SENTIMENT ANALYSIS OF TEXT AND EMOJI DATA FOR TWITTER NETWORK

Paramita Dey
Department of Information Technology, Government College of Engineering & Ceramic Technology

Soumya Dey
Department of Information Technology, Government College of Engineering and Ceramic Technology

Sentiment Analysis of Text and Emoji Data for Twitter Network

Paramita Dey, Soumya Dey*

Department of Information Technology, Government College of Engineering & Ceramic Technology, India

Abstract

Twitter is a social media platform where users can post, read, and interact with 'tweets'. Third party like corporate organization can take advantage of this huge information by collecting data about their customers' opinions. The use of emoticons on social media and the emotions expressed through them are the subjects of this research paper. The purpose of this paper is to present a model for analyzing emotional responses to real-life Twitter data. The proposed model is based on supervised machine learning algorithms and data on has been collected through crawler "TWEEPY" for empirical analysis. Collected data is pre-processed, pruned and fed into various supervised models. Each tweet is assigned to sentiment based on the user's emotions, positive, negative, or neutral.

Keywords: Twitter, Sentiment analysis, Opinion mining, Machine learning, Naive bayes (NB), Support vector machine (SVM), Bag of words (BOW) model, TF-IDF model

1. Introduction

A Japanese telecom association created emoticons in 1990 as a way to encourage children to use their pager company. "Emoticon" is Japanese for "picture character". Later it is introduced to western world when clients find a concealed console in the Apple iPhone implied for the Japanese market. Emoticons are utilizing web based life to express conclusions. For example, an association can use evaluation assessment to perform measurable looking over to survey how their things are capable by their clients, without conveying surveys or in various ways inconvenience their clients. Typical techniques for concluding inclination on Twitter have focused on whether a tweet is sure or pessimistic, generally called twofold (or polar) feeling examination. Genuine pictures inserted in content and ought not to be mistaken for then collected data is gone through preprocessing and pruning data is then fed emojis (face impersonations made utilizing ASCII characters). Today emoticons are accessible as a discretionary composed language in many cell phones and have turned into a significant part in this regular correspondence, generally used to add enthusiastic prompts to instant messages.

1.1. Background

Natural language processing (NLP) uses the technique of sentiment analysis, commonly referred to as opinion mining, to ascertain if input is positive, negative, or neutral. The emotions expressed in text are examined through sentiment analysis. It is often used to research product reviews, survey result, and customer feedback. Sentiment analysis is useful in many situations, such as social media monitoring, reputation management, and customer experience. For example, researching dozens of product reviews can provide insightful information about product features and pricing. Sentiment analysis focuses on the polarity of text (positive, negative, or neutral), but goes beyond polarity to identify specific moods and emotion (angry, happy, sad), urgency (urgent, not urgent), and even intent (interested vs. not interested).

1.2. Purpose

Sentiment analysis, also known as opinion mining, is useful for film reviews, product reviews, customer service reviews and feedback on any event. It enables us to evaluate whether a product or service is good, bad or preferred. It can be used to determine how people feel about a particular event or person, as well as to determine the polarity of the text, positive, negative, or neutral. Text is divided into different emotional categories using sentiment analysis, which is a type of text classification.

Social media, such as Twitter, has grown in popularity as a result of the rise in social awareness. Twitter is a popular social networking site where anyone can post tweets about any event. It is a public forum where people can freely express their thoughts, feelings, and opinions. People have Twitter accounts as a result of low internet costs, low-cost portable devices, and growing social importance.

The majority of them use Twitter to discuss various incidents. People use Twitter to express their thoughts and feelings in the age of social media. As a result, Twitter has a massive amount of data. We know that a tweet can only be 140 characters long, so people can write tweets with the appropriate emotion or feeling for each word 240+ million active users use this website. About 500 million tweets are generated every day. Tweets are often used to express Twitter's feelings about a particular topic. Twitter provides unlimited access to data.

1.3. Problem statement

The following two subtasks are assigned to the main problem in this paper:

Phrase Level Sentiment Analysis:

Determine whether a word or phrase has positive, negative or neutral impact in the context of a message if it appears in the social media.

Emotion Level Sentiment Analysis:

Determine whether an emoticon/text has a positive, negative, or neutral emotion by analyzing how it makes others feel or what they think about a certain occurrence.

1.4. Scope

A shorthand for facial expressions, such as:-), is an emoticon. It enables the writer to express his feelings, moods, and emotions while also incorporating nonverbal elements into a written message. It aids in attracting the reader's attention and improves and enhances the message's comprehension. The emoji (from Japanese, "picture," and moji, "character") is created to allow for more expressive messages using modern communication technology. Topics covered include celebrations, buildings and vehicles, weather, food and beverages, plants and animals, and emotions, feelings and activities.

