Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 154 (2019) 332–340

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Congress of Information and Communication Technology, ICICT 2019

Implementation of Machine Learning Algorithms in Arabic Sentiment Analysis Using N-Gram Features

Donia Gamal *, Marco Alfonse, El-Sayed M. El-Horbaty, Abdel-Badeeh M.Salem

Computer Science Department, Faculty of computer and information sciences, Ain Shams University Cairo, Egypt

Abstract

Sentiment analysis (SA) is a scholarly process of extricating and classifying individuals' emotions and feedbacks expressed in source text content. It is one of the pursued subfields of Computational Linguistics (CL) and Natural Language Processing (NLP). The evolution of social media based applications has generated a big amount of personalized reviews of different related information on the Web in the form of tweets, status updates, and many others. Several approaches have come into the spotlight in recent years to accomplish SA, the most part of SA researches have been applied utilizing the English language. SA in Arabic online social media may be slacking behind commonly because of the difficulties with handling the morphologically complex Arabic natural language and the lack and absence of accessible tools and assets for extracting Arabic opinions from the text. This research is aimed to analyze the collected twitter posts in different Arabic Dialects and a comparison between the various algorithms used for SA with various n-gram as a feature extraction method. The measurement of the performance of different algorithms is evaluated in terms of recall, precision, f-measure, and accuracy. The experiment results show that unigram with Passive Aggressive (PA) or Ridge Regression (RR) gives the highest accuracy 99.96 %.

Keywords: Arabic Dialect Sentiment Analysis; Sentiment Classification; Twitter; Opinion Mining; Machine Learning; Applied Informatics; N-gram.

* *E-mail address:* donia.gamaleldin@cis.asu.edu.eg

10.1016/j.procs.2019.06.048

1. Introduction

Sentiment analysis, also called Opinion Mining (OM) is the field that investigates and analyzes individuals' reactions and responses towards an entity (e.g. Blogs, movies, products, DVD, books…) utilizing text analysis algorithms to determine individual textual attitude1. SA acts like an effective and powerful tool for individuals to extricate the essential information, also to aggregate and mixture the collective sentiments of the reviews. Utilizing SA, variances in stock prices could be predicted2, political election race preferences can be observed closely3, and even groups' interactions could be observed and followed which provides many advantages and benefits4. As individuals, there is always a tendency to consult close friends and relatives about items before purchasing them. Organizations are constantly eager to accumulate public preferences to attract higher numbers of clients. This is normally done by distributing surveys ahead of a diversified group of people. Software engineers and business intelligence researchers have stated these problems beginning from 19585, and attempted to design the proper algorithms that can acquire opinions evaluations about a specific entity and examine them in order to constitute these reviews in a clean and a quick manner.

OM could be seen as a classification task that decides whether a specific document or textual content was written to express a positive or negative sentiment. Sentiment Classification (SC) could be useful in commercial enterprise intelligent applications and recommender frameworks6. There are likewise potential applications for filtering messages7. Researchers have proposed several approaches for SA. In general, there are two principle approaches, the primary one is utilizing Machine Learning (ML) algorithms, which are presented in this paper, and the other one is utilizing Lexicon Based (LB) approach works with respect to that contextual sentiment orientation is the total of the opinion orientation of every word or phrase accessible and available8.

Interaction and communication through online social networking have become progressively popular and well known in the whole world as they see it as an imperative and essential tool for sharing a different point of view and information about various subjects besides expressing opinions uninhibitedly and openly9. This research focuses on the Middle East region and according to the Arab Social Media Report in 2017, Twitter had averaged over 849.1 million tweets produced on a monthly basis10.

This paper documents the comprehensive study between various ML algorithms with various values of N-gram and using different dialects of Arabic tweets as a dataset for determining the polarity of each tweet's sentiment.

The rest of the paper is organized as follows: Section 2 introduces the related work of Arabic SA; the proposed methodology is explained in Section 3. Section 4 shows the experiments and results. Discussion of these results is illustrated in Section 5 and Section 6 includes the conclusions and future work of the research.

