Arabic Sentiment Analysis using WEKA a Hybrid Learning Approach

Sarah Alhumoud, Tarfa Albuhairi and Mawaheb Altuwaijri
College of Computer and Information Science, Al-Imam Muhammad Ibn Saud Islamic University, Riyadh, Saudi Arabia

Keywords: Sentiment Analysis, Data Mining, Machine Learning, Supervised Approach, Hybrid Learning Approach.

Abstract: Data has become the currency of this era and it is continuing to massively increase in size and generation rate. Large data generated out of organisations' e-transactions or individuals through social networks could be of a great value when analysed properly. This research presents an implementation of a sentiment analyser for Twitter's tweets which is one of the biggest public and freely available big data sources. It analyses Arabic, Saudi dialect tweets to extract sentiments toward a specific topic. It used a dataset consisting of 3000 tweets collected from Twitter. The collected tweets were analysed using two machine learning approaches, supervised which is trained with the dataset collected and the proposed hybrid learning which is trained on a single words dictionary. Two algorithms are used, *Support Vector Machine* (SVM) and *K-Nearest Neighbors* (KNN). The obtained results by the cross validation on the same dataset clearly confirm the superiority of the hybrid learning approach over the supervised approach.

1 INTRODUCTION

Online data is doubling in size every two years (Gantz and Reinsel, 2011). The amount of online data generated in 2013 was 4.4 Zettabytes (ZB), and in 2020 it will reach 44 ZB (Gantz and Reinsel, 2011). Individual users are the main source, contributing 75% to the overall produced data (EMC, 2011). Big data is described by the 3'Vs model: variety, velocity and volume. Data variety indicates both structured and unstructured data such as email, video, audio, images, click streams, logs, posts or search queries. Velocity refers to the speed needed to process and store the huge and complex data, to respond to the increasing and continuous requests. Volume indicates the massive size of generated data (Sagiroglu and Sinanc, 2013).

Social networks such as Twitter and Facebook, which are popular means for communication, are important sources for big data that could be harvest and analysed. Twitter, a micro blogging social network, founded in 2006, enables users to freely, easily, and instantaneously express, reach, and share opinions and feelings in public in an SMS style text, called tweets. Each tweet has 140 characters or less (Twitter, 2015). In 2014, a study showed that there are more than 5.8 Arab users (Arab Social Media Report, 2014) out of 255 million users from all over the world (Twitter, 2014). Based on a study done by Twitter in 2015 it has been shown that there are 500 million tweets per day (Twitter, 2015) while 10.8 million tweets of them are written in Arabic (Arab Social Media Report, 2014) as shown in Figure 1. Saudi users produced 40% of all tweets in the Arab world (Arab Social Media Report, 2014).

The ability to extract meaning out of available data is a valuable asset in leading organizations and companies. Knowing clients behaviour, feedback and opinion in order to improve services and products. Although organisations could conduct interviews directly with clients or distribute questionnaires to collect clients' feedback. That drains a considrable amount of time, effort and cost. In addition, it may not serve as a precise indication to actual costumers' behaviour and preferences as the questionnaire may not cover all needs or it may not be answered thoughtfully and accurately. Moreover, clients may not express their immediate feedback openly and timely in a questionnaire compared to what they do in open, personal and global social network like Twitter. Data mining is the process of extracting data or knowledge from a large amount of data.

Moreover, the data mining is analysing and searching data or knowledge (Witten et al., 2011).

Classification and clustering are two important methods to data mining. Classification aims to find a

Alhumoud, S., Albuhairi, T. and Altuwaijri, M..
Arabic Sentiment Analysis using WEKA a Hybrid Learning Approach.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR*, pages 402-408
ISBN: 978-989-758-158-8

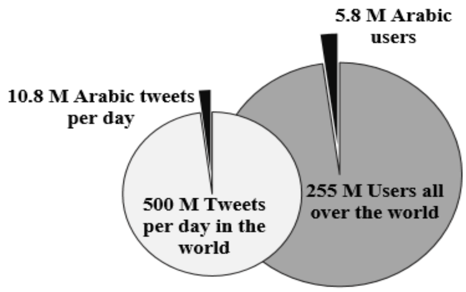

Figure 1: Twitter Arab usage.

classes by feeding it with classified data, training it to learn what make each data of the given class (Han et al., 2000). Clustering takes the input and puts it into clusters while the objects of the same cluster have similar factors (Han et al., 2000). One of the data mining fields is text mining (Han et al., 2000). Text mining uses the same process of data mining but is specialised in data presented as text. Text mining can help in topic categorization and sentiment analysis.