In this paper we have elaborated about the related research in section 2. The emoji lexicon is introduced in section 3. The methodology of emotion analysis of emoticon is discussed in section 4. Experimental results of the empirical study is discussed in section 5. Finally, in section 6, the paper is concluded.

2. Related research

A large number of studies on sentiment analysis based on text were performed in recent years. Yogesh Chandra et al. [1] performed sentiment analysis using a machine learning classifier. Use of polarity-based full sentiment analysis and deep learning modes to categorize consumer tweets as "positive" or "negative". N. Kasture et al. [2] offered a consistent arrangement for taking care of emotions on various social networking sites. The analysis of the text's emotions used integrated grammar, annotation, lexicon construction, and semantic networks. S. Bhuta et al. [3] described the essential methods for categorizing emotions and safeguarding data. With several classifiers, including Naive Bayes, Support Vector Machine (SVM), and others, the select vector function's accuracy in classifying electronic products was evaluated. Subhashini L et al. [4] introduced the results of an extensive survey of the current literature on opinion mining. It also discussed how to represent knowledge in opinions, classify them, and extract text features from opinions that contain uncertainty or noise. Mowlaei ME et al. [5], proposed a methodology for sentiment classification using adaptive aspect-based lexicons. The authors discussed about two ways to build two dynamic lexicons that would help classify emotions based on their aspects: a plan based on genetic algorithms and statistics. Kumar KN et al. [6] developed a dynamic vocabulary that may be automatically updated and offered more precise grading for context-related notions. Zvarevashe K et al. [7] utilized various vocabularies from various word dictionaries to arrange each part of the survey.

Sentiment research has historically been used in a variety of industries, including the stock market, airlines, hotels, and healthcare. Mayur's et al. [8] detailed explanation of the techniques used to complete this challenge and for sentiment analysis.

The methodologies utilized were then assessed, contrasted, and investigated in order to develop a thorough understanding of both their benefits and drawbacks. Finally, in order to determine future directions, the difficulties of sentiment analysis were considered. CH. Rayala Vinod Kumar et al. [9] initially conducted sentiment analysis to classify highly unstructured Twitter data. Secondly, the proposed method was then put up against various #tag live tweet Sentiment Analysis (VLS), Naive Bayes, and Convolution Neural Networks. The research methodology's third section covered the algorithm's operation. The consequences of the analysis were delivered by the Naive Bayes, VLS and CNN algorithm.

Though there are several analyses have been conducted on sentiment analysis, most of them are focused into the text analysis. Emoticon (Emoji), along with text to emphasize the emotion or used separately can be used for further analysis. After comparison, we elaborate the effectiveness of our suggested methodology for both text and emoji sentiment analysis.

3. Emoji lexicon

A number of Japanese mobile operators were the first to use emoji. A set of emoji's is of the type is shown in Fig. 1.

A mobile operator offered a set of 176 emojis for use in messaging. Numerous emoji characters were added to the Unicode character set (Unicode 6.0) in 2010, enabling their widespread use. There were 1126 different emoji characters allowed by Unicode 9.0.

Emoticon characters are put away and addressed similarly as all Unicode characters since they are a subset of a bigger Unicode character set. To draw in additional clients, a few sites have made their own textual styles to show emoticon characters in a more appealing manner. Contingent upon the number framework, programming language, and even documentation, Unicode characters might have various qualities for portrayal.

Data collection methods for the Sentiment Analysis (SA) model vary depending on the tools used. Contingent upon the device, researchers might observe an alternate rendition of the emoticon character in the gathered information. For instance, the Twitter API for R gives these characters in an alternate arrangement from the Hadoop Flume administration. The correct Unicode representation of the software should be used when searching for a specific emoji character.

The Sentiment Analysis of emoji model [10] proposed a method for assigning a score to emoji characters based on the sentiment score of the text and assigning a score between -1 and 1 to the emoji character. Native speakers of the text's language received the plain text. The speakers had the text a negative, natural, or positive rating. The average score for each emoji character was calculated using the sum of all text scores.

Consequently, we have developed an emoji Lexicon that includes sentiment scores ranging from 0 to 1 as well as the emoji's rank. This lexicon can aid in the parsing of text and the detection of emoji characters in data gathered from microblogging services or social media platforms.

4. Methodology

A stepwise methodology for sentiment analysis is elaborated in following subsections. Data is collected through Twitter API, preprocesses and then analyzed using machine learning analysis as shown in Fig. 2.

