iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** I **Month of publication:** January 2020

DOI: http://doi.org/10.22214/ijraset.2020.1056

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 8 Issue I, Jan 2020- Available at www.ijraset.com

Big Data Analytics using Supervised Learning: A Comprehensive Review of Recent Techniques

Wedjdane Nahili[1], Khaled Rezeg[2], Okba Kazar[3]
[1, 2, 3]*Computer Science, LINFI laboratory, Biskra University, Algeria*

Abstract: ***With the availability of text data in various forms on social media platforms, text mining, and sentiment analysis has received huge attention. The task of deriving information from this volume of data to extract knowledge is very complex and expensive because it is usually unstructured and contains noise. Recently, there is a growing need for implementing various approaches and models for efficiently processing this type of data and extracting useful information. This process is known as sentiment analysis, which includes: data gathering, data pre-processing, feature engineering and labeling, finally the application of various natural language processing and machine learning algorithms. This paper provides an overview of the most recent methods used in text mining and sentiment analysis along with their detailed description and a discussion of obtained results.***
Keywords: Sentiment analysis; business intelligence; big data analytics; natural language processing; text analytics

I. INTRODUCTION

Sentiment analysis SA is an active research field in Natural Language Processing (NLP), where people's emotions, opinions, and sentiments towards different entities like products, services, and organizations are studied and analyzed. Sentiment analysis is important for companies, organizations and individual persons [1]. Sentiment analysis can be applied to various sectors such as E-commerce, banking, mining social media websites like Facebook, Twitter and so on. By taking advantage of sentiment analysis and text mining, organizations can gain consumer insight from the responses (reviews) about their products and services. This can be further used to study customers' satisfaction with the services and in case of complaints and issues, finding the possible reasons for that. The main domain where sentiment analysis is applied is recommendation systems, for instance, consumers' likes, dislikes, and comments are collected and analyzed and later used to provide Youtube recommendations. In this paper, we provide an extensive study of several most recent sentiment analysis techniques that had been applied to various domains in a multilingual format and from different resources. Typically, a sentiment analysis framework architecture is divided into the following tasks: data acquisition, data pre-processing, data normalization, data conversion also known as vectorization, features selection, and finally applying NLP and machine learning algorithms. The main purpose of this paper is to provide a literature review on recent trends in text mining and sentiment analysis along with a comprehensive description of each approach. For instance, consumer review mining and application to Business Intelligence (BI) and Big Data analytics are the current successful applications. The contribution of this review is significant for many reasons. Initially, this review provides an elaborate extensive study of a large number of recent articles according to the techniques used. This perspective aims to help researchers with good knowledge of the sentiment analysis field to use these methods and choose a suitable technique for their applications. Second, the different techniques of SA that we studied are categorized with brief details (see Table.1) of the used algorithms and their emerging sources. This can provide newcomers a broad landscape on the entire SA field. Finally, the publicly available benchmarks data sets are described and categorized depending on their domain application. This paper is organized as follows: Section 2 includes the review of recent research in the SA field and a summary of the articles. Section 3 tackles the research gaps in the studied articles and Section 4 discusses the experimental results of the surveyed articles, and finally, the conclusion and future work are presented in Section 5.

II. LITERATURE REVIEW

A. Document level

In [8] sentiment analysis was used to perform natural language processing aiming to detect the polarity of a text document. At first, they tackled a binary classification problem where only positive and negative sentiments were discriminated against. Various machine learning techniques were fitted with this problem. Their results were easily reproducible for they used IMDB dataset. A simple and powerful method was proposed for sentiment analysis. They had combined three conceptually different baseline models: first one based on language models, the second one based on consecutive models of sentences and the last one based on the BOW (Bag of Words) model. This paper helped in determining how to use this in standard generative language models. They included a code that is available publicly at http://github.com/mesnilgr/iclr15.

[4] stated that in sentiment analysis the polarity of a given text is determined either using a machine learning approach or lexicon-based approach. In their work, they used machine learning approaches such as Naive Bayes and SVM. To perform their sentiment analysis task the classifiers they executed on the datasets were Naïve Bayes, Support Vector Machine (SVM) and K-Nearest Neighbour (KNN (k=10)). Their results showed that SVM gave the highest score in terms of precision and KNN gave the highest recall. They used 10-fold cross-validation to test the datasets. They demonstrated that the precision achieved by SVM 75.25% was the best precision and the recall got by KNN 69.04% was the best recall. Therefore bigger datasets were required to provide the best classification results.

