Chapter 8
Sentiment Analysis of Twitter Data Through Machine Learning Techniques

Asdrúbal López-Chau, David Valle-Cruz and Rodrigo Sandoval-Almazán

Abstract Cloud computing is a revolutionary technology for businesses, governments, and citizens. Some examples of Software-as-a-Services (SaaS) of cloud computing are banking apps, e-mail, blog, online news, and social networks. In this chapter, we analyze data sets generated by trending topics on Twitter that emerged from Mexican citizens that interacted during the earthquake of September 19, 2017, using sentiment analysis and supervised learning, based on the Ekman's six emotional model. We built three classifiers to determine the emotions of tweets that belong to the same topic. The classifiers with the best accuracy for predicting emotions were Naive Bayes and support vector machine. We found that the most frequent predicted emotions were happiness, anger, and sadness; also, that 6.5% of predicted tweets were irrelevant. We provide some recommendations about the use of machine learning techniques in sentiment analysis. Our contribution is the expansion of the emotions range, from three (negative, neutral, positive) to six in order to provide more elements to understand how users interact with social media platforms. Future research will include validation of the method with different data sets and emotions, and the addition of new artificial intelligence techniques to improve accuracy.

Keywords Cloud computing · Sentiment analysis · Machine learning · ML · Twitter · Naive Bayes · Ekman's model

A. López-Chau
Autonomous University of the State of Mexico, 55600 Valle Hermoso, Zumpango, Estado de México, Mexico
e-mail: alchau@uaemex.mx

D. Valle-Cruz (✉) · R. Sandoval-Almazán
Autonomous University of the State of Mexico, Instituto Literario # 100, Toluca, Mexico
e-mail: davacr@uaemex.mx

R. Sandoval-Almazán
e-mail: rsandovala@uaemex.mx

M. Ramachandran and Z. Mahmood (eds.), *Software Engineering in the Era of Cloud Computing*, Computer Communications and Networks,
https://doi.org/10.1007/978-3-030-33624-0_8

8.1 Introduction

Cloud computing is a revolutionary technology for businesses, governments, and citizens [1], providing different types of services based on consumer Internet services. It represents a model for enabling ubiquitous, convenience, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction [2]. There are three cloud computing delivery models: Software-as-a-Service (SaaS) where the consumer uses an application, but does not control the computer system; Platform-as-a-Services (PaaS), the consumer uses a hosting for his/her applications; and Infrastructure-as-a-Service (IaaS), where the consumer uses fundamental computing resources such as processing power.

Some examples of cloud computing are banking apps, e-mail, blog, online news, and social networks. Social networks represent an ideal platform for establishing relationships with customers or users, and for understanding the interaction between them, but in an unstructured form. One way to understand how users behave on social networks is through sentiment analysis.

Furthermore, e-governance uses information and communication technologies (ICT) to provide government services and information exchange for developing government to citizen interaction. However, the traditional e-governance solutions are incapable of fulfilling the current need because of its increasing demand, application complexity, infrastructure management, cost overhead, and other technical challenges. Emerging technologies such as cloud computing, big data, and machine learning could overcome these challenges using the modern approach for computing, storage, and data processing, because they provide unique features to e-governance: lower cost, scalability, easy management, disaster recovery, accountability, resource provisioning, distributed storage, data analytics, mobility, etc. [3].

Social networks represent a communication media useful to express feelings and thoughts, meet other people, do business, or speak about someone or something. Among these social networks, the most used ones are Facebook, Twitter, YouTube, and Instagram [4]. Each of these social networks offers different types of content generation on text, images, and video. These data nurture big data, which contains a large amount of information, valuable for decision making of political actors, businesses, and government organizations [5].

There are artificial intelligence techniques and tools for the automatic analysis of data generated in social media in the areas of machine learning, natural language processing, and text mining which allow finding patterns in data or relationships [6, 7]. Sentiment analysis of Twitter posts has attracted the attention of many scholars [8, 9] since the unstructured data generated every day, in this kind of social media, contain valuable information for decision making in organizations and how to react to certain types of events.

This chapter aims to present our experience in analyzing Twitter data sets with sentiment analysis techniques. Unlike the vast previous research, which only

considers three categories: negative, positive, or neutral to identify emotions, our proposal, using the Ekman's emotion model, classifies six types of emotions: joy or happiness, anger, sadness, surprise, disgust, and fear [10]. The traditional use of this model is the detection of emotions in facial expressions, but in this chapter, we used it to identify sentiment analysis from Twitter posts. As part of the framework, we provide some recommendations on how to deal with the problems found on implementing systems for sentiment analysis based on our experience analyzing several Mexican data sets in recent local elections, Mexico's 2017 earthquake, and the presidential campaign [11–13].

This chapter consists of five parts: the first section explains the introduction related to sentiment analysis on Twitter, as well as the purpose of the study. The second section consists of three parts: a literature review related to social networks, Ekman's model, and sentiment analysis literature. The third section describes our proposal of a method based on classification methods to predict sentiments using Ekman's model. The fourth section describes the results of classifiers systems. The final section describes conclusions, experiences, and future research.

8.2 Literature Review

This literature review section is divided into three parts: social networks; Ekman's model; and sentiment analysis. These are explored in the following sections.

8.2.1 Social Networks

Social networks changed communication and represent a technological tool, useful to disseminate any kind of information through the world. With the development of 5G, wireless and Internet connections, it will enable different kinds of new applications, more personalized, connected and interactive services become available with resource-limited mobile terminals [14], and it is easier to express feelings and thoughts and communicate ideas to other people. More and more data will be generated on Internet which represent ground gold for all kind of organizations. For this reason, it is important to study how to analyze big data through techniques such as sentiment analysis.

In this context, different data sources feed the big data every day. Devices such as wearables, smartphones, tablets, and personal computers allow access to programmed applications that maintain us immersed in a hyperconnected world. Internet of Things, sensors, data clouds, Google searches, Amazon and eBay shopping, and the use of social media are some examples that generate much information that cannot be analyzed with traditional tools.

In particular, social networks generate large amounts of data about each person, where they express their tastes, feelings, and moods; social networks have become a

sensor in real time. With the advent of social networks, users create records of their lives by daily posting details of activities they perform, events they attend or live, places they visit, pictures they take, and things they enjoy, want, and feel. Social network is a platform where millions of people interact, share opinions, and express their feelings.

The most used social networks worldwide are Facebook, YouTube, WhatsApp, Instagram, and Twitter. Twitter allows us to publish content in a microblogging format, quickly, compactly, and in real time. Differently, Facebook allows us to use six basic impressions (like, love, haha, wow, sad, and angry). Twitter has no mechanism to express impressions, but several researchers have developed studies on the sentiment analysis on Twitter, to understand what the users of this social media express.