4.1. Twitter API

Twitter offers REST and Streaming APIs. The REST API is made up of two APIs: the REST API itself and the Search API (whose distinction is completely because of their history of advancement). REST APIs and streaming APIs are contrasted by: The streaming API encourages
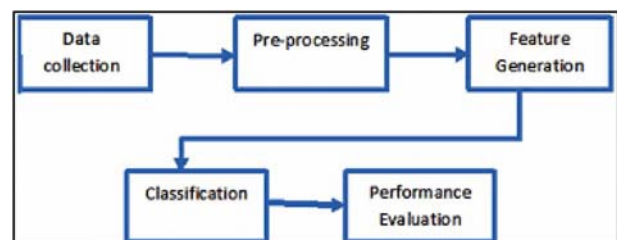

Fig. 1. Screenshot of Emoji's template.

Fig. 2. Flow diagram of Sentiment Analysis.

long−term relationships and provides information that is essentially continuous. The rate-limited REST APIs support transient relationships (one can download a specific measure of information however not more every day). Access to information on Twitter, such as notifications and user data, is made possible using REST APIs. Twitter does not, in any case, make information available that is older than a week or so. REST access is thus limited to data that hasn't been tweeted in more than seven days. As a result, while Streaming API gives access to information as it is being tweeted, REST API enables access to information that has already been collected [6].

4.2. Data processing

Data are processed through several steps as elaborated in Fig. 3.

Data accumulating - In this phase the information to be analyzed is slid from various sources such as blogs, social systems (Twitter, MySpace, so forth) depending on the area of application. We need to provide information in this company collected from the Twitter API.
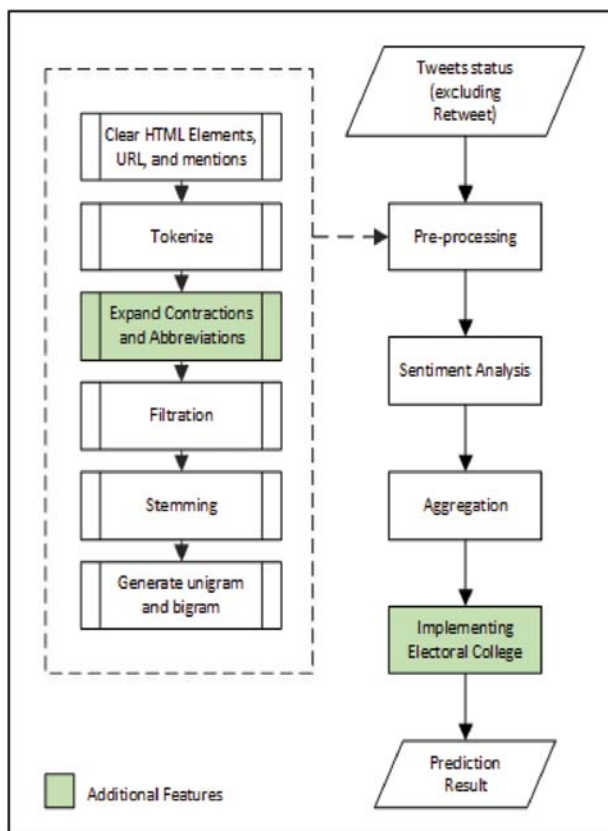

Fig. 3. Step diagram.

Pre-arranging - In this degree, the gained information is cleaned and prepared for supporting it into the classifier. Cleaning carries extraction of articulations and images. Cleaning included Keyword and symbol extraction. Correct uppercase and all lowercase to typical cases, evacuate non-English archives, and banish, important spaces, tabs, so on.

Training Data - A hand-stamped collection of records is set up with the aid of maximum general applied transparently supporting method.

Classification - This is the point of interest of the complete technique. Contingent on the prerequisite of the application SVM or Naïve Bayes is conveyed for investigation. The classifier (following completing the coaching) is in shape to receive to the consistent tweets/content material for supposition extraction motive.

Consequences - Outcomes are plotted ward at the kind of portrayal selected for instance diagrams, frames, and so on. Execution tuning is achieved earlier than the presence of the evaluation.

4.3. Tokenization and stemming

Tokenization is the procedure by which enormous amount of content is partitioned into littler parts called tokens. Regular language preparing is utilized for structure applications, for example, content arrangement, shrewd chat-bot, nostalgic examination, language interpretation, and so forth. It winds up indispensable to comprehend the example in the content to accomplish the above-expressed reason. These tokens are useful for coming across such examples simply as is considered as a base develop for stemming and lemmatization. Stemming and lemmatization are frameworks for content standardization (also known as word standardization) in the field of work language management that are used to organize content, phrases, and reviews for further preparation. Since the 1960s, software planning has broken down and calculated stemming and lemmatization. In this educational activity, you will learn about stemming and lemmatization in a rational framework that covers the inspiration, multiple well-known tests, applications of stemming and lemmatization, as well as instructions on how to stem and lemmatize words, phrases, and data using the Python nltk group, which is the standard language toolbox.

