The Korean Institute of Communications and Information Sciences

Available online at www.sciencedirect.com

ScienceDirect

ICT Express 6 (2020) 357–360

www.elsevier.com/locate/icte

An algorithm and method for sentiment analysis using the text and emoticon

Mohammad Aman Ullah*, Syeda Maliha Marium, Shamim Ara Begum, Nibadita Saha Dipa

Department of Computer Science and Engineering, International Islamic University Chittagong (IIUC), Chittagong 4318, Bangladesh

Received 2 February 2020; received in revised form 15 July 2020; accepted 16 July 2020
Available online 25 July 2020

Abstract

People nowadays use emoticons in their text increasingly in order to express their feelings or recapitulate their words. Earlier machine learning techniques only involve the classification of text, emoticons or images solely where emoticons with text have always been neglected, thus ignored lots of emotions. This research proposed an algorithm and method for sentiment analysis using both text and emoticon. In this work, both modes of data were analyzed in combined and separately with both machine learning and deep learning algorithms for finding sentiments from twitter based airline data using several features such as TF–IDF, Bag of words, N-gram, and emoticon lexicons. This research demonstrates that whenever emoticons are used, their associated sentiment dominates the sentiment conveyed by textual data analysis. Also, deep learning algorithms are found to be better than machine learning algorithms.

Keywords: Machine learning; Deep learning; Classification; Emoticon; Text

1. Introduction

The age of getting meaningful insights from social media data has now arrived with the advance in technology. Traditionally, analysis of sentiment has been done on text, but a large amount of data is now being uploaded as reviews, images, emoticons, and videos. By inspecting these data, the sentiment of the public toward a specific matter can be analyzed, examined and discovered [1,2]. Over the years, people considered emoticons as a medium of communication that is used in texts or solely to dedicate one's sentiment in an efficient manner. Emoticons are symbolic expressions consisting this type of tokens such as \:", \=", \-", \)", or \(" and commonly represent facial expressions. Emoticons can be read either sideways, like \:-("(a sad face), like \(^^)" (a happy face) and with different order to recognize the presentation [3].

Monitoring these symbols of emoticons along with the text is immensely necessary to get the actual sentiments like, happiness, frustration, anger, sadness, etc. which are then classified in positive, negative and neutral. The general researcher of sentiment analysis (SA) deals with either text or emoticons. Maximum researches in SA from social media data have been executed on text using machine learning (ML) algorithms [4]. However, SA on both text and emoticon has been mostly ignored due to the lack of resources and complexity of emoticons. Text analysis is one of the prominent researchable areas as it extracts sentiment using different ML and deep learning (DL) techniques with the help of recent technologies [5,6]. But, research shows DL was rarely applied in SA on text and emoticon data combination. Therefore, this research has analyzed the text and emoticons separately and in combination to find the sentiments. In addition, the research developed an emoticon lexicon, analyzed the sentiments by applying emoticon lexicons along with some text features such as TF–IDF, bag-of-words, and n-gram using both ML and DL algorithms.

The paper includes the following topics such as related works in Section 2. In Section 3, problem statement and objectives are described. Section 4 explains the methods and algorithms of this research. Section 5 gives summary results and corresponding discussion. A comparison with the existing system is discussed in Section 6. Finally, conclusion and future works are included in Section 7.

2. Related works

Modern sentiment analysis researches are now on different domains and languages [7]. In [8], a classification of

[*] Corresponding author.
E-mail address: amanullah@cse.iiuc.ac.bd (M.A. Ullah).

Peer review under responsibility of The Korean Institute of Communications and Information Sciences (KICS).

https://doi.org/10.1016/j.icte.2020.07.003

tweets data was considered with the environment using the unsupervised learning methods. The sentiment lexicon derives different kinds of emoticons with the use of emoji-based ideograms. A limited number of standard emoticons were detected from twitter data and found the corresponding sentiments. In [6] an ensemble sentiment classification technique was applied with the help of different classification methods like Naive Bayes (NB), SVM, Decision Tree, and Random Forest (RF) algorithms. They experimented with six classification techniques and the given ensemble approach was trained and tested using dataset of 12 864 tweets with 10 fold evaluation. The performance of the proposed ensemble approach gave an output of these individual classifiers in the airline service Twitter dataset which could improve the overall accuracy.