The rise of social networks has generated today a tremendous interest among Internet users, organizations, and researchers. Data from these social networking sites can be used for several purposes, like prediction, marketing, or sentiment analysis [15]; some other researchers have developed personalized recommendation systems based on learning automata and sentiment analysis [16]. The millions of tweets received every day could be subjected to sentiment analysis, but handling such a huge amount of unstructured data is a tedious task to take up [15]. Since data generated in social networks are in an unstructured way, analytics solutions that mine structured and unstructured data are important as they can help organizations gain insights not only from their privately acquired data but also from large amounts of data publicly available on the Web [17].

The development of computational intelligence techniques enables these platforms to become modern large-scale laboratories in which the development of intelligent emotion-aware applications can be incubated to maximize the quality of computerized solutions [18]. The optimum solution for this kind of problem is analyzing the information available on social network platforms and performing sentiment analysis [19]. The study of social networks using sentiment analysis allows identifying patterns in large data sets.

8.2.2 Ekman's Model

People express emotions provoked by the events they live in, the environment in which they are immersed, their personality, and the experiences they have lived. An emotion is an alteration by a shock or impulse in the brain caused by impressions of senses, ideas, or memories. Emotions are shown through facial and body expressions, the tone of voice, among other characteristics. Nowadays, social networks invite us to express our emotions and feelings through texts, images, videos, or any multimedia element. In order to understand and classify emotions, different researchers and psychologists have provided answers to questions such as: How do we have emotions and what does it cause to have these emotions [20]? Some

answers are simplistic but very concise in the classification of emotions [21] and some others consider different factors to identify a wide variety of emotions [22].

Sreeja and Mahalakshmi [20] classify theories of emotions into three main categories: physiological: where the response within the human body is responsible for the generation of emotions; neurological: where brain action leads to emotional reactions; and cognitive: where thoughts and other mental activities form emotions [20]. Affective computing scientist has developed computational solutions to identify and react to user emotional states, in this sense, the representation of emotions is designed in two main ways, categorically: the generated emotion is selected from a set of emotions and labeled; dimensional: the representation of emotions is based on a set of quantitative measures using multidimensional scales.

In the categorical models, there are important representations such as: Ekman's [18] basic emotions and Navarasa models; on the other hand, in the dimensional models have been designed representations such as circumplex, Plutchik, Pad, and Thayer [20]. In categorical models, emotions are identified by a class label such as anger, disgust, fear, joy, sadness, and surprise; and they are easy to understand. In contrast, in dimensional models, it is necessary to quantitatively define the value of each emotion, in addition to defining combined emotional states of different numerical levels.

Several studies have carried out sentiment analysis with the help of a frame of reference or model of emotions [22–24]. These emotional models make it possible to identify different kinds and numbers of emotions. However, one of the most commonly used in the area of affective computing, intelligent agents, and sentiment analysis is the Ekman's model which is based on emotions generated in facial expressions, emotions that have been identified in humans and inherited from ancestral times.

Ekman classifies emotions into six types: anger, fear, disgust, surprise, joy, and sadness (see Fig. 8.1). These six emotions are basic and universal for facial expressions since they are defined as adaptations selected by biological mechanisms with evolutionary value and, in a general way, since when expressing any of the six emotions, the same facial features are found [21]. These emotions are presented in social media posts because social media users express themselves depending on the event they are living, and the event generates an impression or reaction that some people display on their social media.

The six emotions proposed by Ekman are classified into positive and negative, depending on the reaction expected, and the event they are living. Positive emotions are produced by reacting to pleasant events or people's liking, such as joy and surprise. Negative emotions are the result of an event that people do not like, such as anger, fear, sadness, and disgust. Humans present a combination of these emotions by reacting to events that happen in their daily lives. For example, if someone receives a birthday present that he or she has wanted for some time, the emotions of surprise and joy will be present, but if a person is frightened, he or she may show fear, anger, and disgust, but also surprise.

Although Ekman's model is used to identify emotions in people's expressions and for modeling intelligent agents, there is some research in the area of sentiment

Fig. 8.1 Basic emotions according to Paul Ekman

analysis that has adopted this model to classify emotions in texts generated in blogs, social media, and Web pages [25–27]. For the purpose of this research, we adapted the Ekman's model to classify emotions in large data sets generated on Twitter, during the earthquake of September 19, 2017, in Mexico.

8.2.3 Sentiment Analysis

Traditionally, sentiment analysis studies have been based on classifying or identifying the polarity of Twitter posts, classifying sentiment as positive, negative, or neutral, getting very good results in the precision of the data analysis, and predictions carried out. However, applying a classification based on a most varied number of emotions is a difficult task.

Sentiment analysis is a technique which involves natural language processing, text analysis, and data mining [28, 29]. The sentiment content of the text is characterized by using techniques such as natural language processing (NLP), statistics or any of the machine learning methods. Sentiment analysis can also be proceeded by based on rule-based classifier or supervised learning [30]. Sentiment analysis of

short texts and reviews available on different social networking sites is challenging because of the limited contextual information. Based on the sentiments and available opinions, developing a recommendation system is an interesting concept, which includes strategies that combine the small text content with prior knowledge [31].

The increase of smartphone and tablet applications allows users to interact on different service platforms at any time through mobile Internet, social media, cloud computing, and others. However, there are very few studies of classification methods applied to this area [28]. Nowadays, large volumes of data are in an unstructured manner, and it is very difficult to perform operations in unstructured data. So, the data need to be structured and organized before any analysis. The sentiment analysis technique is used to analyze the sentiments of a user based on text analysis [32].

The technology within text analytics comes from fundamental fields including linguistics, statistics, and machine learning. In general, modern text analytics uses statistical models, coupled with linguistics and emotional theories, to capture patterns in human languages in such a way that machines can understand the meaning of texts and perform various text analytics tasks. Text mining in the area of sentiment analysis helps organizations to uncover sentiments and improve their customer relationship management [33, 34]. This is useful to identify patterns on data, and for decision making.

Some of the techniques used to develop sentiment analysis on Twitter have been artificial neural networks, vector support machines, logistic regression models, Bayes' theorem, decision trees, and fuzzy logic. In this chapter, we analyze data sets generated by trending topics on Twitter that emerged from the Mexican citizens that interact during the earthquake of September 19, 2017, using sentiment analysis, supervised learning, and based on the Ekman's six emotional model.

8.3 Methodology

In this section, we introduce the proposed methodology to perform sentiment analysis of Twitter posts. This methodology is summarized in Fig. 8.2.

Each one of the steps presented in Fig. 8.2 is explained in the following subsections.