4.4. Lemmatization

Lemmatization is the manner closer to changing over a phrase to its base plan. The separation among Stemming and Lemmatization can't abstain from
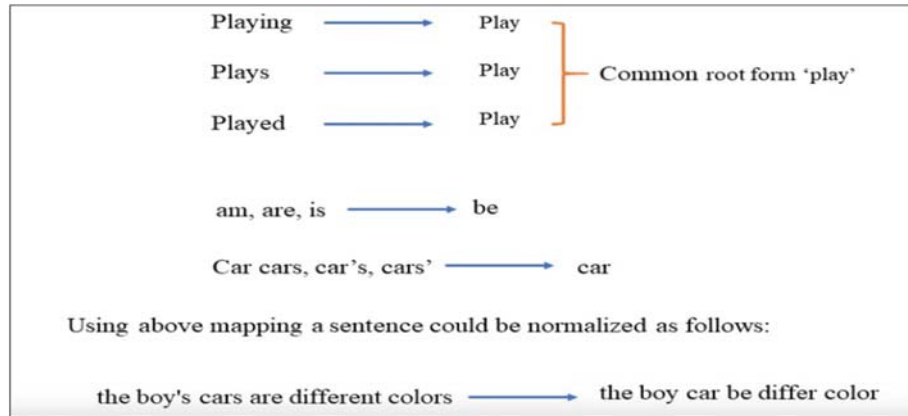
Fig. 4. Example of stemming.

being, Lemmatization ponders the exceptional situation and changes over the word to its basic base plan, while Stemming just expels the last a few characters, regularly prompting mixed up repercussions and spelling mishandles.

For example, Lemmatization would accurately distinguish the base type of 'caring' to 'mind', while, Stemming would remove the 'ing' part and convert it to vehicle.

'Caring'- > Lemmatization->'Care'.

'Caring' - > Stemming -> 'Car'.

4.5. Stop words

Stop words are the exceptionally familiar words like 'if', 'however', 'we', 'he', 'she', and 'they'. We can typically eliminate these words without changing the semantics of a text and doing so frequently (however not generally) works on the exhibition of a model (see Fig. 4).

4.6. Model

4.6.1. BOW model

Bag of Words (BOW) is a technique to concentrate highlights from content/emoticon documents as shown in Fig. 5. Those highlights may be applied for training machine getting to know algorithms. It makes a vocabulary of all of the special words happening in each one of the reviews within the instruction set. In basic terms, it's a gathering of words to speak to a sentence with word check and for the most part slighting the request in which they show up [11,12] (see Fig. 6).

BOW is an approach widely used with natural language processing, information retrieval from documents, document classifications. On a high level, it involves the following steps.

4.6.2. TF-IDF model

TF-IDF stands for "Term Frequency — Inverse Data Frequency". The term frequency (tf) function tells us how often a word appears in each archive of the corpus. It is the ratio between the frequency of a word's appearance in a report and the total number of words in that report. It grows as the number of events containing that word within the report grows. Every archive has a unique tf [11].

$$tf_{i,j} = n_{i,j} / \sum_k n_{i,j}$$

The Inverse Data Frequency (idf) method is used to calculate the prevalence of rare words across all corpus reports. A high IDF score is given to words that appear seldom in the corpus. It is provided by the circumstance below.

$$idf(w) = \log(N / df_t)$$

Combining these two results in the TF-IDF score(w) for a word in a corpus document. The outcome of tf and idf is as follows:

Fig. 5. BOW model.

Fig. 6. Support vector machine algorithm.

