Analysis Using Ensemble of CNN and Bi-LSTM Models[40]	IMDB reviews with 50,000 reviews and SST2 Dataset that contained binary sentiment analysis	CNN, LSTM and Bi-LSTM models	Comparing each separate LSTM and CNN models and their combination	LSTM - 80% CNN – 80.2% LSTM+CNN - 80.5%
Multimodal Sentiment Analysis to explore the structure of emotions[50]	A dataset consists 1,009,534 posts on Tumblr		LSTM for textual content Inception model for images compare them with a multi-modal neural network (LSTM+Inception)	Inception Model - 36% LSTM – 69% Multi-modal neural network - 72%
Recursive Deep Models for Semantic Compositionality Over a Sentiment Tree bank[48]	Dataset of Stanford Sentiment Treebank	Deep Analysis models like NB, SVM, BiNB ,Vec Avg, RNN , MV-RNN, RNTN	Comparison of the deep models in sentiment analysis (NB, SVM, BiNB ,VecAvg, RNN , MV-RNN, RNTN)	NB – 81.8% SVM – 79.4% BiNB – 83.1% VecAvg – 80.1% RNN – 82.4% MV-RNN – 82.9%

				RNTN – 85.4%
Detection and classification of social media-based extremist affiliations using sentiment analysis techniques[51]	Dataset of 20,000 tweets produced from revolutionist user accounts	CNN+LSTM+GRU	Testing different combinations and GRU, CNN and LSTM to gain more accurate results	LSTM + CNN – 93%
Halal Products on Twitter: Data Extraction and Sentiment Analysis Using Stack of Deep Learning Algorithms[52]	Dataset of Tweets over 10 years on halal tourism and halal cosmetics	CNN, LSTM, RNN	Testing and the CNN,LSTM and RNN models and their combination to gain the most accurate model	CNN+LSTM - 93.78%.
An image-text consistency driven multimodal sentiment analysis approach for social media[53]	A set of social media pictures and textual content	CNN, SVM	Testing the data set with existing models and proposed model	CNN+SVM – 87%

CNN - Convolutional Neural Network

NB – NaiveBayes

SVM - Support vector machine

LSTM - Long Short-Term Memory

GRU - Gated Recurrent Unit

BiNB – bi-gramNB

VecAvg - Vectors and ignores word order

RNN – Recursive Neural Network

RNTN – Recursive Neural Tensor Network

MV-RNN - Matrix-Vector RNN

Conclusion

There are several varieties of Twitter data for instance pictures, textual contents, videos, etc and according to the data type and the final required result, the sentiment analysis method should be selected. The elementary knowledge about sentiment analysis methods is mentioned in the review paper with the highlights of real-time data analysis methods. The techniques and tools of real-time Twitter data analysis based on geo-location data are demonstrated using a simple model. The methods of data collection, data pre-processing, feature extraction, and methods of sentiment data are hierarchically discussed in this paper. The numerous models and techniques used by several researchers are depicted comparatively with mentioning the accuracy of each method. Clear understanding and proper practice of appropriate sentiment analysis methods may lead to higher accuracy and performance in final results. This paper depicts the basic theory of several sentiment analysis methods including machine learning techniques. The objective of this paper is to provide a proper conceptual understanding of the sentiment analysis techniques. In future work, the development of a novel machine-learning algorithm to analyze real-time Twitter data will be conducted.

Acknowledgement

I would like to convey my gratitude to Dr. KPN Jayasena and Dr. Iraj Ratnayake who advised and guided me on this review paper.

References

[1] H. Tankovska, "Most popular social networks worldwide as of January 2021, ranked by number of active users," 2021. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed Apr. 10, 2021).

[2] G. D. Devi and S. Kamalakkannan, "Literature Review on Sentiment Analysis in Social Media: Open Challenges toward Applications," Test Eng. Manag., vol. 83, no. 7, pp. 2466–2474, 2020.