[5] implemented a new feature type to check its contribution in document-level sentiment analysis. They attained best results on dataset produced by [6] containing 2000 reviews with 91.6% accuracy than by [7] containing 50000 reviews with 89.87% accuracy. They also applied sentiment analysis on a dataset containing 233600 reviews and their proposal achieved 93.24% accuracy. In this paper, an experimental study on sentiment polarity classification had been conducted. First of all, a rating based feature had been described which was based on the regression model, and learned from an external independent dataset of 233600 movie reviews. Afterward, the contribution of both machine learning and rating based criteria were used to achieve accuracy of 91.6% and 89.87% on the datasets from different domains. These results showed that rating based feature was more efficient for sentiment classification on polarity reviews and by adding bi-gram and tri-gram features the performance could be improved.

[9] presented that the reviews and blog datasets obtained from the social networking sites were unsystematic and need a classification for meaningful information. These reviews could be classified as positive, negative and neutral using supervised machine learning methods. In this particular work, they introduced four different machine learning algorithms: NB (Naive Bayes), ME (maximum entropy), SGD (stochastic gradient descent) and SVM (support vector machine) for sentiment classification. They used precision, recall, F-measure, and accuracy as evaluation metrics for the proposed models. This paper helped in classifying the movie reviews using supervised machine learning algorithms which was further applied to the IMDB dataset using the n-gram approach. The results of this paper gave several insights; they concluded the classification accuracy decreases as the value of n increases in the n-gram way; better accuracy was obtained when TF-IDF and count vectorizer techniques were combined. Following a deeper study, they faced some limitations due to the small size of tweets, reviews or comments and the fact that this type of data includes punctuation symbols and words like “greatttt, fineee” as they don't have proper meaning. To solve the problem a lexicon was created for sentiment classification after removing the stop words to select the best feature. And last, hybrid machine learning techniques were also considered for better accuracy (see Table.2 below)

[10] analyzed online audits and film ratings using a content-based sentiment analysis approach. Supervised machine learning strategies were used to group these reviews. For conclusions three different machine learning algorithms were considered; SVM, ME, NB and these were based on evaluation parameters such as accuracy, review, f-measure, and precision. In this paper for classifying film reviews from the rottentomatoes dataset, the authors used the n-gram method where different machine learning techniques had been suggested. This research main conclusion was that in comparison with other research works their results obtained the best accuracy.

B. Sentence Level

[3] proved that for text classification different alternatives of machine learning algorithms show a large variation in their performance. They also demonstrated how Naive Bayes (NB) was more suitable than Support Vector Machines (SVM) for a small part in sentiment tasks in addition to bi-gram results that showed constant improvements in tasks. They also proposed a new SVM variant that showed better consistent results on datasets and resulting in this information they demonstrated NB and SVM variants. This paper resulted in several conclusions: Multinomial Naive Bayes (MNB) was a better choice for sentiment analysis tasks; SVM was more preferred on long reviews; the performance of bi-grams depends on the sentiment tasks; They also concluded that NB-SVM generated the best results and Bernoulli Naive Bayes (BNB) produced poor results compared to MNB.

In [11] Alomari claimed that Arabic tweets pose a good opportunity for opinion mining research but they were set back due to lack of sentiment analysis resources or challenges in Arabic language text analysis. Their work included an Arabic Jordanian twitter corpus in which the tweets were labeled as positive or as negative. These tweets were analyzed using different supervised machine learning approaches. Several techniques experiments were conducted using different weight schemes, stemming and n-grams. This initiative showed that the SVM classifier using TF-IDF through bi-grams feature was better as compared to the Naive Bayesian classifier. The main objective was to examine the machine learning approach for Arabic sentiment analysis. Firstly, the authors collected a new publicly available Arabic tweet corpus containing 1,800 tweets written in the Jordanian dialect. Afterward, various n-grams with different weighting schemes and stemming techniques were used to compare SVM and NB classifiers. Following an

experimental study they finally concluded that the SVM classifier using a combination of TF-IDF weighting scheme with stemmer through bi-grams showed 88.72% accuracy and 88.27% f-score, their model performed better than other Arabic sentiment analyses research results. In the work of [12], sentiment analysis was applied to analyze and extract the polarity of sentiment from product reviews (laptop and restaurant) collected in the SemEval 2014 Task 4 dataset. They conducted an aspect-based sentiment analysis approach which consisted of studying specific aspects of products such as food, service, price, and ambiance (see Table.1). This research was conducted following three phases; data pre-processing which involved part-of-speech (POS) tagging, feature selection using Chi-Square for, it has been proven to speed up the computation time in the classification process, and classification of sentiment polarity of aspects using Naïve Bayes classifier. Based on their evaluation results, the proposed system gave promising output and was able to perform aspect-based sentiment analysis with its highest F1-Measure of 78.12%.