$$w_{I,j} = tf_{i,j} \times \log(N / df_i)$$

where $tf_{i,j}$ = number of occurrences of i in j.
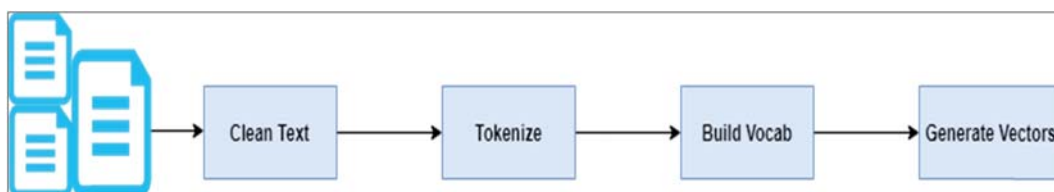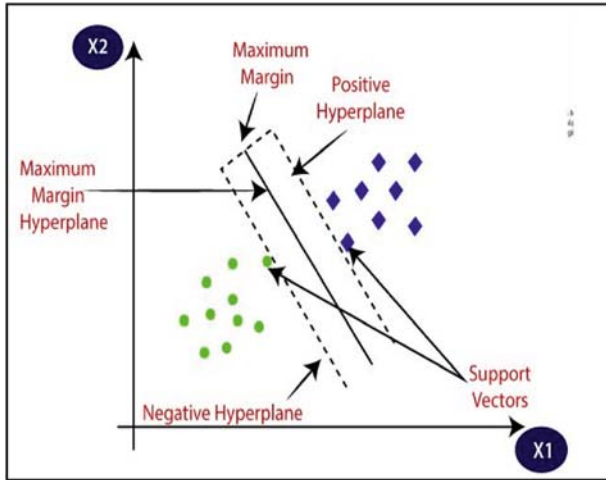
dfi = number of documents containing ith word.

N = total number of documents.

4.7. Classifier

For the empirical study, we are using two classifiers. Brief details about the algorithms are discussed in following subsections.

4.7.1. Naïve Bayes classification

Naive Bayes is the fastest request estimation, which is sensible for an awesome snippet of statistics. Naive Bayes classifier is absolutely utilized in numerous applications, as an example, unsolicited mail separating, message accumulating, assessment, and recommender systems. It includes Bayes' theory of chance for estimates for the dark class [11]. The request has two phases, a getting to know level, and the assessment degree. The classifier prepares its model on a given dataset at the mastery level, and in the evaluation level, it assesses how well the classifier is working. Precision, error, correctness, and audit ability are just a few of the characteristics that are taken into consideration while evaluating performance.

Naive Bayes classifier expects that the impact of a particular component in a category is liberated from various components. As an instance, a credit up-and-comer is captivating or now not depending upon his/her repayment, past improvement and alternate history, age, and vicinity.

Whether or not these components are structured, those features are at this factor contemplated uninhibitedly. This assumption that is known as class contingent possibility, where P(h) denotes a high likelihood of speculation (irrespective of what the information). P(D) is the probability of the data (no matter what the hypothesis). P(h|D) is the possibility of principle h given the information D. That is referred to as returned probability and P(D|h) denotes the probability of data D for the reason that the speculation h changed into legitimate.

4.7.2. Support vector machine (SVM)

For grouping and relapse issues, support vector machine (SVM) is a general managed learning strategy. The motivation behind the SVM calculation is to decide the ideal line or choice limit for characterizing N-layered space into classes with the goal that new information focuses can be effectively positioned in fitting areas. Hyper plane is the most ideal decision line. The extreme point/vector that contribute to the formation of the hyper plane are selected using SVM.

5. Results and analysis

The end result can be proven in a pie chart representing the share of hashtags with positive, negative and zero sentiment. Null hash tags represent hash tags that have been given a value of zero. SVM and Naive Bayes algorithm are used for result prediction. SVC(C = 1.0, gamma = "scale") linear kernel is used for SVM, and multinomial Naive Bayes is used for Naive Bayes. Support Vector Machine provides more accurate results than the Naive Bayes technique in our research. Then we calculate the accuracy the usage of the Naïve Bayes and Support Vector Machine Algorithm. Accuracy value of Naive Bayes is 0.73 and SVM is 0.79. Subsequent we compare the emotion of a textual content and the emojis present in that textual content. Maximum emotions in a text and its emojis are the same and a few emotions in a textual content and its emojis are different (see Table 1).

5.1. Results for BOW model

The result of BOW model is represented as Table 2. We construct a vector, which represents whether a word in each sentence is a frequent phrase or not. If a word in a sentence is a frequent phrase, we set it as 1; else we set it as 0.