2. Related Work

This section affords preceding work that deals with SA. The listed works below differentiate in their preprocessing steps, analysis methods, and the structure of opinions and reviews. They are also different in the language of the data.

Omar11, carried out a comparative study concerned with the effectiveness and adequacy of using individual supervised ML classifiers and ensemble algorithms for SA of Arabic customers' feedbacks. The ensemble method was implemented to SC task, pointing at accommodating classification algorithms proficiently and efficiently to formulate and figure a classification algorithm more precise and accurate than the others. They applied the three common text classification algorithms which are Bernoulli Naive Bayes (BNB), Rocchio classifier and Support Vector Machines (SVM). Second, they made a comparative study of two different kinds of ensemble methods which are the voting and meta-classifier combinations. Despite, the experiment results of individual ML classifiers demonstrated that BNB and SVM algorithms outperformed better than other methods. The results also demonstrated that ensemble of classification ML algorithms with meta-learner ensemble algorithm achieved strongly better than all the other individual ML classifier.

Hamouda and Akaichi12 have explored the utility of SC on a novel gathered Tunisian Facebook updates dataset. Utilizing the majority well-known ML algorithms, they carried out a comparative study between the SVM and the Naïve Bayes (NB) algorithms by mixing different N-gram feature extraction. Those algorithms can accomplish high accuracy for classifying opinions when consolidating variant features. The overall highest accuracies of the

proposed methodology are 74.05% for using unigram and bigram with NB and 75.31% while using unigram and bigram with SVM.

Zainuddin and Selamat13 presented the experimental results of performing the SVM on two different label datasets which contain 2000 movie reviews that are used in pang and lee14, and 400 alternative positive and negative reviews of Simon Fraser University (SFU) review datasets utilized in Taboada, et al 15 for their experiments. The N-gram and diverse weighting scheme were utilized to extract the most traditional features. Experimental results show that applying unigram outperforms other n-gram models for both datasets.

Hamdan, et al.16 applied three famous textual content classification algorithms and ran experiments and tests on the in-house created dataset. The three algorithms utilized are SVM, NB and Multilayer Perceptron-Neural Network (MLP-NN). The average measures acquired from performing the stated algorithms demonstrate that the SVM algorithm performs better than NB and MLP-NN. The precision results of SVM, NB and MLP-NN respectively are 77.8%, 75.4%, and 71.7%.

3. The Proposed Methodology

This section introduces a methodology for SA and explains how it handles the preprocessing and classification of Arabic tweets. The methodology presented in Fig. 1, consists of many stages which are: collecting data source, preprocessing, generating various levels of n-gram features, training model using different ML classifiers, and evaluating the performance of ML algorithms.

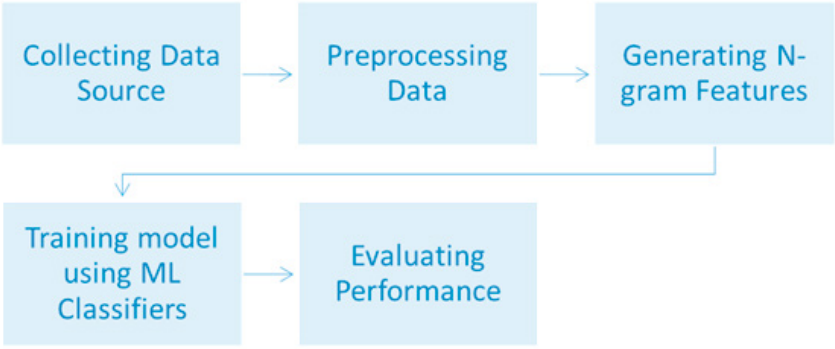

Fig. 1 the proposed methodology

3.1 Collecting data source

The polarity data set utilized is a set of tweet sentences, which have been collected and labeled automatically as a positive or a negative opinion. The tweet sentences are gathered using different Arabic dialects phrases in querying tweets using Tweepy Application Programming Interface (API)17. 438,931 positive and negative tweets are collected. In the crawled tweets, it is found that, there are sarcastic and neutral tweets in between alongside with the certain positive and negative tweets. A tweet is dismissed if it is not obviously adjusted toward a positive or a negative opinion. For the binary classification task, to avoid the imbalanced class distribution, 75,774 positive tweets and 75,774 negative tweets are selected from the dataset.