Sentiment Analysis (SA) is a one of the *Natural Language Processing* (NLP) concepts, which is also called opinion mining (Liu, 2012). This field of computer science is used to extract sentiment out of text giving useful information about the author's tendency towards a specific topic. Two approaches can be used in SA, supervised and unsupervised, those will be described more in detail in the related work section.

SA can be implemented using tools for data mining like RapidMiner (RapidMiner, 2015), R (R-project, 2015), WEKA (WEKA, 2014) and Orange (Orange, 2015).

WEKA is considered as one of the widely used data mining tools (WEKA, 2014), (Choo et al., 2013). WEKA stands for Waikato Environment for Knowledge Analysis (WEKA, 2014), (Choo et al., 2013). WEKA is a none-paid software that provide Graphical User Interface (GUI) and a coding interface with Java code. WEKA could be used to implement multiple data mining functions such as classification, association, clustering, and regression. Some key advantages of WEKA is that it can perform data pre-processing, visualization and feature selection (Choo et al., 2013). In addition, WEKA provides knowledge flow interfaces that make it easier for the user to specify the flow of data though connecting visual components (Jović et al., 2014).

This rest of the paper is divided as following: related work section where a list of similar work presented and discussed, then the methodology section that describes the implementation processes to the sentiment analyser. After, the results and discussions of each approach. Finally, the conclusion section where a conclusion of the presented work is presented there with the future directions.

2 RELATED WORK

There are two learning approaches for sentiment analysis supervised and unsupervised (Medhat et al., 2014) and (Ravi, 2015). The supervised learning (also known as corpus-based approach) relay on machine learning (Vinodhini and Chandrasekaran, 2012) and uses machine learning algorithms such as *Support Vector Machine* (SVM), *Naïve Bayes* (NB), *Decision Tree* (D-Tree) and *K-Nearest Neighbors* (KNN) to build a classifier. Moreover, the supervised learning contains five main stages: building the dataset, building the classifier (model), training the classifier, evaluating the classifier, and using the classifier.

The first stage in supervised learning is build the training data and testing data through giving it labels. The second stage is building the classifier (model) through using one of the data mining algorithms. After that training the classifier using a training dataset that have been built previously. After the classifier has been trained its performance needs to be evaluated using a testing dataset, in this stage the classifier guesses the labels after they have been hidden. Finally, the classifier is used to classify a new dataset without labels.

Unsupervised approach, also known as lexicon-based approach (Liu, 2012) and (Alhumoud et al., 2015). The approach is based on lexicons or dictionaries which can be created manually or automatically (Medhat et al., 2014). Lexicon is a collection of opinion words where each word is associated with a polarity value: +1, -1 or 0 for positive, negative or neutral, respectively (Shoukry and Rafea, 2012) and (Ravi, 2015).

The next subsection will present related studies in data mining, data mining classification and sentiment analysis studies using WEKA.

2.1 General Data Mining

Many researchers have used WEKA to proof their experiments in the fields of data mining. In a paper by (Apala et al., 2013) they have implemented clustering function to perform text mining in twitter using K-means algorithm. Another clustering solution was applied in the work proposed by (Ali and Massmoudi, 2013) they enhanced their results using Gower similarity coefficient. Also, authors (Ahmed and Bansal, 2013) hired clustering to enhance search engines using K-means algorithm. Authors (Parack et al., 2012) have used along with clustering another function that is association. They used clustering to group students while using association for students profiling. Association is a data mining function implemented using WEKA. The work that was proposed by (Chen et al., 2013) have used association on herbs to find the association between them. While (Lekhal et al., 2013) used the association function in case studies related to breast cancer, larynx cancer and other datasets.

2.2 Data Mining Classification

WEKA was used to perform classification to solve problems in different domains. The work that was done by (Dan et al., 2012) has performed text categorization using supervised approach. They used three algorithms to perform text categorization those are, SVM, NB, and D-Tree. They did feature selection to reduce the feature space and improve the classifiers performance using the *Information Gain* (IG) method. The SVM has got the highest accuracy with 95%. Their results shows a high accuracy which were affected by the used method of feature selection.