Table 1. Emoji lexicon sample Entry.

id	user_description	user_location	text	user_name	user_created	user_followers	created	retweet_count	polarity	subjectivity
1	T H I C C latinas only	Trumbull, CT	â†™ #unpopularopinion #lol The inevitability of peace i	plive_calmer	2017-03-11T09:37:24	16	2019-12-07T14:13:54	0	0.225	0.35833333
2	parksosh	FA - RANDOM	RT @smolkjd: are we surprised he's trending on 3 differ	dyompolin	2018-12-16T12:38:58	1137	2019-12-07T14:13:54	0	-0.09107	0.75892857
3	â˘¯¯¡ People,Pets,Art	Tucson	RT @janewells: Jury foreman told us Unsworth did not g	sandi_childs	2013-11-16T11:53:08	4100	2019-12-07T14:13:54	0	0	1
4	Love to express my d	Islamabad, Pakista	@summayasaleem close call :p	iamwakas	2011-04-19T18:08:24	572	2019-12-07T14:13:54	0	0.75	1
5	[Roleplayer] What are you trying to fin	Bubu, @johnnysoeh meet my ooglie friends. They are a	peppermxnts	2019-11-30T15:00:58	13	2019-12-07T14:13:54	0	0.05	0.8875
6	http://GoldenRation	Around.	@madelyndewinter @NewYorker @hulu I donâ€™t hav	NautilusShell	2009-09-08T13:08:55	538	2019-12-07T14:13:54	0	-0.75	1
7	â'€à¸›à'‡à¸™à¸²à¸™à¸²		RT @GOT7Official: I GOT7 6th Generation Photo Teaser	_Support_YOU	2019-03-03T11:46:54	18	2019-12-07T14:13:54	0	0	0
8	â€ð—™ð—œð—â—ð—·	US â€¢ JPN â€¢ KR â	@Lusinners message â€" Arcene.	rxndzvous	2019-07-16T15:03:44	108	2019-12-07T14:13:54	0	0	0
9	DAY6 ðŸŽ˜˜ BTSðŸ'£˜ iKONðŸ à¸à¸±à¸ à	RT @caleridula: Jae : life is really hard sometimes, and i	lammyday6	2018-01-03T04:39:31	34	2019-12-07T14:13:54	0	0.252778	0.79722222
10	i†Œ€‰‰/ëˆë°œiŠ뜮Ž	ë,©ï¿¡¨ 124589ë°²	RT @An_Rock_Sam: Third Krita Practice. I still like her.	EVA_Mk_1	2015-02-03T12:05:33	119	2019-12-07T14:13:54	0	0.25	0.5
11	Thalapathy Vijay loy	Chennai	ðŸŽ‰Active Thalapathy Fans Followers InvitedðŸ'¥	Itz_Starker	2017-04-22T06:24:39	889	2019-12-07T14:13:54	0	0	0
12		California, USA	RT @atensnut: To schedule a speaking - book signing	Stancy11382013	2018-08-18T17:45:05	1740	2019-12-07T14:13:55	0	0	0
13	Fashion . Radio Pers	Harvey, IL	RT @therealblair: Pt2 yo need a Grammy ðŸ'¸(he only 1	dsymoneallure	2015-01-13T07:27:58	101	2019-12-07T14:13:54	0	0	1
14	My sunshine Double B #BamBam , 2014.	RT @BubblebamB: ðŸ'½191207	bias_BamBamonly	2014-08-26T00:00:12	283	2019-12-07T14:13:55	0	0	0

5.2. Result of TF-IDF model

Here, we generate a term frequency matrix as shown in Table 3, in which the rows represent documents and the columns represent unique terms used in each document. Find every instance of a word in each text.

In the second step, use the formula to calculate the inverse document frequency (IDF). Finally, TF matrix and IDF are multiplied in turn. The text is now

Table 2. Result of BOW Model.

	32	33	34	35	36	37	38	39	40
251	0	0	0	0	0	0	0	0	0
252	0	0	0	0	0	0	0	0	0
253	0	0	0	0	0	0	0	0	0
254	0	0	0	0	0	0	0	0	0
255	1	0	0	0	0	0	1	0	0
256	0	0	0	1	0	0	0	0	0
257	0	0	0	0	0	0	0	0	0
258	0	0	0	0	0	0	0	0	0
259	0	0	0	0	0	0	0	0	0
260	0	0	0	0	0	1	0	0	0
261	0	0	0	0	0	0	0	0	0
262	0	0	0	0	0	0	0	0	0
263	0	0	0	0	0	0	0	0	0
264	0	0	0	0	0	0	0	0	0
265	0	0	0	0	0	0	0	0	0
266	0	0	0	0	0	0	0	0	1

Table 3. Result of TF-IDF model.

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0	0	0	0	0	0	0	0	0	0	0.0869361	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0.0314157
2	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0.0781397	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0.0342694	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0.0917886	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0.13075	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0.0341868	0	0.030967	0	0	0	0	0
14	0	0	0.0949551	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0.103016	0	0	0

prepared to be fed into an algorithm for machine learning.

5.3. Confusion matrices for NB classifier

We generate a confusion matrix for our data in NB classifier as shown in Table 4.