3.2 Preprocessing data

Many studies and researches revealed and discussed that pre-processing of textual content can enhance the performance of textual content classification18. The steps associated with data pre-processing are removing all noisy data in tweets such as hashtags, profile pictures, retweets, emoticons, user-names, user mentions, and URLs. The second step is tokenization, removing non-Arabic letters, removing diacritics, and normalizing Arabic analogous letters such as ‘أ’ to be ‘ا’ to decrease uncertainty and ambiguity. Then stop words are removed, such as ‘ان’, ‘كما’ and

‘قد’ etc. Finally, labeling the data is done by counting the number of positive phrases and negative phrases in the textual content of each tweet, and the tweet is labeled according to the higher count.

3.3 Generating N-gram features

N-gram as the feature extractor method is utilized. The N-gram features are frequently applied in textual content classification task19. The n-gram features could be divided into letter n-gram features and word n-gram features as illustrated in table 1. Unigram refers to n-gram of length 1, bigram refers to n-gram of length 2, trigram refers to n-gram of length 3, and so on.

Table 1 Example of N-gram

Sentence / N-gram	"خبر مبهج فعلا جدا "
Unigram	{ "خبر", "مبهج", "فعلا", "جدا" }
Bigram	{("خبر", "مبهج"), ("مبهج", "فعلا"), ("فعلا", "جدا")}
Trigram	{("خبر", "مبهج", "فعلا"), ("مبهج", "فعلا", "جدا")}

3.4 Training model using ML classifiers

SC algorithms can be generally divided into the ML approach, and the LB approach. This research paper focuses on ML approach. ML algorithms recreate the way people gain knowledge from their old experiences to obtain information and apply it in the future decisions making. These ML algorithms are broadly utilized in artificial intelligence and textual content classification. The classification by ML utilization may be summed up in two, Supervised ML (SML) and Un-Supervised ML (USML). SML is learning the model by utilizing the labeled dataset while the USML does not need a labeled dataset. Ten different classifiers are applied which are RR, PA, NB, SVM, BNB, Multinomial NB (MNB), Stochastic Gradient Decent (SGD), Logistic Regression (LR), Maximum Entropy (ME) and Adaptive Boosting (Ada-Boost). SciKit-Learn (SKLearn)20 and Natural Language Tool Kit (NLTK)21 have been used in developing these experiments. These libraries are loose and free which awards the programmer the ability to utilize different ML algorithms. The purpose of utilizing such algorithms lies on their powerful ability to manage the textual content categorization where the quantity of features is immense22.

3.5 Evaluating performance

The metrics utilized here for assessing and evaluating the classifiers’ performance are accuracy, precision, recall, and f-measure as defined in equations. (1)– (4) and Table 223.

Table 2 Confusion Matrix

	Predicted	
	Positive tweets	Negative tweets
Actual positive tweets	Number of True Positive Tweets (T_{PT})	Number of False Negative Tweets (F_{NT})

Actual negative tweets	Number of False Positive Tweets (F_{PT})	Number of True Negative Tweets (T_{NT})

The accuracy is described as the rate of correctly classified tweets. It is characterized by the formula (1).

$$Accuracy = \frac{T_{PT} + T_{NT}}{T_{PT} + T_{NT} + F_{PT} + F_{NT}} \quad (1)$$

The precision is the number of tweets correctly classified as its right class out of all the tweets classified as that class. It is measured for each class. Precision is characterized by the formula (2).

$$Precision = \frac{T_{PT}}{T_{PT} + F_{PT}} \quad (2)$$

The Recall expresses the number of correctly classified tweets of a class out of all tweets of that class. The formula for measuring the recall is characterized in (3).

$$Recall = \frac{T_{PT}}{T_{PT} + F_{NT}} \quad (3)$$

The standard F-measure also known as a balanced F-score (F1 score) is the harmonious mean of precision and recall. The formula for measuring the F-measure is characterized in (4).