8.3.1 Data Collection

The large number of posts that are made at any time in social media, such as Twitter and Facebook, have attracted the attention of researchers to apply methods for sentiment analysis. Although there are public data sets to test and compare these methods (see, for example, https://www.kaggle.com/kazanova/sentiment140,

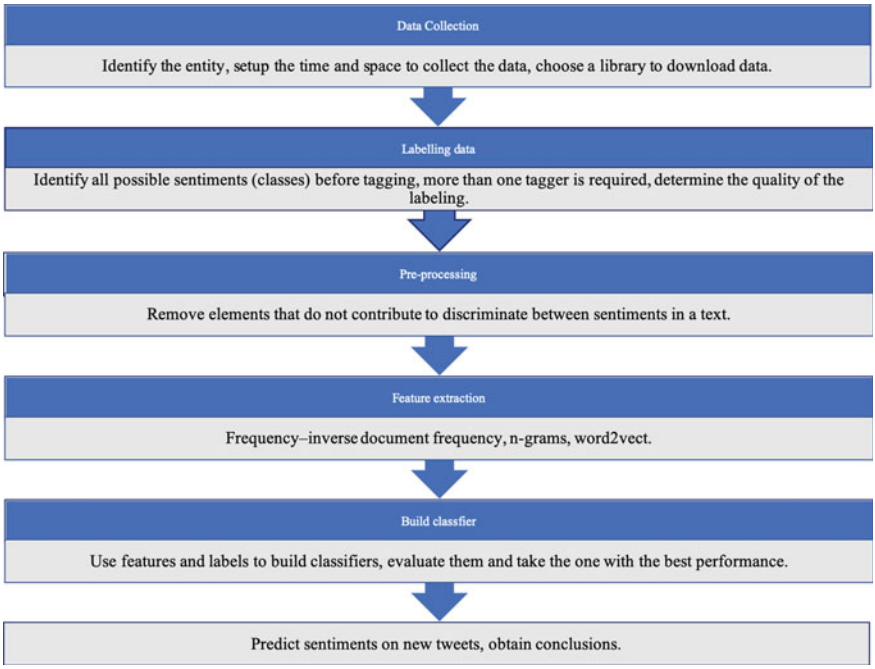

Fig. 8.2 Summary of the methodology

www.kaggle.com/eliasdabbas/5000-justdoit-tweets-dataset, and https://data.world/datasets/twitter), it is now possible to download data directly from social media to analyze data. To achieve the above goal, it is necessary to register as a developer in the social media from which you wish to download data and use special libraries. As part of the methodology, we recommend to take into account the following recommendations for this phase:

- Identify the entity. This can be an event, topic, service, institution, person, or thing that users of social media post about. These users use the symbol# before a keyword or a series of words to refer to a specific topic in a message. On the other hand, the symbol @ is usually used before the username to refer someone or something (a brand, for example). These two symbols can help to group publications related to an entity.
- Set up the time and space to collect data. Data downloaded from social media (in this case: microblogging sites) can correspond to an interval, or they can be collected in real time. According to the literature, the most common approach is to apply sentiment analyses methods on data of a period. Also, location can be important in some cases, for example, in the case of earthquakes, tsunamis, etc.
- Choose a library to download the data. Several alternatives are available, depending on the chosen programming language. For the R programming language, the packages twitterR or retweet are good options. For the Python

programming language, there are many libraries; some of them are the following: python-twitter, tweepy, TweetPony, Python Twitter Tools, Twitter Search, TwitterAPI, and Birdy. It is recommended to make sure that the coding (UTF-8 for example) is correct to avoid getting data corrupted or with incorrect symbols.

8.3.2 Labeling Data

One of the crucial steps to apply sentiment analysis successfully is the labeling of data. In order to label data, it is necessary that texts are read carefully and then assigned to each of the labels that correspond to a specific sentiment. Therefore, this step is one of the most difficult ones, as well as time-consuming.

Before starting the labeling, it is necessary that all the labels that will be used have been defined, including one for cases where it is not possible to accurately determine the sentiment in the analyzed text (for example, "ambiguous" label). Each tagger (it is recommended to have more than one) must first determine if the analyzed text is relevant. If so, the entity and the aspects or attributes of the entity in the text must be identified. Then, the labels (sentiments) that best represent the full text are assigned. It is important to point out that one tweet can have different sentiments.

Once the data have been labeled, it is necessary to carry out a type of evaluation to determine the quality of the labeling. One of the metrics is the Kappa index, which is used to know the degree of agreement between the taggers.

The number of labels of each type should be balanced to avoid a poor performance of machine learning methods. It has been suggested that there are around 3000 texts labeled.

8.3.3 Preprocessing of Texts

The texts post in social media can contain emoticons, numbers, exclamation and question marks, and other symbols. In some languages, such as Spanish, some words have accented vowels. It has been studied that most of these elements do not contribute to discriminate between sentiments in a text. Therefore, in the preprocessing step, these elements are removed or transformed from texts. Basic preprocessing for sentiment analysis includes the following tasks:

- Remove numbers, exclamation marks, question marks, and punctuation marks.
- Identify mentions (@UserName), these can be substituted by the keyword USERNAME.
- Identify topics (#topic), these can be substituted by the keyword TOPIC.
- Remove URLs, these begin with the string "http".

- Remove html tags, these begin and end with "<" and ">", respectively.
- Remove other symbols, such as $, %, ^, *.
- Change accented vowels by the same vowels without accents.
- Apply a stemming algorithm, such as snow-ball.
- Most of these previous tasks can be implemented using a programming language, and therefore done automatically.

A more detailed preprocessing can be applied to the texts, for example:

- Identify misspelled words and correct them using a dictionary. Identify words that contain extra letters [for example: "Fueraaaa" (get out), "Goool," (goal)], and remove these extra letters.
- Words transformation, these include many variants, for example:
  - In some texts in Spanish, letters in words are substituted by numbers or symbols, (for example "H014" instead of "HOLA," Hello in English). Detecting these symbols and transforming the string into a known word is one task of the preprocessing phase.
  - It is common to interchange one word instead of another; quite similar to synonyms. Some authors present a Spanish specific lexicon of social networks. It is a list of words in Spanish that is used in social networks, and that is understood with a completely different meaning to the common one. Identifying these words and substituting them with other ones makes the text clearer to understand.

8.3.4 Feature Extraction

Machine learning methods produce better results when they are fed with characteristics extracted from a text, instead of providing them with the raw text. The features extracted can be very simple such as identifying the presence of terms (bag-of-word or BoW), or more complex such as lexical and syntactic features.

For this research, we used a lexicon-based feature which consists of counting sentiment terms in each document. From these frequencies, derived features can be obtained, such as a ratio of term frequency on the document, the ratio of term frequency on the whole corpus, and the absolute value of the difference between both previous ratios.

Word2vect calculates the distribution probability of terms in a document. This technique can discover semantic relations among terms in the corpus. It is computed by training a neural network, and the vectors that represent each word are the synaptic weights.

One of the most common features used for sentiment analysis is the term frequency–inverse document frequency. It is a matrix that contains the inverse of the frequency of terms that occur in a set of documents. Therefore, each entry of the

matrix has a low value for terms that occur very frequently in the document set, and a high value for those that occur rarely.