The first column of the dataset has the value of (55 + 12) = 67. The dataset's second column's total column values are (42 + 91) = 133. Predicting the first column as the second column resulted in more errors than the other way around.

5.4. Confusion matrices for SVM classifier

We generate a confusion matrix for our data in SVM classifier as shown in Table 5.

Table 4. Confusion matrix for Naïve Bays.

	0	1
0	55	42
1	12	91

Table 5. Confusion matrix for SVM.

	0	1
0	87	16
1	22	57

The first column of the dataset has the value sum (87 + 22) = 109. The dataset's second column's total column values are (16 + 57) = 73.

5.5. Accuracy of SVM and Naïve Bays algorithm

The accuracy is calculated as (TP + TN)/ (TP + TN + FP + FN).

Table 6 compares the outcomes of the two applied algorithms. SVM accuracy is 79.12, compared to Naive Bayes accuracy of 73. SVM's precision and recall values are 78.08 and 72.15, while Naive Bayes' values are 68.42 and 88.34.

5.6. Outcomes of analyzing the sentiment of tweets

5.6.1. Positive events with emoji

We picked the "FIFA World Cup" events to concentrate on the utilization of emoticon characters in event with positive feelings since individuals get positive sentiments when they welcome the "FIFA World Cup". The hashtag "#fifa" was used to collect about 1000 tweets.

Table 6. Comparative chart of algorithm accuracy.

	SVM	Naïve Bayes
Accuracy	79.12	73
Precision	78.08	68.42
Recall	72.15	88.34
F1 Measure	74.99	77.11

Fig. 7. Pie Chart of the analyzed tweets.

Fig. 9. Result for compare the emotions of a text & its emojis which is present in that text.
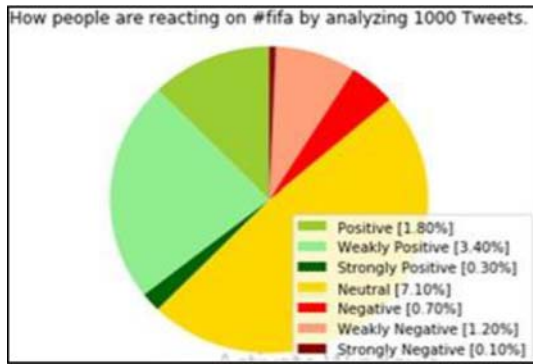
We get a General Report-weakly positive from this Pie Chart. The pie charts each percentage positive, weakly positive, strong positive, negative, weakly negative, strong negative and null sentiment hash tags are shown in a different color in Fig. 7.

From the pie chart, it is clear that 1.80% of people thought it was positive and 3.40% of people thought it was weakly positive. Again 0.30% of people thought it was strongly positive.

Contrary to the pie chart, 0.70% thought it was negative and 1.20% thought it was weakly negative. Again 0.10% of people thought it was strongly negative. In the end, 7.10% of people thought it was neutral.

5.6.2. Negative events with emoji

We picked the "NRC in India" events to concentrate on the utilization of emoticon characters in event with positive feelings since individuals get negative sentiments when they welcome the "NRC in India". The hashtag "#NRC" was used to collect about 1000 tweets.

From this Pie Chart we get General Report-weakly Negative as shown in Fig. 8.
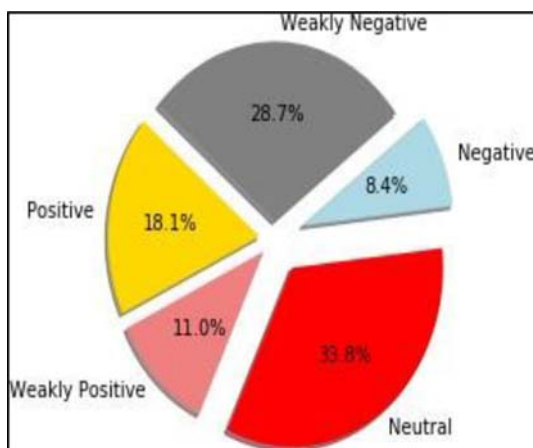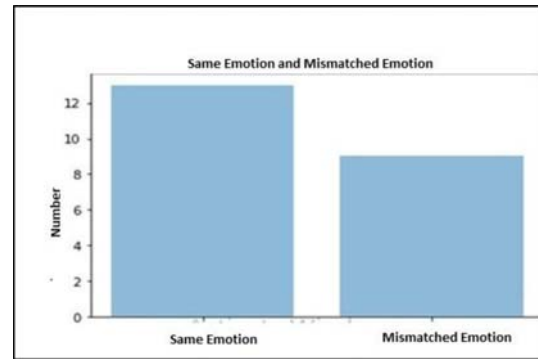

Fig. 8. Pie Chart of the analyzed tweets.