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

4. Experiment Results and Discussion

One of the most standard and popular methods of validating ML classifiers is the K-fold cross validation where K is an integer value. A comparison of the performances using different feature sets on the ten different ML algorithms is done using the 10-fold cross validation. The comparative evaluation and analysis based on the acquired accuracy, precision, recall and f-measure results utilizing n-gram features are shown in the tables (2)-(5).

Table 3 Measurement of Accuracy

ML Classifier	Single Fold			10-Fold		
	N-gram					
	Unigram	Bigram	Trigram	Unigram	Bigram	Trigram
NB	95.22%	83.27%	66.24%	95.91%	82.90%	64.78%
BNB	98.00%	63.30%	55.29%	97.73%	65.10%	56.19%
MNB	98.19%	82.89%	66.22%	98.03%	82.37%	64.78%
ME	93.53%	79.41%	60.01%	94.18%	80.07%	60.84%
Ada-Boost	73.76%	59.20%	53.25%	72.95%	59.34%	53.38%
SVM	99.31%	77.83%	57.73%	98.86%	79.29%	58.61%

LR	98.96%	82.63%	62.36%	98.61%	82.63%	59.42%
SGD	99.11%	77.75%	65.93%	98.59%	76.61%	63.86%
RR	99.09%	77.87%	58.04%	99.95%	99.96%	98.84%
PA	99.15%	77.28%	58.27%	99.96%	99.96%	98.84%

Table 4 Measurement of Precision

ML Classifier	Single Fold			10-Fold		
	N-gram					
	Unigram	Bigram	Trigram	Unigram	Bigram	Trigram
NB	95.62%	84.18%	77.61%	96.15%	83.53%	76.28%
BNB	98.00%	78.02%	76.17%	97.73%	78.74%	76.41%
MNB	98.21%	83.36%	66.22%	98.03%	82.89%	76.17%
ME	93.93%	80.98%	74.27%	94.44%	81.53%	74.92%
Ada-Boost	78.95%	76.74%	75.53%	77.00%	76.52%	75.36%
SVM	99.32%	80.72%	75.92%	98.87%	81.72%	76.15%
LR	98.98%	83.85%	75.53%	98.61%	83.50%	75.88%
SGD	99.12%	83.14%	77.51%	98.59%	82.32%	77.39%
RR	99.09%	80.83%	75.85%	99.95%	99.96%	98.86%
PA	99.16%	80.45%	75.59%	99.96%	99.96%	98.86%

Table 5 Measurement of Recall

ML Classifier	Single Fold			10-Fold		
	N-gram					
	Unigram	Bigram	Trigram	Unigram	Bigram	Trigram
NB	95.22%	83.27%	66.22%	95.90%	82.94%	64.96%
BNB	98.00%	63.31%	55.30%	97.73%	65.11%	56.21%
MNB	98.18%	83.36%	66.21%	98.03%	82.40%	64.96%
ME	99.31%	79.41%	60.03%	94.18%	80.07%	60.86%
Ada-Boost	73.75%	59.18%	53.23%	72.95%	59.32%	53.35%
SVM	99.31%	77.84%	57.75%	98.86%	79.30%	58.63%
LR	98.97%	82.63%	58.57%	98.60%	82.63%	59.44%
SGD	99.11%	77.74%	65.91%	98.59%	76.60%	63.84%
RR	99.08%	77.88%	58.07%	99.95%	99.96%	98.84%
PA	99.15%	77.28%	58.29%	99.96%	99.96%	98.84%