Regardless of the feature extracted, each document (or tweet in our case) is transformed into a vector. We have assigned a label (sentiment) to some of these vectors that are set manually. This way, we obtain a labeled data set, and we use it as the input of machine learning methods.

8.3.5 Classification Methods for Sentiment Analysis

Classification methods are supervised learning methods of machine learning. This means that they need labeled data to build a model that is called a classifier. For sentiment analysis, each sentiment (positive, negative, and neutral in most and simple cases, or joy, anger, sadness, surprise, displeasure, and fear in Ekman's model) is considered a category or class. The purpose of classifiers is to predict the class of previously unseen data. Therefore, they are used to determine the polarity or sentiment of opinions post on social networks as Twitter. Most common classifiers for sentiment analysis are the following:

- Support vector machine. It is a classifier that computes the optimal separating hyperplane. It solves a quadratic programming problem to compute the hyperplane with maximal margin.
- Naive Bayes. It is a probabilistic model that considers that each variable is independent of the rest.
- Decision tree. Decision trees are classifiers whose structure resembles a flowchart. A classifier of this type is induced by partitioning the input space recursively, up to a level of purity of each partition is satisfied.
- Logistic regression. It is a statistical procedure that estimates relationships between attributes and classes. Logistic regression is quite similar to linear regression, but is oriented for categorical outputs.
- Neural network. It is a classification method inspired by the human brain. The most popular training method for neural networks is backpropagation.

Currently, some libraries facilitate the generation of classifiers without the need to implement them from scratch. Some of these libraries for the Python programming language are scikit-learn, NLTK, and SciPy.

8.3.6 Evaluation of Classifiers

One of the techniques most commonly used for evaluating classification methods is 10-cross-validation. In this technique, the data set is partitioned into two subsets; one of them is the training set, used to build the prediction model (train a classifier).

Table 8.1 Confusion matrix for two classes

	Actual class			
		Positive	Negative	Total
Prediction	Positive	TP	FP	Number instances predicted as positive
	Negative	FN	TN	Number instances predicted as negative
	Total	Number of actual positive instances	Number of actual negative instances	*N*: Number of instances

The other subset is the test set, used to observe the prediction of a classifier and to compare it against the true value. This process is repeated 10 times to obtain an average of the performance of a classifier.

It is useful to build a confusion matrix to assess the performance of classifiers. The entries of this matrix contain the counts of the actual categories or classes in a data set and the number of correct and incorrect predictions made by a classifier. Table 8.1 shows the confusion matrix for the simplest case of two classes that are called the positive and the negative class.

In Table 8.1, TP is the number of instances (tweets, in our case) of the positive class that is predicted as positive by a classifier; TN is the number of instances of the negative class that is predicted as negative by a classifier; and FP/FN is the number of instances of positive/negative class that is predicted as negative/positive by a classifier. For these two entries, the classifier commits an error.

Based on the values of a confusion matrix, the following measures can be computed:

- Classification accuracy is the percentage of correct predictions. Classification accuracy = (TP + TN)/(N).
- Precision measures how accurate the classifier to predict positive instances is. Precision = TP/(TP + FP).
- Recall measures how accurate the classifier to predict negative instances is. Recall = TP(TN + FN).
- F1 score combines precision and recall in one formula. F1 score = 2 * (Precision * Recall/(Precision + Recall)).

8.3.7 Using Classification Methods for Sentiment Analysis

Once a classifier has been evaluated, it can be applied to predict new instances. It is recommended to build more than one classifier, and then choose the one with the best performance. Depending on the problem being solved with machine learning methods, the data can have two or more classes. The first case is a binary classification problem; the second case is a multiclass problem.

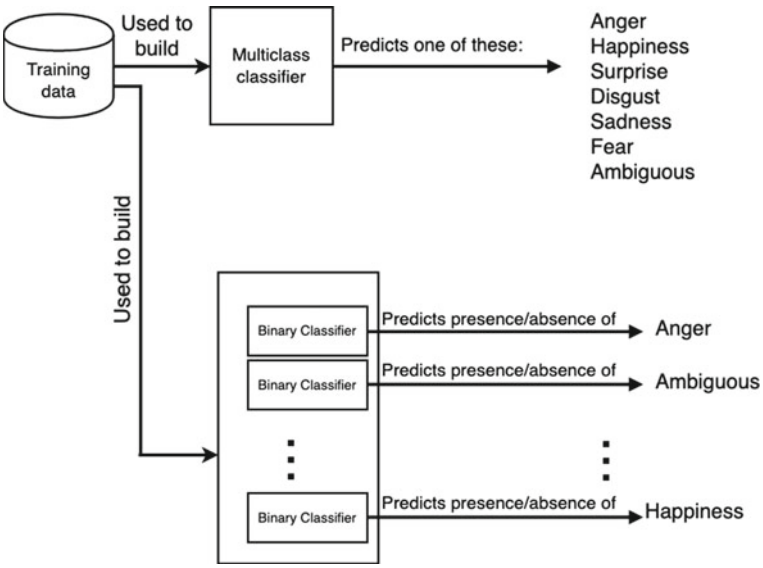

Fig. 8.3 Classifier architecture for sentiment analysis

If there are more than two classes (sentiments) in the analyzed data, then there are at least two possibilities for applying classification methods. The first option is to merge all classes in one categorical attribute. In our case, the possible values of it are the seven classes (six from Ekman's emotions and the class ambiguous). As stated before, it is a problem of multiclass classification. It is well-known that this type of problem is more complex than a binary class classification problem, and that the performance of classifiers is usually better for the latter. The second option—the one used in this chapter—is to consider each sentiment independently of the others. This is a binary classification problem. Therefore, instead of predicting one of the sentiments for a new document (tweet), the presence/absence of only one of them is predicted with each classifier. Figure 8.3 shows the two explained approaches.

8.4 Results

As an example of the application of the methodology explained above, we used data about Mexico's earthquake occurred on September 19th, 2017. For our experiments, we downloaded a data set (corpus) that correspond to each trend identified during this earthquake, these trends are #TvAztecaMiente (TV Azteca Lies), #TodosSomosMexico (WeAreAllMexico), #Terremoto (Earthquake), #TemblorMx (Earthquake Mx), #SomosMexico (WeAreMexico), #skyAlertMx (SkyAlertMx), #SismoMX (EarthquakeMX), #SismoMexico2017 (EarthquakeMexico2017), #Sismo (Earthquake), #RoboComoGraco (StealLikeGraco), #PueblaSigueDePie

(PueblaStillStanding), #PrayForMexico, #PartidosDenSuDinero (PoliticalParties GiveTheirMoney), #PartidosDenNuestroDinero (PoliticalPartiesGiveOurMoney), #Millenials, #MexicoEstaDepie (PueblaStillStanding), #Jojutla (Jojutla), #Fuerza Mexico (ForceMexico), #FuerteMexico (FortMexico), #FinDelMundo (EndOf World), #AyudaMexico (HelpMexico), #AyudaCdMx (HelpMxCity), #Alerta Sismica (SeismicAlert), and #SismoMX (EarthqueakeMX).