The sentiment of Twitter messages was pre-processed (cleaning and stemming) and extracted by exhibiting results of a NB algorithm using R and Rapid Miner and categorized in neutral, negative and positive sentiments and finally summarized the results in [9]. A total of 1 298 395 tweets on United Airlines controversy were used in this research. The researchers in [10] have classified microblogging instances and determine positive, negative and neutral sentiments. The main contribution is the automatic collection of corpus and sentiment analysis. Their proposed techniques are said to be efficient and perform better than previously proposed methods. In [11], the authors assessed 2080 Dutch tweets and forum messages, which all have been manually annotated for sentiment and contain emoticons. In this corpus, paragraph-level accounting for sentiment implied by emoticons remarkably surpasses sentiment classification accuracy. This indicated that emoticons association dominates the sentiment conveyed by textual cues and forms a good proxy for intended sentiment whenever emoticons were used. For detecting the sentiment of Twitter messages they investigated the utility of linguistic features. They examined the usefulness of existing lexical resources also the features that capture information about the informal and creative language noticed in microblogging.

3. Problem statement and objectives

Sentiment analysis is ever-growing sub-field of natural language processing. There are many research works that have addressed the problems of SA from text, emoticon, images, and audio or videos separately [1,6]. Very few researches have been done on emoticon for finding sentiments [3]. Moreover, related works section shows there is a scope for further extension in the field of SA with both text and emoticon. Therefore, the objectives of this research are:

- ✓ Sentiment analysis using bi-mode (text and emoticons) on social media data.
- ✓ Developing emoticon lexicon
- ✓ To improve the classification accuracy of sentiment analysis by using both ML and DL algorithms.

4. Methodology

Many corpora in different languages are available online for SA purposes. The data for this research was collected from different comments on twitter airline reviews. The dataset contains 14 460 reviews among which 2363 are positive, 3099 are neutral and 9178 are negative reviews. In addition, the dataset contains 732 emoticons among which 220 are classified positive, 78 neutral and 434 are negative. We have also created and used an emoticon lexicon which contains different types of emotions that are vastly used on social media. The data were then processed to help enhance the effectiveness of SA. This step as well as total analysis has taken place in two steps. In the first step, the data were tokenized, and the stop words, URL's, and digits were removed, but we did not remove punctuation and emoticons. This pre-processed data were used in finding sentiments with emoticons. For finding the sentiments, important features such as TF–IDF, Bag-of-Words, n-grams, emoticon lexicons were extracted and selected. In the second step, punctuations and emoticons were removed and sentiments were calculated without emoticon and using the same features except emoticons.

In the case of combined text and emoticon data, the sentiment was analyzed using some ML and DL algorithms such as support vector machine (SVM), NB, RF, logistic regression (LR), Long short term memory (LSTM) and convolutional neural network (CNN). For ML model, the study has split the dataset into two parts, where, 80% was used as training, and 20% was used as validation dataset. The study has used stratified 10-fold cross validation to estimate the accuracy. The dataset was split into 10 parts, where, 9 parts were train and 1 was used as test. The steps were continued for all train–test splits combinations.

In the case of LSTM, this study worked first on embedding layers and the further works on LSTM cells which consists of recurrent connections to the network so that information about the sequence of words can be concluded in the data. To deal with both short and long reviews, all reviews were padded or truncated to a specific length; where, length is defined by sequence length (number of steps for LSTM layer). Besides, training, validation, testing have been done by creating data loaders, data iterator for this data. The study has used generator function for batching the data. Finally, LSTM cells go to a sigmoid output layer and gave the accuracy. In the case of CNN, all the training sentences were padded and the embeddings matrix was passed to embedding layer. Five different filter sizes are applied to each comment, and GlobalMaxPooling1D layers were applied to each layer. All the outputs were then concatenated. A dropout layer dense the dropout and final dense layer is applied. As the training is done on small data set, in just a few epochs out model over fitted it and with just three iterations the accuracy is achieved.

The proposed algorithm starts with pre-processing the airline input tweet data by tokenizing, stop words, URL, and digit removing and saving the data in the file 'f1_fil'. This file is further pre-processed to remove the punctuations and emoticon and is saved in file 'f2_fil'. The algorithm then extracts the features such as TF–IDF, Bag of words, n-gram, and emoticons

Table 1
Comparison of results in different modes and algorithms.

ML or DL algorithms	Text & emoticon				Only text			
	Precision	Recall	F-Score	Accuracy	Precision	Recall	F-Score	Accuracy
SVM	0.73	0.71	0.74	0.78	0.74	0.69	0.71	0.78
Random Forest	0.73	0.66	0.68	0.76	0.71	0.63	0.65	0.75
Naive Bayes	0.59	0.57	0.60	0.52	0.59	0.59	0.77	0.64
Logistic Regression	0.74	0.66	0.68	0.78	0.76	0.68	0.71	0.77
LSTM	0.85	0.89	0.86	0.89	0.84	0.86	0.87	0.88
CNN	0.81	0.83	0.82	0.81	0.78	0.75	0.79	0.79

(using emoticon lexicon) from file ‘f1_fil’ and same features except emoticon from ‘f2_fil’. For each feature from F1 and F2, the ML (NB, SVM, RF, LR) and DL (LSTM, CNN) algorithms are applied and their scores are recorded. Finally, it compares the best ML result with DL results and selects the best one. The total processing and analysis were conducted in python (see Fig. 1).