Table 6 Measurement of F-Measure

ML Classifier	Single Fold			10-Fold		
	N-gram					
	Unigram	Bigram	Trigram	Unigram	Bigram	Trigram
NB	95.21%	83.16%	62.37%	95.90%	82.83%	60.50%
BNB	98.00%	57.78%	44.20%	97.73%	60.42%	45.87%
MNB	98.18%	82.83%	62.42%	98.03%	82.31%	60.53%
ME	93.51%	79.14%	53.17%	94.17%	79.83%	54.43%
Ada-Boost	72.53%	51.20%	40.22%	71.90%	51.48%	40.48%
SVM	99.31%	77.31%	48.79%	98.86%	78.89%	50.31%
LR	98.97%	82.48%	50.32%	98.60%	82.51%	51.79%
SGD	99.11%	76.81%	61.92%	98.59%	75.53%	58.77%
RR	99.09%	77.33%	58.07%	99.95%	99.96%	98.84%
PA	99.15%	76.67%	49.82%	99.96%	99.96%	98.84%

From the experiments above it is observed that PA and RR give the highest performance with different n-gram features. Using Unigram with PA achieved the highest accuracy, precision, recall and f-measure as shown in Fig.2.

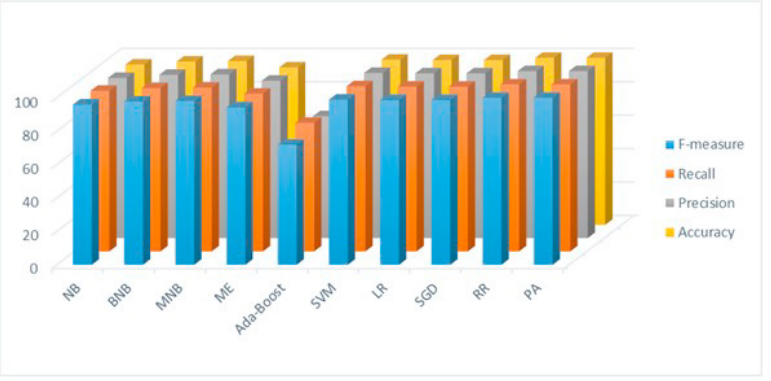

Fig. 2 Unigram 10-fold Measurements

From Fig.2 it can be noticed that also LR, SGD, RR have almost above 98% in all terms of metrics. Applying Ada-Boost achieved the lowest value in all evaluation metrics, as it is less than 78%.

Using Bigram with PA and Bigram with RR achieved the same value in all evaluation metric and it is the highest value compared to other ML algorithms as displayed in Fig. 3.

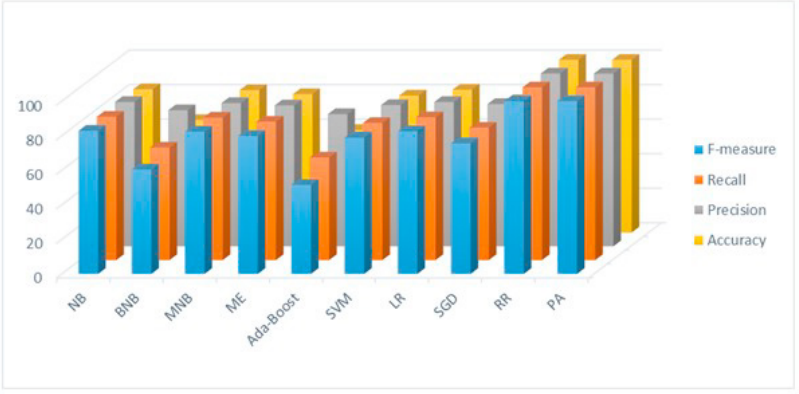

Fig. 3 Bigram 10-fold Measurements

From Fig.3 it can be observed that NB, SVM, and LR have almost equal accuracy which exceeds 82%. Applying Ada-Boost achieved the lowest accuracy with 59.38%.