We focused on labeling manually a portion of tweets, avoiding to label retweets of the corpora downloaded. In our case, for each corpus, we selected about 10% of the tweets randomly. Each tweet was read independently by three people to identify the presence/absence of each one of the following: anger, happiness, surprise, disgust, sadness, fear, and ambiguous. We chose as the label of tweets the opinion of the majority of taggers. The column "only tweets" of Table 8.2 shows the number of tweets for each corpus. After labeling the data, the number of times each feeling was detected is shown in the corresponding column ("Anger," "Happiness," etc.) of Table 8.2.

It can be seen in Table 8.2 that the frequency values for sentiments in these corpora are very low. For example, for the corpus #TVAztecaMiente, 119 tweets were labeled; the sentiment surprise was present in 85.7% (102 out of 119) of them. However, disgust and fear are only 0.8% (1 out of 119) of tweets read and labeled. According to the labeled corpora, the predominated emotions were sadness, surprise, and happiness. There was also much ambiguous information, as some Twitter users posted information that was not relevant for the event (13.9%).

We applied preprocessing to text removing numbers, exclamation marks, question marks, and punctuation marks from tweets, as mentioned earlier. On the other hand, although many features can be extracted from the analyzed tweets, it is recommended to begin with the simplest features. We computed the term "frequency-inverse document frequency matrix" with the preprocessed texts. This matrix was used to feed three classification methods. The purpose was to build classifiers to identify automatically the sentiments presented in not labeled tweets.

We built three classifiers from data, Naive Bayes (NB), decision tree (DT), and support vector machine (SVM). The parameters of DT and SVM were tuned using the grid search technique; NB classifier does not have any parameter to optimize. Table 8.3 shows the performances of three classifiers built with each corpus. The method to obtain these performances was 10-cross-validation.

In Table 8.3, the best performances among the three classifiers are marked in bold; the white spaces mean that the corresponding classifier could not produce a response. This is usually the case for data with only one class. It is important to clarify that the classification accuracy is not the only evaluation that needs to be applied to classifiers. Especially, for cases of imbalanced data sets, more measures such as precision, recall, and F1 score can be necessary. Naive Bayes and support vector machine were the classifiers with the best accuracy.

The classifiers achieve the following performance on average for each emotion, for anger: 91.4%, joy: 78.8%, surprise: 81.3%, disgust: 89.7%, sadness: 76.8%, fear: 66.1%, and for ambiguous: 50.3% (see Table 8.3).

Table 8.2 Sentiments found on the corpora labeled manually

Corpus	Total count with RT	Only tweets	Anger	Happiness	Surprise	Disgust	Sadness	Fear	Ambiguous	**Sentiments found**
#TvAztecaMiente	5323	1133	37	11	102	1	41	1	5	193
#TodosSomosMexico	6340	2965	2	256	225	3	11	3	9	500
#Terremoto	8922	1078	15	182	148	1	639	71	107	1056
#TemblorMx	2648	752	24	26	49	0	242	22	25	363
#SomosMexico	8707	1292	0	92	16	5	11	0	6	124
#skyAlertMx	252	252	0	7	6	0	10	10	3	33
#SismoMX	9992	1440	13	175	178	26	181	3	29	576
#SismoMexico2017	10,000	2220	0	0	10	549	114	64	361	737
#Sismo	10,000	781	30	133	177	0	107	73	24	520
#RoboComoGraco	10,000	1470	251	75	54	36	69	27	38	512
#PueblaSigueDePie	3171	791	73	199	164	80	236	155	4	73
#PrayForMexico	10,000	1779	0	0	0	180	427	458	138	1065
#PartidosDenSuDinero	10,000	1695	0	0	114	45	107	0	20	266
#PartidosDenNuestroDinero	10,000	2699	123	52	102	0	33	4	9	314
#Millenials	4180	2838	33	166	149	7	15	8	52	378
#MexicoEstaDepie	10,000	1305	21	92	294	6	572	45	5	1030
#Jojutla	10,000	629	2	15	2	0	18	5	30	42
#FuerzaMexico	10,000	1458	3	72	6	0	23	3	40	107
#FuerteMexico	10,000	1207	32	202	213	82	123	95	473	747
#FinDelMundo	10,000	4247	40	150	125	108	38	28	62	489
#AyudaMexico	10,000	1871	38	151	157	216	191	71	223	824
#AyudaCdMx	10,000	1231	12	47	24	1	23	10	56	117
#AlertaSismica	10,000	1622	15	53	92	15	48	53	15	275
#SismoMX	9992	1440	13	175	178	26	181	3	29	676
Total	196,356	37,404	777	2132	2421	1307	3224	1057	1763	10918
Percentage			6.1%	16.8%	19.1%	10.3%	25.4%	8.3%	13.9%	

Table 8.3 Performances of classifiers on each corpus

Corpus	Classifier	Anger (%)	Happiness (%)	Surprise (%)	Disgust (%)	Sadness (%)	Fear (%)	Ambiguous (%)
#Todossomosmexico	NB	**99.22**	**53.91**	53.91	**98.44**	**96.88**	**100.00**	**99.22**
	DT	**99.22**	52.34	49.22	**98.44**	94.53	**100.00**	97.66
	SVM	**99.22**	48.44	**54.69**	**98.44**	**96.88**	**100.00**	**99.22**
#Terremoto	NB	**98.80**	**79.20**	**86.00**	**100.00**	**67.60**	**94.00**	**91.60**
	DT	**98.00**	76.00	82.00	**100.00**	66.00	91.60	86.80
	SVM	**98.80**	**79.20**	85.60	**100.00**	60.00	**94.40**	88.40
#Temblormx	NB	95.29	89.41	**83.53**	72.94	**96.47**	**92.94**	
	DT	91.76	**90.59**	**83.53**	**74.12**	92.94	82.35	
	SVM	**95.29**	89.41	81.18	70.59	95.29	**92.94**	
#Somosmexico	NB	**69.70**	**93.94**	**90.91**	**90.91**	**93.94**		
	DT	18.18	87.88	**90.91**	**90.91**	75.76		
	SVM	**69.70**	**93.94**	**90.91**	**90.91**	**93.94**		
#Skyalertmx	NB	66.67	**100.00**	77.78	**66.67**	**88.89**		
	DT	**77.78**	66.67	**88.89**	44.44	**88.89**		
	SVM	66.67	**100.00**	77.78	**66.67**	**88.89**		
#SismoMX	NB	**97.67**	**62.79**	**64.34**	**95.35**	**65.12**	**100.00**	**95.35**
	DT	94.57	60.47	63.57	90.70	59.69	99.22	89.92
	SVM	**97.67**	61.24	62.79	**95.35**	**65.12**	**100.00**	**95.35**
#Sismomexico2017	NB	**98.81**	**88.14**	**93.28**	**68.77**			
	DT	98.02	82.61	88.93	58.10			
	SVM	**98.81**	**88.14**	**93.28**	67.98			
#Sismo	NB	**96.95**	**70.23**	68.70	**78.63**	**88.55**	**93.89**	
	DT	94.66	67.18	59.54	75.57	83.97	90.08	
	SVM	**96.95**	67.18	68.70	**78.63**	**88.55**	**93.89**	