Algorithm: SA with and without emoticon

Input: Airline Review Dataset
Output: Sentiment (Positive/Negative)
Notation: Feature→F, Machine learning→ML, Deep Learning→DL, TF-IDF→TI, Bag of words→BOW, n-gram→ng, and emoticons→E

```
1.  Begin:
2.  f1_fil = tokenize, stop words, url, and digit removing
    f2_fil = remove punctuation and E from 'f1_fil'
3.  F1←Extract the features TI, BOW, ng, and E from f1_fil
    F2←Extract the features TI, BOW, ng,from f2_fil
4.  Find the score by ML and DL algorithms
      for each F1 in f1_fil do
        while i=1…n do
          while j=1…n do
              ML_score [i][j]= score applying ML;
              DL_score [i][j]= score applying DL;
          end while
        end while
      end for
      for each F2 in f2_fil do
        while i=1…n do
          while j=1…n do
              ML_score1 [i][j]= score applying ML;
              DL_score1 [i][j]= score applying DL;
          end while
        end while
      end for
5.  Compare the result of ML and DL algorithms obtained for
    both 'f1_fil' and 'f2_fil' and select the best one
6.  end
```

5. Results and discussions

In this research, the experiments were conducted on two modes (text, and emoticon) and two ways (only text, and text & emoticon). The data in later ways were analyzed using both ML and DL algorithms, but, the textual data were analyzed only with ML algorithms. The experimentations were conducted using the python programming language. The results obtained from the experiments are shown in Tables 1 to 3 and in Fig. 2. Table 1 shows the comparison between the results obtained using different modes and algorithms. It is evident that, analyzing text & emoticon at a time with the adaptation of emoticon lexicons and other features as mentioned earlier is better than analyzing only text. But, both the ways have very little differences of 1–3% on the scale of performance metrics. The classification accuracy of both SVM and LR prevails to be better than those of the other two ML classifiers such as NB and RF. Overall, the performance of NB was found to be very poor for this analysis. Whereas, in the case of DL, LSTM shows the better classification accuracy than CNN. In general, DL algorithms are outperforming ML algorithms. DL achieved higher classification accuracy as it solved the problem end to end and automatically extracts features.

Table 2
Comparison between machine and deep learning on accuracy.

	Algorithms	Accuracy
Machine Learning (ML)	SVM	78%
	Logistic Regression	78%
	Random Forest	76%
	Naïve Bayes	52%
Deep Learning (DL)	LSTM	89%
	CNN	81%

The result in Table 2 and Fig. 2 reveals that, DL algorithms such as LSTM and CNN are performing better than all ML algorithms with 89% and 81% accuracy respectively in the case of combined text and emoticon data. These two algorithms are also working better when dealing with only text. The differences in accuracy are 1–9%. The overfitting issue placed by these models is resolved by adopting different hyper-parameter tuning techniques such as reduce the network’s capacity (RNC), regularization (R), and dropout layers (DO). However, very insignificant improvement is achieved. The model with RNC and DO starts overfitting later than the baseline model. However, in case of R model, overfitting starts in the same epoch as the baseline model.

6. Comparison with existing system

This study has conducted a review of many existing works and found very few of them are related to this research. Table 3 shows a comparison between the existing and proposed systems. The existing study [1] shows that, with textual data, they have achieved 57% accuracy, but, this study achieves 78% accuracy which is far better than the existing work. With combined textual and emoticon data, the accuracy of 89% has been obtained by this study as compared to 65% and 84% respectively in the existing study [6,12].