[3] F. Neri, C. Aliprandi, F. Capeci, M. Cuadros, and T. By, "Sentiment analysis on social media," Proc. 2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2012, no. August, pp. 919–926, 2012, doi: 10.1109/ASONAM.2012.164.

[4] R. K. Dey, D. Sarddar, I. Sarkar, R. Bose, and S. Roy, "A Literature Survey On Sentiment Analysis Techniques Involving Social Media And Online Platforms," vol. 9, no. 05, 2020.

[5] S. Mukherjee and P. Bhattacharyya, "Sentiment Analysis : A Literature Survey," ArXiv, vol.

abs/1304.4, no. April, pp. 1–52, 2013.
[6] K. F. M. Panguila and J. Chandra, "Sentiment analysis on social media data using intelligent techniques," Int. J. Eng. Res. Technol., vol. 12, no. 3, pp. 440–445, 2019.
[7] S. Ge, H. Isah, F. Zulkernine and S. Khan, "A Scalable Framework for Multilevel Streaming Data Analytics using Deep Learning," 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), 2019, pp. 189-194,doi: 10.1109/COMPSAC.2019.10205.
[8] O. Almatrafi, S. Parack, and B. Chavan, "Application of Location-Based Sentiment Analysis Using Twitter for Identifying Trends Towards Indian General Elections 2014," 2015.
[9] M. Alghalibi, A. Al-Azzawi, and K. Lawonn, "Deep Attention Learning Mechanisms for Social Media Sentiment Image Revelation," Int. J. Comput. Commun. Eng., vol. 9, no. 1, pp. 1–17, 2020, doi: 10.17706/ijcce.2020.9.1.1-17.
[10] V. Aggarwal and G. Kaur, "A review:deep learning technique for image classification," Accent. Trans. Image Process. Comput. Vis., vol. 4, no. 11, pp. 21–25, 2018, doi: 10.19101/tipcv.2018.411003.
[11] M. K. Singh, P. G. S. Baluja, and D. P. Sahu, "Understanding the Convolutional Neural Network &it ' s Research Aspects in Deep Learning," vol. 5, no. Vi, pp. 867–871, 2017.
[12] V. Podgorelec, Š. Pečnik, and G. Vrbančič, "Classification of similar sports images using convolutional neural network with hyper-parameter optimization," Appl. Sci., vol. 10, no. 23, pp. 1–24, 2020, doi: 10.3390/app10238494.
[13] D. Sharma, M. Sabharwal, V. Goyal, and M. Vij, "Sentiment analysis techniques for social media data: A review," Adv. Intell. Syst. Comput., vol. 1045, no. September, pp. 75–90, 2020, doi: 10.1007/978-981-15-0029-9_7.
[14] G. Murray and G. Carenini, "Subjectivity detection in spoken and written conversations," Nat. Lang. Eng., vol. 17, no. 3, pp. 397–418, 2011, doi: 10.1017/S1351324910000264.
[15] P. Tyagi, S. Chakraborty, R. C. Tripathi, and T. Choudhury, "Literature Review of Sentiment Analysis Techniques for Microblogging Site," SSRN Electron. J., 2019, doi: 10.2139/ssrn.3403968.
[16] C. C. S. E. Khatib, D. C. S. E. Kamble, A. P. M. T.-K. M, B. R. C. C. S. E, and G. N. S. C. S. E, "Social Media Data Mining For Sentiment Analysis," pp. 373–376, 2016.
[17] W. Hua, Z. Wang, H. Wang, K. Zheng, and X. Zhou, "Understand Short Texts by Harvesting and Analyzing Semantic Knowledge," IEEE Trans. Knowl. Data Eng., vol. 29, no. 3, pp. 499–512, 2017, doi: 10.1109/TKDE.2016.2571687.
[18] Z. N. Gastelum and K. M. Whattam, "State-of-the-Art of Social Media Analytics Research," Pacific Northwest National Laboratory (U.S.), Richland, WA (United States), Jan. 2013. doi: 10.2172/1077994.
[19] Y. Y. Desai, To Study The Social Media Sentimental Analysis Using Facebook As Platform, no. 10. Master of Philosophy in Business Management, D. Y. Patil University, Navi Mumbai, 2017.
[20] Y. Mejova and P. Srinivasan, "Exploring Feature Definition and Selection for Sentiment Classifiers," in Association for the Advancement of Artificial Intelligence (www.aaai.org), 2011, no. January, pp. 546–549.
[21] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," Proc. 12th Int. Conf. World Wide Web, WWW 2003, no. March, pp. 519–528, 2003, doi: 10.1145/775152.775226.
[22] A. Giachanou and F. Crestani, "Like it or not: A survey of Twitter sentiment analysis methods," ACM Comput. Surv., vol. 49, no. 2, 2016, doi: 10.1145/2938640.
[23] G. Magesh and P. Swarnalatha, "Analyzing customer sentiments using machine learning techniques," Int. J. Civ. Eng. Technol., vol. 8, no. 10, pp. 1829–1842, 2017.
[24] N. A. S. Abdullah, N. I. Shaari, and A. R. A. Rahman, "Review on sentiment analysis approaches for social media data," Journal of Engineering and Applied Sciences, vol. 12, no. 3. pp. 462–467, 2017, doi: 10.3923/jeasci.2017.462.467.
[25] S. R. C. J. A. Gualtieri R. F. Cromp, and L. F. Johnson, "Support vector machine classifiers as applied