(continued)

Table 8.3 (continued)

Corpus	Classifier	Anger (%)	Happiness (%)	Surprise (%)	Disgust (%)	Sadness (%)	Fear (%)	Ambiguous (%)
#Robocomograco	NB	**56.15**	**87.69**	**90.77**	**91.54**	**91.54**	**96.92**	**90.00**
	DT	53.85	82.31	86.15	86.15	84.62	**92.31**	87.69
	SVM	54.62	**87.69**	**90.77**	**91.54**	**91.54**	**96.92**	**90.00**
#Pueblasiguedepie	NB	86.36	71.72	73.74	89.90	63.64	76.26	**99.49**
	DT	87.88	72.22	72.22	88.38	61.11	72.22	98.99
	SVM	**90.40**	**79.29**	**76.26**	**91.41**	**70.20**	**79.80**	**99.49**
#Prayformexico	NB	90.11	70.34	**71.91**	**93.26**			
	DT	82.70	64.27	62.25	88.99			
	SVM	**90.34**	**71.01**	**71.91**	**93.26**			
#Partidosdensudinero	NB	**92.92**	**98.11**	**93.63**	**98.58**			
	DT	87.74	96.70	90.09	97.41			
	SVM	**92.92**	**98.11**	**93.63**	**98.58**			
#PartidosDenNuestroDinero	NB	60.00	**84.29**	**88.57**	**95.71**	**97.14**		
	DT	**65.71**	77.14	80.00	**95.71**	95.71		
	SVM	48.57	**84.29**	**88.57**	**95.71**	**97.14**		
#Millenials	NB	89.52	**80.95**	69.52	**100.00**	**95.24**	**99.05**	84.76
	DT	**94.29**	75.24	**79.05**	**100.00**	93.33	98.10	85.71
	SVM	90.48	76.19	77.14	**100.00**	95.24	**99.05**	**87.62**
#Mexicoestadepie	NB	**97.60**	92.80	70.80	**99.60**	**66.00**	**95.20**	**99.60**
	DT	97.20	86.80	66.80	98.40	54.80	91.20	99.20
	SVM	**97.60**	**93.20**	**71.20**	**99.60**	60.00	**95.20**	**99.60**
#Jojutla	NB	**94.44**	**77.78**	**100.00**	**77.78**	**94.44**	50.00	
	DT	**94.44**	38.89	**100.00**	**77.78**	**94.44**	38.89	
	SVM	**94.44**	**77.78**	**100.00**	**77.78**	**94.44**	**55.56**	

(continued)

Table 8.3 (continued)

Corpus	Classifier	Anger (%)	Happiness (%)	Surprise (%)	Disgust (%)	Sadness (%)	Fear (%)	Ambiguous (%)
#Fuerzamexico1	NB	**97.30**	45.95	**94.59**	**86.49**	**97.30**	**72.97**	
	DT	**97.30**	**54.05**	**94.59**	83.78	**97.30**	67.57	
	SVM	**97.30**	51.35	**94.59**	**86.49**	**97.30**	**72.97**	
#Fuerzamexico2	NB	95.56	**95.56**	73.33	**100.00**	**93.33**		
	DT	**97.78**	88.89	**80.00**	**100.00**	**93.33**		
	SVM	95.56	**95.56**	75.56	**100.00**	**93.33**		
#Fuerzamexico3	NB	**98.40**	**92.00**	**86.00**	**100.00**	**97.60**	**98.80**	**99.20**
	DT	**98.00**	90.40	82.80	**100.00**	95.20	98.40	98.00
	SVM	**98.40**	**92.00**	85.60	**100.00**	**97.60**	**98.80**	**99.20**
#Fuerzamexico4	NB	92.00	**68.00**	**94.00**	**88.00**	**92.00**	**94.00**	**84.00**
	DT	**94.00**	62.00	86.00	84.00	**92.00**	92.00	76.00
	SVM	92.00	38.00	**94.00**	**88.00**	**92.00**	**94.00**	**84.00**
#Fuerzamexico5	NB	**97.50**	**57.50**	**77.50**	**97.50**	**82.50**	**90.00**	**90.00**
	DT	**97.50**	**57.50**	75.00	95.00	80.00	82.50	72.50
	SVM	**97.50**	**57.50**	**77.50**	**97.50**	**82.50**	**90.00**	**90.00**
#Fuertemexico	NB	**96.80**	**82.40**	**75.20**	**91.60**	**84.80**	**90.40**	**70.00**
	DT	95.60	70.80	67.60	90.80	82.40	86.80	64.00
	SVM	**96.80**	**82.40**	**75.20**	**91.60**	**84.80**	**90.40**	62.40
#Findelmundo	NB	**92.86**	**68.75**	**76.79**	**77.68**	**90.18**	**97.32**	83.93
	DT	**92.86**	60.71	66.07	71.43	87.50	95.54	**84.82**
	SVM	**92.86**	63.39	**76.79**	**77.68**	**90.18**	**97.32**	83.93
#AyudaMexico	NB	**96.15**	**86.15**	**86.92**	**80.77**	80.77	**93.08**	75.77
	DT	90.38	75.77	76.54	71.92	66.92	91.92	68.46
	SVM	**96.15**	**86.15**	**86.92**	**80.77**	**81.54**	**93.08**	**76.15**

(continued)

Table 8.3 (continued)

Corpus	Classifier	Anger (%)	Happiness (%)	Surprise (%)	Disgust (%)	Sadness (%)	Fear (%)	Ambiguous (%)
#Ayudacdmx	NB	**87.18**	**63.16**	**89.74**	**92.31**	**97.44**	**58.97**	
	DT	82.05	57.89	84.62	84.62	92.31	51.28	
	SVM	**87.18**	**63.16**	**89.74**	**92.31**	**97.44**	**58.97**	
#Alertasismica	NB	**97.26**	**80.82**	68.49	**93.15**	**80.82**	**80.82**	**97.26**
	DT	93.15	65.75	57.53	90.41	78.08	76.71	91.78
	SVM	**97.26**	79.45	**69.86**	**93.15**	**80.82**	79.45	**97.26**
#SismoMX	NB	**97.67**	62.79	**64.34**	**95.35**	**65.12**	**100.00**	**95.35**
	DT	95.35	**63.57**	61.24	90.70	60.47	99.22	89.92
	SVM	**97.67**	61.24	62.79	**95.35**	**65.12**	**100.00**	**95.35**
Average of the bests		**91.44**	**78.81**	**81.29**	**89.76**	**76.84**	**66.06**	**50.32**