Using Trigram as a feature extractor achieved the lowest accuracy among different ML algorithms compared with Unigram and Bigram as presented in Fig. 4.

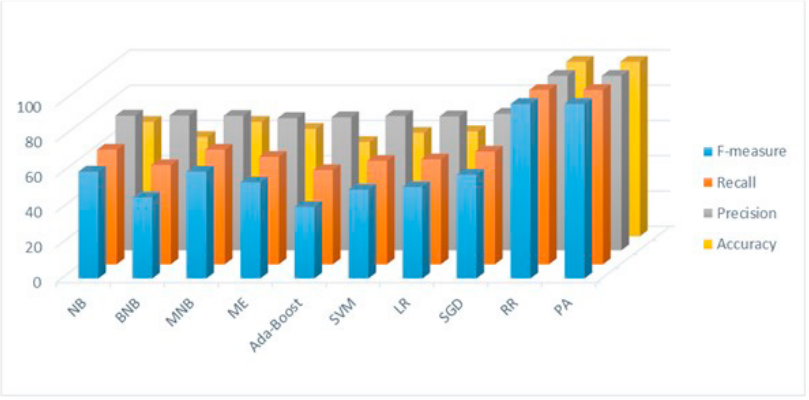

Fig. 4 Trigram 10-fold Measurements

From Fig.4 it can be seen that PA and RR achieve the highest accuracy which exceeds 98%. The lowest accuracies were for applying SVM, Ada-Boost, LR and BNB less than 60%.

It is noticed that using unigram outperformed using bigram or trigram with different ML algorithms on an Arabic tweets dataset. The maximum accuracy of 99.96% is obtained for unigram using PA and RR. On expanding the feature extraction procedure for bigram and trigram, it can be concluded that bigram gives preferable and better accuracy and precision than trigram. Expanding the procedure to higher order n-gram complicates the procedure prompting over-fitting[24].

5. Conclusions and Future Work

This work has concerned with SA in Arabic textual content. A dataset of different Arabic dialects, which consists over 151,500 tweets/comments, was collected and automatically labeled. The NB, LR, ME, PA, RR, SVM, MNB, Ada-Boost BNB, and SGD classifiers were used to extract and discover the polarity of a given tweet. A 10-fold cross validation was utilized to divide the data into a separated training set and testing set. The best evaluation metric values are achieved by PA and RR using unigram, bigram, or trigram is 99.96%.

For future work, certainly there are numerous ways that this work can and will be progressed and improved. Firstly, the dataset can be extended by including a Franco-Arabic dialect. Semi-supervised learning algorithms can be utilized to sentiment analysis in Arabic text as these algorithms have been applied effectively to different languages.