Authors (Dass et al., 2014) have applied classification in the field of diseases. They have tried to predict the class of lung cancer using supervised approach. J48 algorithm has achieved 99.7% accuracy through building eight rules for classification. Another work by (Saraç and Özel, 2013) was done using WEKA to classify web pages. They added *Firefly Algorithm* (FA) to enhance the process of feature selection and used J48 algorithm to build the classifier. Their addition made a difference through reducing time for classification with no loss in accuracy. Researchers (Shah and Jivani, 2013) have studied 699 instances from the Hospital of Wisconsin University. They tested three classification methods those are NB, *Random Forest* (RF) and KNN. Their results tells that the accuracy of NB is close to the other two 95.9% and takes less execution time. Finally the paper by (Thabtah et al., 2011) they used dataset of 415 documents to perform text categorization on Arabic text through using four algorithms: C4.5, RIPPER, PART and OneRule. The results shows that OneRule algorithm comes in the last rank with less accuracy while C4.5, RIPPER, PART have got similar accuracy results.

2.3 SA in WEKA

WEKA has been used for SA purposes by lots of papers and researches. Proposed work by (Jin et al., 2014) has used WEKA to build their classifier. They first collected 2250 tweets with the keyword Obama using Twitter *Application Program Interface* (API), in specific stream API. They used NB, SVM, D-Tree and RF. The experiments show that RF performance was the best among the others. They used a 10-fold cross-validation to evaluate the classifiers. The results have raised when combination of temporal, punctuations, emoticons and PMI-IR values were used as features. While the negation feature has decreased the accuracy of the classifiers. Another work made by (Shoukry and Rafea, 2012) were they used also Twitter API to collect a dataset of 4000 Arabic tweets and they used 1000 tweets as a training set. They pre-processed the dataset before feeding it to the classifier by removing user-names, pictures, hashtags, URLs, and all non-Arabic words. They used WEKA to set the n-gram size and build the classifiers. They used two types of n-gram: unigram and unigram with bigram. For the classifiers they used SVM and NB. Their experiment was of two parts: first part is to find the effect of unigram and unigram with bigram on the accuracy, and the second part is to show the effect of removing stop words along with using unigram and unigram with bigram. The first experiment shows that there is no effect on the accuracy when unigram used alone or when combining unigram with bigram. The second experiment revealed that the effect of removing stop words is not noticeable. Author (V, 2014) has collected about 5574 SMSs. He applied filtering process on the SMS's and used two algorithms SVM and D-Tree. The goal of the research is to find the best tokenizer among the three that they used. The results state that *AlphabeticTokenizer* has the best effect on the classifiers where SVM got 92% and D-Tree got 89%.

3 METHODOLOGY

The hybrid learning approach was applied using WEKA package with Java code, the results were compared against those of the WEKA GUI.

Hybrid learning approach contains five main stages: building the dataset, building the classifier (model), training the classifier, evaluating the classifier, and using the classifier to get overall sentiment of a new dataset as shown in Figure 2.

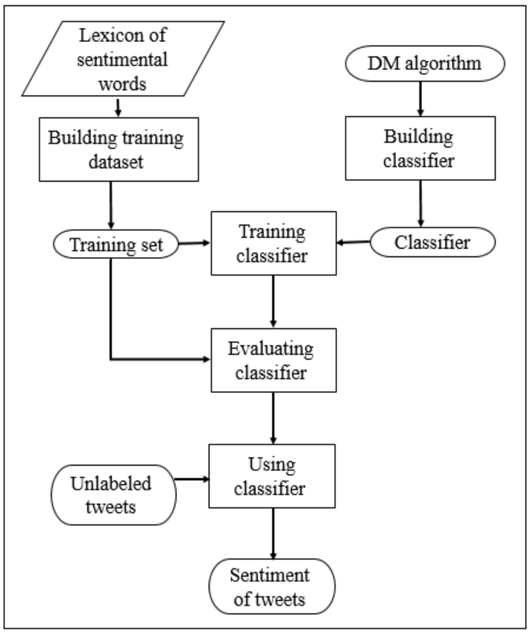

Figure 2: Hybrid learning stages.