Table 8.4 Predictions for each corpus

Corpus	Anger	Happiness	Surprise	Disgust	Sadness	Fear	Ambiguous
#Todossomosmexico	0	3748	0	0	0	256	0
#Terremoto	0	30	17	0	7564	0	236
#Temblormx	0	110	1	2221	162	2	0
#Somosmexico	9859	0	0	5	0	0	0
#Skyalertmx	21	0	10	0	0	0	0
#SismoMX	0	243	93	0	222	0	0
#Sismomexico2017	0	0	0	1608	0	0	0
#Sismo	0	264	594	42	21	0	0
#Robocomograco	4083	0	0	0	29	7	0
#Pueblasiguedepie	0	0	0	0	0	0	0
#Prayformexico	0	0	126	0	0	0	0
#Partidosdensudinero	0	0	130	0	0	0	0
#PartidosDenNuestroDinero	4344	1	0	0	0	0	0
#Millenials	342	100	1084	13	13	0	0
#Mexicoestadepie	0	0	0	0	6902	0	0
#Jojutla	0	0	0	6	0	0	0
#Fuerzamexico1	0	5842	0	1	0	0	0
#Fuerzamexico2	256	9821	1126	0	0	205	0
#Fuerzamexico3	0	9000	9	0	0	0	0
#Fuerzamexico4	283	1553	0	5	0	1	0
#Fuerzamexico5	0	9746	0	0	0	0	0
#Fuertemexico	0	3	11	0	11	9	4449

(continued)

Table 8.4 (continued)

Corpus	Anger	Happiness	Surprise	Disgust	Sadness	Fear	Ambiguous
#Findelmundo	8	501	384	15	0	0	1455
#AyudaMexico	0	6	5	43	0	0	0
#Ayudacdmx	0	5	0	1	1	17	0
#Alertasismica	0	259	0	0	0	0	0
#SismoMX	0	4150	93	0	222	125	0
Total	19196	45382	3683	3960	15147	622	6140

We use the best classifier for each sentiment on each data set to obtain the results shown in Table 8.4. It is interesting to mention that in most of these predictions, the classifiers with the best performances made a similar number of predictions about the presence of a sentiment, concerning the taggers. This allowed us to claim that the results presented in Table 8.4 agree with the data labeled manually (Table 8.2).

Based on Table 8.4 (row named "total"), it is possible to claim that almost half of the emotions in data correspond to happiness (48.2%). Anger represents 20.4% of all predicted emotions, and sadness represents 16.1%. This can be attributed to the fact that a large part of the tweets was related to solidarity, support, and help; another part with claims toward mass media, political parties, and governors; and others related to bad news about dead or disappeared persons.

There were very few emotions of surprise (3.9%), disgust (4.2%), and fear (0.7%) in all the analyzed data because much of the data were collected after the earthquake. Ambiguous tweets represented the 6.5% of data. Figure 8.4 shows a graphic summary of the sentiments found in the data analyzed through machine learning techniques.

The behavior of emotions is explained because the information was downloaded after the earthquake. Many of the Twitter users were happy because they were safe, and they found lost people or pets and also because of the actions carried out by the civil society. Unfortunately, there were many people who, due to the conditions of the event, could not interact on Twitter.

8.5 Conclusions and Future Research

Social media is an important source of data that nurture the big data every day. Most of the works on sentiment analysis only consider three possible cases for each opinion or post of users of social media. These are positive, negative, and neutral. The use of emotional frameworks, such as Ekman's model, allowed defining the emotions generated in social media more precisely, as they are related to human behavior.

It is useful to analyze identified emotions in order to study the reactions of social media users to certain kinds of events, people, services, or products, as well as their posture and the effects that they generate.

In this chapter, we explained a methodology to apply sentiment analysis on data from Twitter through machine learning techniques. As an example, we analyzed the emotions on tweets after an earthquake event in Mexico. Based on Ekman's model, we found that the most frequent predicted emotions were happiness, anger, and sadness, and 6.5% of predicted tweets were irrelevant. We built three classifiers to determine the emotions of tweets that belong to the same topic, and Naive Bayes and support vector machine were the classifiers with the best accuracy for predicting emotions.

Based on previous experience with sentiment analysis, we can suggest the following recommendations:

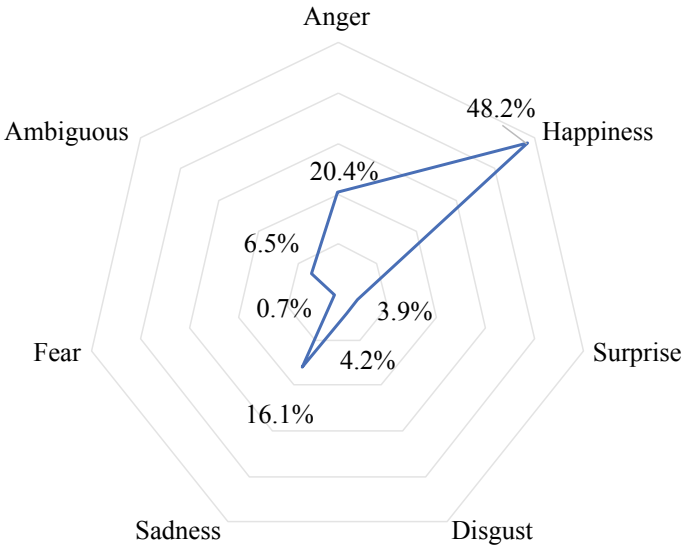

Fig. 8.4 Summary of sentiments in the analyzed corpora

- Preprocessing is a necessary phase for successfully applying machine learning methods for sentiment analysis. Most of the steps of preprocessing can be implemented by software.
- Labeling documents is another important phase before applying machine learning methods for sentiment analysis. The recommendation is to label enough data to produce accurate models. How much is enough data? Some authors suggest about 3000 labels [35]. Although they report that increasing this number damages the performance of classifiers, it is possible that this happens because if data of the same class have a different distribution, then the decision boundary is more complex. That is, the more data are added, the more concepts related to the same class must be discovered by the classification method. This causes the performance of the classifiers to be diminished.
- In general, the imbalance of classes affects the performance of classification methods dramatically. If certain sentiment or emotion predominates (majority class) and there are just a few of the other sentiments (minority class), the classifier will predict the majority class most of the time, or even always. In these cases, applying a balancing method, such as SMOTE [36], or labeling more data to balance the number of sentiments can be helpful.
- A larger number of sentiments or classes in data diminish the performance of classification methods. The greater the number of classes, the lower the performance of a classifier. The recommendation is to merge classes or get rid of the ones that are not relevant before using a supervised machine learning method. The approach used in this chapter is to create a classifier for each class, and then apply a binary classification method;

- It is recommended to build more than one classifier, each one of a different type, then measure their performances. If the classifiers are not achieving good results, extract more features from text and repeat the process.