6. References

1. Gamal, Donia, Marco Alfonse, El-Sayed M. El-Horbaty, and Abdel-Badeeh M. Salem. "A comparative study on opinion mining algorithms of social media statuses." Proceedings of Intelligent Computing and Information Systems (ICICIS), 2017 Eighth International Conference on, pp. 385-390. IEEE, 2017.
2. Bollen, Johan, Huina Mao, and Alberto Pepe. "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena." Proceedings of International Aaai Conference On Web And Social Media (Icwsm), No. 11, pp: 450-453, 2011.
3. Petz, Gerald, Michał Karpowicz, Harald Fürschuß, Andreas Auinger, Václav Stříteský, and Andreas Holzinger. "Opinion mining on the web 2.0–characteristics of user generated content and their impacts." In Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data, pp. 35-46. Springer, Berlin, Heidelberg, 2013.
4. Hagen, Matthias, Martin Potthast, Michel Büchner, and Benno Stein. "Twitter sentiment detection via ensemble classification using averaged confidence scores." Proceedings of European Conference on Information Retrieval, pp. 741-754. Springer, Cham, 2015.
5. Chen, Hsinchun, Roger HL Chiang, and Veda C. Storey. "Business intelligence and analytics: from big data to big impact." International Journal of Management Information Systems Research Center (MIS) quarterly, Vol. 36, No. 4, pp: 1165-1188, 2012.
6. Malouf, Robert, and Tony Mullen. "Taking sides: User classification for informal online political discourse." International Journal of Internet Research: Electronic Networking Applications and Policy, Vol.18, No. 2, pp: 177-190, 2008.
7. Glance, Natalie, Matthew Hurst, Kamal Nigam, Matthew Siegler, Robert Stockton, and Takashi Tomokiyo. "Deriving marketing intelligence from online discussion." Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 419-428. ACM, 2005.
8. Ravi, Kumar, and Vadlamani Ravi. "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." International Journal of Knowledge-Based Systems, Vol. 89, pp: 14-46, 2015.
9. Alwakid, Ghadah, Taha Osman, and Thomas Hughes-Roberts. "Challenges in Sentiment Analysis for Arabic Social Networks." International Journal of Procedia Computer Science, Vol. 117, pp: 89-100, 2017.
10. http://www.arabsocialmediareport.com/ [last access in July 2018]
11. Omar, Nazlia, Mohammed Albared, Adel Qasem Al-Shabi, and Tareq Al-Moslmi. "Ensemble of classification algorithms for subjectivity and sentiment analysis of Arabic customers' reviews." International Journal of Advancements in Computing Technology, Vol. 5, no. 14, pp: 77-85, 2013.
12. Hamouda, Safa Ben, and Jalel Akaichi. "Social networks' text mining for sentiment classification: The case of Facebook'statuses updates in the 'Arabic Spring'era." International Journal Application or Innovation in Engineering and Management, Vol. 2, no. 5 pp: 470-478, 2013.
13. Zainuddin, Nurulhuda, and Ali Selamat. "Sentiment analysis using support vector machine." Proceedings of In Computer, Communications, and Control Technology (I4CT), pp. 333-337. IEEE, 2014.
14. Pang, Bo, and Lillian Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." Proceedings of the 42nd annual meeting on Association for Computational Linguistics, p. 271. Association for Computational Linguistics, 2004.
15. M. Taboada, C. Anthony, and K. Voll, "Methods for creating semantic orientation dictionaries," Proceedings of Language Resources and Evaluation (LREC), pp. 427–432, 2006
16. Mohammad, Adel Hamdan, Tariq Alwada'n, and Omar Al-Momani. "Arabic text categorization using support vector machine, Naïve Bayes and neural network." GSTF Journal on Computing (JoC), Vol. 5, no. 1, 2018.
17. Roesslein, Joshua. "tweepy Documentation." [Online] http://tweepy. readthedocs. io/en/v3 5 ,2009.
18. Krouska, Akrivi, Christos Troussas, and Maria Virvou. "The effect of preprocessing techniques on twitter sentiment analysis." Proceedings of In Information, Intelligence, Systems & Applications (IISA), 2016 7th International Conference on, pp. 1-5. IEEE, 2016.
19. Zhai, Zhongwu, Hua Xu, Bada Kang, and Peifa Jia. "Exploiting effective features for chinese sentiment classification." International Journal of Expert Systems with Applications, Vol. 38, no. 8 , pp:9139-9146, 2011.
20. http://scikit-learn.org/ [last access in July 2018]
21. https://www.nltk.org/ [last access in July 2018]
22. Joachims, T. "Text categorization with support vector machines: Learning with many relevant features". Machine Learning: ECML-98, Lecture Notes in Computer Science, 1398, pp 137-142, 2005.
23. Moraes,Rodrigo, João Francisco Valiati, Wilson P. Gavião Neto." Document-level sentiment classification: An empirical comparison between SVM and ANN", International Journal of Expert Systems with Applications, Vol 40, pp. 621 – 633,2013
24. Tripathi, Gautami, and S. Naganna. "Feature selection and classification approach for sentiment analysis." International Journal of Machine Learning and Applications, Vol. 2, No. 2, pp: 1-16, 2015.