3.1 Building the Training Dataset

The training dataset was built from rows of single sentimental words, instances and their label. A labelled word tagged positive or negative according to the word sentiment. The training dataset contains 3690 sentimental words, 1370 words were positive and the remaining 2320 words are negative. The training dataset contains 1000 MSA sentimental words that could be found in (NLP4Arabic, 2012) and 2690 Saudi dialect sentimental words were added manually with the help of 1000 tweets from the sports domain collected form Twitter. Datasets need to be represented in either *Comma Separated Value* (CSV) or *Attribute-Relation File Format* (ARFF). The latter, ARFF, is WEKA's default format and is an ASCII text file describing a list of instances sharing a list of attributes (WEKA, 2015). While the former is an accepted file format in WEKA this research uses the latter. Additionally, ARFF, consumes less memory than the former, it is faster, and better for analysis as it includes metadata about columns headers.

In the supervised approach, each line contains one instance, a tweet. This instance holds several words that do not affect the sentiment but causes confusion in the classifier, hence, decreasing the accuracy. If these words are removed from the instance, only sentimental words will remaining which is similar to lexicon in unsupervised approach. Using sentimental lexicon as a training dataset aids in minimising the classifier confusion. Therefore, the hybrid learning incorporates the advantages of a data mining algorithm in supervised approach and lexicon based approach in unsupervised approach to better teach the classifier. The main difference between supervised and hybrid learning is in building the training dataset. This difference has a noticeable effect in the results.

The hybrid learning also avoids the pre-processing cost associated with the supervised approach. Building a training dataset in the supervised learning requires several steps, first, collecting n instances, second pre-processing them with O(n) time complexity, third normalizing them with O(n) time complexity as well. While the hybrid learning approach does not require the second nor third steps, minimising overall overhead.

3.2 Building the Classifier

Second step to the classification is building the classifier. In hybrid learning approaches, two different experiments were conducted, each experiment used one of the data mining algorithms to build a classifier, model. SVM and KNN data mining algorithms were used in this research. These algorithms work efficiently in text classification and they showed superior performance in previous related studies (Khasawneh et al., 2013), (Shoukry and Rafea, 2012) and (Abdulla et al., 2013).

3.3 Training the Classifier

Training the classifier is the third step in hybrid learning approach. Data mining algorithms in WEKA are not compatible with string data type. For this reason, a filter *StringToWordVector* is used to convert string attributes to numeric. This research used *NGramTokenizer* tokenizer which splits a string into n-gram with min and max gram. Only unigram could be applied because the training set contains one word in each row.

3.4 Evaluating the Classifier

Evaluating the classifier is the fourth step, in which the training dataset is used as testing dataset to produce the expected label of each tweet. The 10-fold cross validation technique was used to evaluate the classifier; since cross validation is more suitable for small datasets. Cross validation evaluates the classifier multiple times by specifying the fold value using the training set as a testing set. The result of the evaluation is measured by computing the precision, and recall. The precision and recall equations are presented below:

$$\text{Precision} = TP / (TP + FP) \quad (1)$$

$$\text{Recall} = TP / (TP + FN) \quad (2)$$

Where TP, FP, TN, and FN are true positive, false positive, true negative and false negative, respectively. In precision and recall, the highest level of performance is equal to one and the lowest is zero (Sokolova et al., 2006).

3.5 Using the Classifier

Fifth step comprises the classification of a new dataset by using the produced classifier to expect overall sentiment polarity of each tweet in the dataset. The dataset must be normalized, and filtered using the *StringToWordVector* filter and *NGramTokenizer* tokenizer which is used in training the classifier to match words correctly. Classification accuracy was measured by computing the number of correct classified tweets.

4 RESULTS AND DISCUSSIONS

This research applied two experiments, hybrid learning and supervised using the same dataset. Both hybrid learning and supervised learning was implemented using two classifiers: SVM and KNN. The supervised classifier was trained on sports domain using 1000 tweets. Table 1 shows a comparison between the accuracy of hybrid learning and supervised classifiers when classifying a new dataset contains 1000 tweets in sports domain. Where “Sup” and “Hyb” denote supervised learning and hybrid learning, consecutively. SVM in hybrid learning was outperforming SVM in supervised by 6% in average, also KNN in hybrid learning was outperforming that of the supervised by 15%. An explanation to this superiority is related to the *minimum term frequency* (MTF). Selecting a suitable MTF is a key factor in the overall accuracy of the classifier. In other words, a word that is repeated the same number of the MTF; WEKA algorithm considers it a sentimental word. *SetMinTermFreq()* is a WEKA function used to specify MTF and it is set to one by default. The issue is that most tweets have general words that may be repeated with a frequency more than one. Consequently, the classifier considers those words as sentimental words based on the threshold set by one by default in the *SetMinTermFreq()*. This will cause mistakes in classifying some tweets because the repeated words will appear in both positive and negative tweets decreasing the accuracy of the classifiers. Hybrid learning, proposed in this research implies a new technique in building the training dataset. That is training the classifier on a set of one labelled sentimental word, as said in section 3.1. This technique increases the accuracy of the classifiers.