The main purpose of this chapter is to provide insights into the use of sentiment analysis methods. A second contribution is the expansion of the emotions range, from three (negative, neutral, positive) to six in order to provide more elements to understand how users interact with social media platforms. A third contribution is the analysis of the data sets from Mexico's 2017 earthquake and expanding the understanding of Mexican emotions on social networks. Future research will include validation of the method with different data sets and emotions; the addition of new artificial intelligence techniques to improve accuracy; and also new research paths to clean data are considered along with other techniques for computing emotions. We hope this contribution will foster the use of methods and techniques to understand emotions in social media in the future.

References

1. Almarabeh T, Majdalawi YK, Mohammad H (2016) Cloud computing of e-government
2. Sasikala P (2012) Cloud computing and E-governance: advances, opportunities and challenges. Intl J Cloud Appl Comput (IJCAC) 2(4):32–52
3. Jadhav B, Patankar A (2018) Opportunities and challenges in integrating cloud computing and big data analytics to e-governance. Int J Comput Appl 180(15):6–11
4. Lohr S (2012) The age of big data. New York Times, 11 (2012)
5. Liebowitz J (2001) Knowledge management and its link to artificial intelligence. Expert Syst Appl 20(1):1–6
6. Fan W, Bifet A (2013) Mining big data: current status, and forecast to the future. ACM SIGKDD Explor Newsl 14(2):1–5
7. Wu X, Zhu X, Wu GQ, Ding W (2014) Data mining with big data. IEEE Trans Knowl Data Eng 26(1):97–107
8. Sandoval-Almazán R, Valle-Cruz D (2016) Understanding network links in Twitter: a Mexican case study. In: Proceedings of the 17th international digital government research conference on digital government research. ACM, pp 122–128
9. Shaikh S, Feldman LB, Barach E, Marzouki Y (2017) Tweet sentiment analysis with pronoun choice reveals online community dynamics in response to crisis events. In: Advances in cross-cultural decision making. Springer, Cham, pp 345–356
10. Vo BKH, Collier NIGEL (2013) Twitter emotion analysis in earthquake situations. Intl J Comput Linguist Appl 4(1):159–173
11. Sandoval-Almazan R (2019) Using twitter in political campaigns: The case of the PRI candidate in Mexico. In: Civic engagement and politics: concepts, methodologies, tools, and applications. IGI Global, pp 710–726
12. Sandoval-Almazán R, Valle-Cruz D (2018) Towards an understanding of Twitter networks: the case of the state of Mexico. First Monday, vol 23, no 4
13. Sandoval-Almazan R, Valle-Cruz D (2018) Facebook impact and sentiment analysis on political campaigns. In: Proceedings of the 19th annual international conference on digital government research: governance in the data age. ACM, p 56
14. Chen M, Zhang Y, Li Y, Mao S, Leung VC (2015) EMC: emotion-aware mobile cloud computing in 5G. IEEE Netw 29(2):32–38

15. Preethi G, Krishna PV, Obaidat MS, Saritha V, Yenduri S (2017) Application of deep learning to sentiment analysis for recommender system on cloud. In 2017 international conference on computer, information and telecommunication systems (CITS). IEEE, pp 93–97
16. Krishna PV, Misra S, Joshi D, Obaidat MS (2013) Learning automata based sentiment analysis for recommender system on cloud. In: 2013 international conference on computer, information and telecommunication systems (CITS). IEEE, pp 1–5
17. Assunção MD, Calheiros RN, Bianchi S, Netto MA, Buyya R (2015) Big data computing and clouds: trends and future directions. J Parallel Distrib Comput 79:3–15
18. Karyotis C, Doctor F, Iqbal R, James A, Chang V (2018) A fuzzy computational model of emotion for cloud based sentiment analysis. Inf Sci 433:448–463
19. Ali F, Kwak D, Khan P, Islam SR, Kim KH, Kwak KS (2017) Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling. Transp Res Part C: Emerg Technol 77:33–48
20. Sreeja PS, Mahalakshmi GS (2017) Emotion models: a review. Int J Control Theory Appl 10:651–657
21. Ekman PE, Davidson RJ (1994) The nature of emotion: fundamental questions. Oxford University Press, Oxford
22. Plutchik R (1965) What is an emotion? J Psychol 61(2):295–303
23. Munezero MD, Montero CS, Sutinen E, Pajunen J (2014) Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. IEEE Trans Affect Comput 5(2):101–111
24. Nakamura (1993) Kanjo hyogen jiten [dictionary of emotive expressions]. Tokyodo, Teluk Intan
25. Yang Y, Jia J, Zhang S, Wu B, Chen Q, Li J et al (2014) How do your friends on social media disclose your emotions? In: Twenty-eighth AAAI conference on artificial intelligence
26. Wang Y, Pal A (2015) Detecting emotions in social media: a constrained optimization approach. In: Twenty-fourth international joint conference on artificial intelligence
27. Cambria E, Livingstone A, Hussain A (2012) The hourglass of emotions. In Cognitive behavioural systems. Springer, Berlin, pp 144–157
28. Zhang L, Hua K, Wang H, Qian G, Zhang L (2014) Sentiment analysis on reviews of mobile users. Procedia Comput Sci 34:458–465
29. Zhang L, Hua K, Wang H, Qian G, Zhang L (2014) Sentiment analysis on reviews of mobile users. Procedia Comput Sci 34:458–465
30. Prabowo R, Thelwall M (2009) Sentiment analysis: a combined approach. J Informetr 3 (2):143–157
31. Kumar M, Bala A (2016) Analyzing Twitter sentiments through big data. In: 2016 3rd international conference on computing for sustainable global development (INDIACom). IEEE, pp 2628–2631
32. Subramaniyaswamy V, Vijayakumar V, Logesh R, Indragandhi V (2015) Unstructured data analysis on big data using map reduce. Procedia Comput Sci 50:456–465
33. Al-Kabi M, Al-Ayyoub M, Alsmadi I, Wahsheh H (2016) A prototype for a standard arabic sentiment analysis corpus. Int Arab J Inf Technol 13(1A):163–170
34. Kune R, Konugurthi PK, Agarwal A, Chillarige RR, Buyya R (2016) The anatomy of big data computing. Softw Pract Exp 46(1):79–105
35. Sidorov G et al (2013) Empirical study of machine learning based approach for opinion mining in tweets. In: Batyrshin I, González Mendoza M (eds) Advances in artificial intelligence. MICAI 2012. Lecture notes in computer science. Springer, Berlin, vol 7629
36. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Int Res AI Access Found 16:321–357