The pie chart depicts each percentage positive, weakly positive, negative, weakly negative, and zero sentiment hash tags in a different color, as shown in Fig. 8. From the above pie chart, it is clear that 28.7% of people thought it was weakly negative and 8.4% of people thought it was negative. Contrary to the pie chart, 18.1% thought it was positive and 11.0% thought it was weakly positive. In the end 33.8% of people thought it were neutral. We have discussed two pie charts based on two topics, one is "NRC in India" and the other is "FIFA". From Fig. 7, we get the appearance of positive sentiment rather than negative sentiment. Thus the analysis leads to positive sentiment. From Fig. 8, we get the feedback of negative sentiment rather than positive sentiment and thus lead to negative sentiment. For neutrals, this represents zero total tweets/hashtag's.

5.7. Result and analysis

In Fig. 9, bar chart represents the emotion of the text and its emojis for same emotion and some cases where the emotion of the text and its emojis are different. The numbers of identical emotions as compare to mismatched emotion are in the ratio of 12.5:8.

6. Conclusion

In this paper we analyzed the utilization of emoticon characters on informal communities especially in social media. The impacts of emoticon on text mining and correlation between them are empirically studied. We've seen a few worldwide positive and negative trending topics to check whether there is a distinction between the emotions between emoticons and related text.

We found that utilizing emoticons while dissecting feelings further develops the general opinion score. Although emoji characters can be used to express both negative and positive emotions, it is studied that emojicon in emotion analysis improves the

expression of positive opinions and overall emotional score when compared to negative opinions. Using the crawler, tweet is collected from the Tweeter. The tweet is then retrieved from the Twitter information that goes through feature removal.

Sentiment Analysis techniques have been there for more than ten years, and organizations are now using them as a significant tool for crucial advanced arranging and moving. This move is additionally because of the progression in information stockpiling, access and examination empowered through enormous information structures. Emoji analysis can provide a fresh perspective to the well-known field of research that is text sentiment analysis. Emojis are frequently used with text to emphasize or counterpoint context. For a comprehensive interpretation of the emotion on online social networks, it is crucial to relate text and emoji.

Disclosure statement

The authors report there are no competing interests to declare.

References

[1] Chandra Yogesh, Jana Antoreep. Sentiment analysis using machine learning and deep learning. In: 7th international conference on computing for sustainable global development (INDIACom); 2020.

[2] Kasture N, Bhilare P. An Approach for Sentiment analysis on social networking sites. Computing Communication Control and Automation (ICCUBEA); 2015. p. 390—5.

[3] Bhuta S, Doshi A, Doshi U, Narvekar M. Are view of techniques for sentiment analysis of twitter data, issues and challenges in intelligent computing techniques (ICICT). 2014. p. 583—91.

[4] Subhashini L, Li Y, Zhang J, Atukorale AS, Wu Y. Mining and classifying customer reviews: a survey. Artif Intell Rev 2021;54:6343—89.

[5] Mowlaei ME, Abadeh MS, Keshavarz H. Aspect-based sentiment analysis using adaptive aspect based lexicons. Expert Syst Appl 2020;148:113234.

[6] Naresh Kumar KE, Uma V. Intelligent sentinet-based lexicon for context-aware sentiment analysis: optimized neural network for sentiment classification on social media. J Supercomp 2021;77:12801—25.

[7] Zvarevashe K, Olugbara OO. A framework for sentiment analysis with opinion mining of hotel reviews. In: 2018 Conference on information communications technology and society (ICTAS). IEEE; 2018. p. 1—4.

[8] Wankhade Mayur, Sekhara Rao Annavarapu Chandra, Kulkarni Chaitanya. A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review; 2022.

[9] Rayala Vinod Kumar CH, Lalitha Bhaskari D, Srinivasa Rao P. "VLS Algorithm: A new approach to sentiment analysis". Int J Technol Res MAY 2020;9(Issue 05):2277—8616. ISSN.

[10] Kralj Novak P, Smailović J, Sluban B, Mozetič I. Sentiment of emojis. PLoS One 2015;10(12).

[11] Bagheri A, Saraee M, de Jong F. Care more about customers: unsupervised domain independent aspect detection for sentiment analysis of customer reviews. Knowl Base Syst 2013;52:201—13.

[12] Fredriksen Valerij, Jahren BrageEkroll. Twitter sentiment analysis - exploring automatic creation of sentiment lexica. NTNU; 2016.