to AVIRIS data," JPL Airborne Earth Sci. Work., no. November 1999, pp. 217–227, 1999.
[26] S. Ray, "Understanding Support Vector Machine Algorithm from Examples." https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ (accessed Apr. 18, 2021).
[27] S. Siddharth, R. Darsini, and M. Sujithra, "Sentiment Analysis on twitter data using Machine Learning," J. Xidian Univ., vol. 14, no. 12, 2020, doi: 10.37896/jxu14.12/039.
[28] N. Mucha, Sentiment Analysis of Global Warming Using Twitter Data. Master of Computer Science, North Dakota State University,United States, December,2018.
[29] S. Sharma, "Artificial Neural Network (ANN) in Machine Learning," 2017. https://www.datasciencecentral.com/profiles/blogs/artificial-neural-network-ann-in-machine-learning (accessed Apr. 19, 2021).
[30] R. Yamashita, M. Nishio, R. Kinh, G. Do, and K. Togashi, "Convolutional neural networks : an overview and application in radiology," Springer, vol. 9, pp. 611–629, 2018, doi: https://doi.org/10.1007/s13244-018-0639-9.
[31] A. O. Tarasenko, Y. V. Yakimov, and V. N. Soloviev, "Convolutional neural networks for image classification," CEUR Workshop Proc., vol. 2546, pp. 101–114, 2019.
[32] S. Ray, "6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R," 2017. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ (accessed Apr. 21, 2021).
[33] D. Rossiter and L. Marc, Project Report Twitter Emotion Analysis. Hong Kong: MSc Information Technology,Hong Kong University of Science and Technology, Hong Kong, July, 2015.
[34] S. Joshi and D. Deshpande, "Twitter sentiment analysis system," arXiv, vol. 180, no. 47, pp. 35–39, 2018, doi: 10.5120/ijca2018917319.
[35] Y. Wang and B. Li, "Sentiment Analysis for Social Media Images," Proc. - 15th IEEE Int. Conf. Data Min. Work. ICDMW 2015, pp. 1584–1591, 2016, doi: 10.1109/ICDMW.2015.142.
[36] K. Z. Aung, "Sentiment Analysis of Students ' Comment Using Lexicon Based Approach," in 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), 2017, pp. 149–154.
[37] I. Gupta and N. Joshi, "Enhanced twitter sentiment analysis using hybrid approach and by accounting local contextual semantic," J. Intell. Syst., vol. 29, no. 1, pp. 1611–1625, 2020, doi: 10.1515/jisys-2019-0106.
[38] P. Cen, K. Zhang, and D. Zheng, "Sentiment Analysis Using Deep Learning Approach," J. Artif. Intell., vol. 2, no. 1, pp. 17–27, 2020, doi: 10.32604/jai.2020.010132.
[39] D. B. Oprean, "Building a Twitter Sentiment Analysis System with Recurrent Neural Networks," Sensors 2021, pp. 1–24, 2021, doi: https://doi.org/10.3390/s21072266.
[40] S. Minaee, E. Azimi, and A. Abdolrashidi, "Deep-Sentiment : Sentiment Analysis Using Ensemble of CNN and Bi-LSTM Models," arXiv, 2019, doi: 1904.04206v1 [cs.CL].
[41] "APACHE KAFKA." https://kafka.apache.org/ (accessed May 14, 2021).
[42] "Filtering Tweets by location." https://developer.twitter.com/en/docs/tutorials/filtering-tweets-by-location (accessed May 13, 2021).
[43] M. Karanasou, A. Ampla, C. Doulkeridis, and M. Halkidi, "Scalable and Real-Time Sentiment Analysis of Twitter Data," IEEE Int. Conf. Data Min. Work. ICDMW, vol. 0, no. Us, pp. 944–951, 2016, doi: 10.1109/ICDMW.2016.0138.
[44] "Apache Cassandra." https://cassandra.apache.org/doc/latest/architecture/overview.html (accessed May 13, 2021).
[45] "Hadoop Distributed File System (HDFS) Architecture – A Guide to HDFS for Every Data Engineer." https://www.analyticsvidhya.com/blog/2020/10/hadoop-distributed-file-system-hdfs-architecture-a-guide-to-hdfs-for-every-data-engineer/ (accessed May 13, 2021).
[46] M. Paolanti et al., "Tourism destination management using sentiment analysis and geo - location information : a deep learning approach," Inf. Technol. Tour., no. 0123456789, 2021, doi: 10.1007/s40558-021-