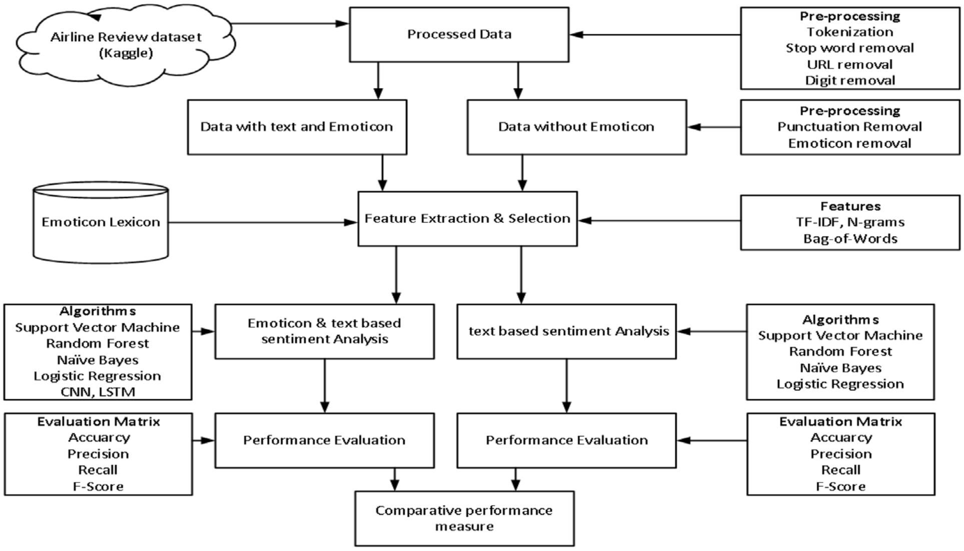

Fig. 1. Proposed methodology.

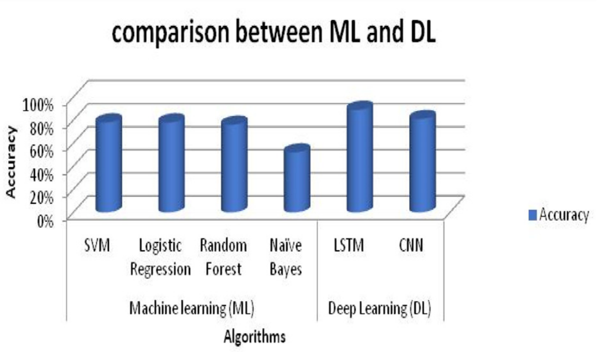

Fig. 2. Comparison between machine and deep learning.

Table 3
Comparison between the existing and the proposed system.

Sentiment analysis mode	Existing system	Proposed system
Text [1]	57%	78%
Text and emoticon [6,12]	65%, 84%	89%

7. Conclusion and future directions

This paper mainly contributes an algorithm, a method, and the emoticon lexicon for analyzing sentiments of social media data (both textual and emoticons); for instance airline data collected from twitter. This research also represents the impact of considering emoticons along with the text while analyzing sentiments. The study was conducted using both ML and DL algorithms. The proposed system applied several features and models on the collected reviews based on text and emoticons to determine the sentiments. The overall result indicates that, considering emoticon along with text has a positive impact on the sentiment analysis. It is also found that, DL algorithms are working better than ML algorithms. Finally, this research concluded by outperforming existing researches. In the future, this research could be extended to the field of multilingual data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] P. Chikersal, S. Poria, E. Cambria, A. Gelbukh, C.E. Siong, Modelling public sentiment in twitter: using linguistic patterns to enhance supervised learning, in: International Conference on Intelligent Text Processing and Computational Linguistics, 2015, pp. 49–65.

[2] K. Kang, C. Yoon, E.Y. Kim, Identifying depressive users in twitter using multimodal analysis, in: 2016 International Conference on Big Data and Smart Computing (BigComp), 2016, pp. 231–238.

[3] A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, U. Kaymak, Exploiting emoticons in sentiment analysis, in: Proceedings of the 28th Annual ACM Symposium on Applied Computing, 2013, pp. 703–710.

[4] P. Yadav, D. Pandya, SentiReview: Sentiment analysis based on text and emoticons, in: 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2017, pp. 467–472.

[5] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, V. Stoyanov, SemEval-2016 task 4: Sentiment analysis in twitter, 2019, arXiv preprint arXiv:1912.01973.

[6] Y. Wan, Q. Gao, An ensemble sentiment classification system of twitter data for airline services analysis, in: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 2015, pp. 1318–1325.

[7] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, et al., Semeval-2016 task 5: Aspect based sentiment analysis, in: 10th International Workshop on Semantic Evaluation (SemEval 2016), 2016.

[8] W. Wolny, Emotion analysis of twitter data that use emoticons and emoji ideograms, 2016.

[9] D.D. Das, S. Sharma, S. Natani, N. Khare, B. Singh, Sentimental analysis for airline twitter data, in: IOP Conference Series: Materials Science and Engineering, 2017, 042067.

[10] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in: LREc, 2010, pp. 1320–1326.

[11] E. Kouloumpis, T. Wilson, J. Moore, Twitter sentiment analysis: The good the bad and the omg!, in: Fifth International AAAI Conference on Weblogs and Social Media, 2011.

[12] Y. Wang, M. Huang, X. Zhu, L. Zhao, Attention-based LSTM for aspect-level sentiment classification, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 606–615.