When using the same SVM classifiers to classify new datasets in new domains, the hybrid learning approach achieved higher accuracy than supervised as shown in Figure 3. Three domains were used, 1000 sports tweets, 500 social and 500 political to measure the accuracy of the classifiers. These tweets are never seen by the classifiers before. Additionally, both social and political domains are new domains to the classifier. Supervised classifier scored a low accuracy when classifying datasets in social or political domains. While hybrid learning classifier was better by more than 20%. This is explained by the training dataset used in the hybrid and supervised approaches. While the former used a one word instances the latter used instances with more than one word increasing classifier confusion. This proves the hybrid learning approach scalability to analyse new domains over the supervised approach. Moreover, increasing the training dataset size in hybrid learning is easier than the supervised because it does not require the pre-processing steps prior to building the training dataset which are required in the supervised approach.

The hybrid learning approach could be implanted on languages other than Arabic, such as English. For the English language, it is even more convenient and easy to use the full features of R SA packages.

Incorrect classification is caused by several factors. One is that some tweets have negation words and this inverts the polarity eliding the opposite polarity. Another is tweets that have two sentiment, and tweets that have ambiguous sentiment. Table 2 shows examples on tweets which have incorrect classification.

The hybrid learning approach’s accuracy can be improved by increasing the size of the training dataset, and by using a words’ stemmer. In additional, negation should be considered to inverse the meaning of word.

Table 1: Accuracy of hybrid learning and supervised classifiers in sports domain.

Dataset Size	SVM		KNN	
	Hyb	Sup	Hyb	Sup
100	92.00%	89.50%	92.00%	84.80%
250	89.40%	80.40%	90.60%	69.40%
500	89.90%	80.50%	90.70%	69.60%
1000	88.50%	83.90%	88.50%	75.40%
Average	90.00%	83.60%	90.50%	74.80%

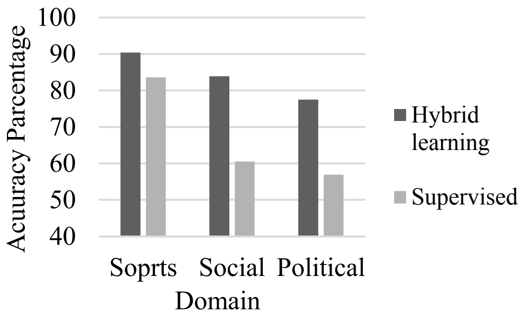

Figure 3: Hybrid learning and supervised classification in several domains.

Table 2: Example on difficulty classification tweets.

Polarity	Problem	Example Tweets in Arabic and in English
Negative	Negation	ما حبيت لعب الفريق اليوم I did not like the game today
Negative	Two sentiments	يحسبونه يوم خسر مهزوم ما دروا انه بطل لو خسر Even when defeated still a champion
Negative	Ambiguous sentiment	المفروض مع كل مباراة يعطونك حبوب ضغط وسكر with each match a prescription is needed

5 CONCLUSIONS

Nowadays it is not enough to own data, but being able understand it efficiently and analyse it in a timely manner gives its owner knowledge and power. This paper presented the new hybrid learning approach for Arabic SA in Twitter examining the classification of randomly collect tweets in three domains. The results confirm that the hybrid learning approach has better accuracy than both supervised and unsupervised approaches. The hybrid learning approach scored an enhancement in accuracy of 6%, 23% and 21% in sports, social and political domains respectively over the supervised approach using SVM. Additionally, KNN classifier in hybrid learning approach outperformed the supervised approach with 15% in accuracy. Proving that the hybrid approach has higher accuracy and scalability over the supervised approach.