00196-4.
[47] S. Liao, J. Wang, R. Yu, K. Sato, and Z. Cheng, "CNN for situations understanding based on sentiment analysis of twitter data," Procedia Comput. Sci., vol. 111, no. 2015, pp. 376–381, 2017, doi: 10.1016/j.procs.2017.06.037.
[48] R. Socher et al., "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in Conference on Empirical Methods in Natural Language Processing, 2013, no. October, pp. 1631–1642.
[49] K. Saranya and S. Jayanthy, "Thumbs up? sentiment classification using machine learning techniques," Proc. 2017 Int. Conf. Innov. Information, Embed. Commun. Syst. ICIIECS 2017, vol. 2018-Janua, pp. 1–5, 2018, doi: 10.1109/ICIIECS.2017.8276047.
[50] A. Hu and S. Flaxman, "Multimodal Sentiment Analysis To Explore the Structure of Emotions," Appl. Data Sci. Track Pap., pp. 350–358, 2018, doi: https://doi.org/10.1145/3219819.3219853.
[51] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and I. Awan, "Detection and classification of social media-based extremist affiliations using sentiment analysis techniques," Human-centric Comput. Inf. Sci., vol. 9, no. 1, 2019, doi: 10.1186/s13673-019-0185-6.
[52] A. Feizollah, S. Ainin, N. B. Anuar, N. A. B. Abdullah, and M. Hazim, "Halal Products on Twitter: Data Extraction and Sentiment Analysis Using Stack of Deep Learning Algorithms," IEEE Access, vol. 7, no. June, pp. 83354–83362, 2019, doi: 10.1109/ACCESS.2019.2923275.
[53] Z. Zhao, H. Zhu, Z. Xue, Z. Liu, and J. Tian, "An image-text consistency driven multimodal sentiment analysis approach for social media," Inf. Process. Manag., vol. 56, no. 6, p. 102097, 2019, doi: 10.1016/j.ipm.2019.102097.
[54] You, Q., Luo, J., Jin, H., & Yang, J. ,"Cross-modality Consistent Regression for Joint Visual-Textual Sentiment Analysis of Social Multimedia", Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, 2016.