REFERENCES

Abdulla, N. Ahmed, N. Shehab, M. & Al-Ayyoub, M. (2013) *Arabic Sentiment Analysis: Lexicon-Based and Corpus-Based*. Proceedings of the IEEE Jordan Conference Applied Electrical Engineering and Computing Technologies (AEECT). Amman, pp. 1–6.

Ahmed, E. & Bansal, P. (2013*) Clustering Technique on Search Engine Dataset using Data Mining Tool*. Proceedings of International Conference on Modeling, Simulation and Applied Optimization, Hammamet, pp. 1 - 5.

Alhumoud, S. Altuwaijri, M. Albuhairi, T. & Alohaideb, W. (2015) *Survey on Arabic Sentiment Analysis in Twitter*. Proceedings of the International Conference on Computer Science and Information Technology (ICCSIT), Paris, pp. 364 – 368.

Ali, B. & Massmoudi, Y. (2013) *K-Means clustering based on Gower Similarity Coefficient: A comparative study*. Proceedings of International Conference on Advanced Computing & Communication Technologies, Rohtak, pp. 86 – 89.

Apala, K. Jose, M. Motnam, S. Chan, C. Liszka, K. & Gregorio, F. (2013*)Prediction of Movies Box Office Performance Using Social Media*. Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagara Falls, ON, pp. 1209 - 1214.

Arab Social Media Report. (2014) *Twitter in the Arab Region*, Dubai School of Government, [Online] Available from: https://shar.es/129INW. [Accessed: 16th June 2015].

Chen, S. Xie, X. Zeng, Z. Yu, J. & Lu, C. (2013*) Study of the Regularities in the Treatment of Psoriasis Vulgaris by TCM: Applying Association Rule Mining to TCM Literature*. Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, Shanghai, pp. 15 - 17.

Choo, T. Abu Bakar, A. Talebi, A. Sundararajan, E. & Rahmany, M. (2013) *Classification modeling on distributed environment*. Proceedings of IEEE Conference on Open Systems, Kuching, pp. 209 - 214.

Dan, L. Lihua L. & Zhaoxin,Z. (2012) *Research of Text Categorization on WEKA*. Proceedings of International Conference on Intelligent System Design and Engineering Applications, Hong Kong, pp. 1129 - 1131.

Dass, V. Abdul Rasheed, M. & Ali, M. (2014) *Classification of Lung cancer subtypes by Data Mining technique*. Proceedings of International Conference on Control, Instrumentation, Energy & Communication, Calcutta, pp. 558 - 562.

EMC. (2011) *The 2011 IDC Digital Universe Study Sponsored by EMC*, [Online] Available from: http://www.emc.com/collateral/about/news/idc-emc-digital-universe-2011-infographic.pdf. [Accessed: 16th June 2015].

Gantz, J & Reinsel, D. (2011) *Extracting Value from Chao*, EMC, [Online] Available from: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf. [Accessed: 16th June 2015].

Han, J. Kamber, M. & Pei, J. (2000) *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Jin, H. Zhu,Y. Jin,Z. and Arora,S (2014) *Sentiment*

Visualization on Tweets Stream Journal of Software. 9 (9). p. 2348-2352.

Jović, A. Brkić, K. & Bogunović, N. (2014) *An Overview Of Free Software Tools For General Data Mining*. Proceedings of the International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, pp. 1112 – 1117.

Khasawneh, R. Wahsheh, H. Al Kabi M. & Aismadi, I. (2013) *Sentiment analysis of arabic social media content: a comparative study*. Proceedings of the 8th International Conference for Internet Technology and Secured Transactions (ICITST). London, pp. 101 - 106.

Lekhal, A. Srikrishna,C. &Vinod,V. (2013*) Utility of Association Rule Mining: a Case Study using WEKA Tool*. Proceedings of International Conference on Emerging Trends in VLSI, Embedded System, Nano Electronics and Telecommunication System, Tiruvannamalai, pp. 1 - 6.

Liu, B. (2012) *Sentiment Analysis and Opinion Mining*. Morgan & Claypool.

Medhat, W. Hassan, A. Korashy H. (2014) *Sentiment analysis algorithms and applications: A survey*, Ain Shams Engineering Journal (5). P. 1093–1113. [Online] Available from: http://www.sciencedirect.com/science/article/pii/S2090447914000550. [Accessed: 28th Aug 2015].

NLP4Arabi. (2012) *Arabic MPQA Subjective Lexicon & Arabic Opinion Holder Corpus*. [Online] Available from: http://nlp4arabic.blogspot.com/2012/05/arabic-mpqa-subjective-lexicon-arabic.html . [Accessed: 17th June 2015].

Orange. (2015) *Data Mining - Fruitful and Fun*, [Online] Available from: http://orange.biolab.si/. [Accessed: 13th July 2015].

Parack, S. Zahid, Z. & Merchant,F. (2012) *Application of Data Mining in Educational Databases for Predicting Academic Trends and Patterns*. Proceedings of IEEE International Conference on Technology Enhanced Education, Kerala, pp. 1 - 4.

R- project. (2015) *The R Project for Statistical Computing*, [Online] Available from: http://www.r-project.org/. [Accessed: 13th July 2015].

RapidMiner. (2015) *Predictive Analytics Reimagined*, [Online] Available from: https://rapidminer.com/. [Accessed: 13th July 2015].

Ravi, K. & Ravi, V. (2015) *A survey on opinion mining and sentiment analysis: Tasks, approaches and applications*, Knowledge-Based Systems (1). P. 1–33. [Online] Available from: http://www.science direct.com/science/article/pii/S0950705115002336. [Accessed: 28th Aug 2015].

Sagiroglu, S. & Sinanc, D. (2013) *Big Data: A review*, Proceedings of the International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, pp. 42 - 47.

Saraç, E. & Özel,S. (2013) *Web Page Classification Using Firefly Optimization*. Proceedings of IEEE International Symposium on Innovations in Intelligent Systems and Applications, Albena, pp. 1 - 5.

Shah, C. & Jivani, A. (2013) *Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction*. Proceedings of International Conference on Computing, Communications and Networking Technologies, Tiruchengode, pp. 1 - 4.

Shoukry, A. & Rafea, A. (2012) *Sentence Level Arabic Sentiment Analysis*. Proceedings of the International Conference on Collaboration Technologies and Systems, Denver, USA, pp. 546 - 550.

Sokolova, M. Japkowicz, N. and Szpakowicz, S. (2006) *Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation*. In Hutchison, D. Kanade, T. Kittler, J. Kleinberg, J.M. Mattern, F. Mitchell, J.C. Naor, M. Pandu Rangan, C. Steffen, B. Terzopoulos, D. Tygar, D. & Weikum, G. (eds.). *AI 2006: Advances in Artificial Intelligence*. Lecture Notes in Computer Science (4304). Australia Springer Berlin Heidelberg.

Thabtah, F. Gharaibeh, O. & Abdeljaber,H. (2011) *Comparison of Rule based Classification Techniques for the Arabic Textual Data*. Proceedings of International Symposium on Innovation in Information & Communication Technology, Amman, pp. 105 - 111.

Twitter. (2014) *About Twitter*, [Online]. Available from: https://about.twitter.com/what-is-twitter. [Accessed: 14th July 2015].

Twitter. (2015) [Online] Available from: https://support.twitter.com/articles/215585-getting-started-with-twitter. [Accessed: 16th June 2015].

Twitter. (2015) *About.Twitter*, [Online]. Available from: https://about.twitter.com/en/company. [Accessed: 31th Aug 2015].

V, U. (2014) *Sentiment Analysis Using WEKA* International Journal of Engineering Trends and Technology. 18 (4). p. 181-183.

Vinodhini G. & Chandrasekaran, RM. (2012) *Sentiment Analysis and Opinion Mining:A Survey*. International Journal of Advanced Research in Computer Science and Software Engineering. [Online] (2). P. 283- 292. Available from: http://www.dmi.unict.it/~faro/tesi/sentiment_analysis/SA2.pdf. [Accessed: 16th June 2015].

WEKA. (2014) *WEKA 3: Data Mining Software in Java*, [Online] Available from: http://www.cs.waikato.ac.nz/ml/weka/index.html. [Accessed: 24th June 2015].

WEKA. (2015) *ARFF (book version)*, [Online] Available from: http://weka.wikispaces.com/ARFF+%28book+version%29. [Accessed: 20th April 2015].

Witten, I. Frank, E. & Hall, M. (2011) *Data Mining Practical Machine Learning Tools and Techniques*. Burlington: Elsevier.