In the work of [17], the goal was to find out why RNN and LSTM models work well for sentiment analysis and how these models work. Their research was based on the principle of compositionality, which states that the meaning of a longer expression depends on the meaning of its predecessors. RNNs were used to perform sentiment analysis because they allow the network to have a memory. Since the author dealt with sequenced text data, having a memory in a network is useful because the meaning of a word depends on the context of the previous text. The main drawback of the RNN is that its capacity of only dealing with short-term dependencies. To deal with this problem they used a combination of both RNNs and LSTM. To illustrate how an LSTM can be used for sentiment classification, a network consisting of four layers: the input layer, the embedding layer, the LSTM layer, and the output layer was built to classify the sentiment of the IMDB dataset. The dataset contained reviews of 50,000 movies. Each movie is labeled with a 0 or a 1, indicating a negative or positive review. Only the top 10,000 most frequently occurring words were used. The model has been trained using word embeddings as a feature. It was very sensitive to overfitting, so the model was stopped from training after the fifth epoch. The final model resulted in a loss and accuracy on the validation set of 0.4366 and 0.8674, respectively. This was already a pretty good result. The parameters in the model have not yet been trained, performing a grid search on the parameters could possibly lead to a better performance of the model.

[18] proposed an ensemble classifier that combines the base learning classifiers such as Naive Bayes, Random Forest, Support Vector Machines (SVM) and Logistic Regression to form a single classifier. Their proposal aimed at improving the performance and accuracy of sentiment classification techniques. The proposed architecture included four modules - (1) Data pre-processing module: for pre-processing the data (2) Feature representation module: for feature extraction from pre-processed tweets, BoW technique was used for converting tweets into numeric representation (3) Sentiment classification using base classifiers: in which different base classifiers were implemented for sentiment analysis and finally (4) Sentiment classification using the proposed ensemble classifier. The implementations were done in Python. The results showed that the proposed ensemble classifier performed better than stand-alone classifiers and majority voting ensemble classifier. Also, as part of their study, the role of data pre-processing and feature representation in sentiment classification technique was explored.

Badr [19] presented an approach for sentiment analysis, which was realized adopting fastText with recurrent neural network variants to represent textual data efficiently. Its main goal was to improve the performance of common Recurrent Neural Network (RNN) variants regarding classification accuracy and deal with large scale extensive data. Besides, a distributed intelligent system for real-time social big data analytics was introduced. The system was developed to ingest, store, process, index, and visualize the voluminous amount of information in real-time. The proposed system followed distributed machine learning with the proposed model for enhancing decision-making processes. Extensive experiments were conducted on two benchmark data sets (Yelp and Sentiment140) demonstrated that their proposal for sentiment analysis outperformed well-known distributed recurrent neural network variants. Their proposal showed higher accuracy and F-score than other state-of-the-art methods. It significantly outperformed other results on both data sets (i.e. Yelp and sentiment140). Given the obtained results, it is reasonable to conclude that their proposal was able to enhance the performance of several existing methods and thus contribute to more efficient models for big data sentiment analysis.

C. Word Level

Kouloumpis [2] demonstrated the impact that feature selection has on the model's performance. They used linguistic features and existing lexical resources used in micro-blogging to detect the sentiment orientation of Twitter messages. From their work, the researchers concluded that POS (Part-of-Speech) features and features from existing sentiment lexicons were useful but not as the features found in Twitter messages. They also came with the conclusion that the training data will be of less benefit if they include micro-blogging features.

So in Law [13] focused their study on the domain of underperformance in large home appliances precisely dishwashers. They developed two domain-specific dictionaries related to dishwasher defects (sparkle and smoke). Their research was very useful for improving the quality of dishwasher appliances. The authors conducted different experiments to detect the defects in the products. In their first experiment, the Afinn lexicon was used to detect the defects but they concluded that the remaining sentiment analysis techniques performed better than uni-gram, bi-grams and tri-grams (see Table.1). From the second experiment they came with the conclusion that in discovering the defects logistic regression, neural network and decision tree classifiers performed better. Using domain-specific terms suggests that the user was satisfied with the product or design and these terms were having a high effect on all the used models. The best results were achieved by Neural Networks and the negative online reviews had an unfavorable effect on sales, brand reputation, and company profits.

The approach implemented in the work of Malik [14] is a modification of the approach stated in [15]. For their extended study, 100 reviews about the product 'Fit-Bit' were crawled from the well-known e-commerce website amazon.com. For their sentiment classification based on an ontology framework they used different attributes along with their corresponding polarity values i.e. (Battery, 2), (Display, 2),(Accuracy, 2), (Waterproof, -3), and (Synchronization, 1). To capture users' preferences over different product aspects and attributes they used the formula for calculating the OGC values defined in [16] where; orientation of comment based on the sentiment value z, group of customer g, and feature or category of product c. Experimental results were based on a random set of reviews which showed that by assigning weights to the attributes depending upon the priority that the user wishes to assign, it has been observed that the proposed ontology model works effectively. The result obtained proved that when the buyer's choice is the most specific the more the decision-making process is accurate.

Table 1. Identification of the main themes across the research surveyed.

Authors/ year	Task	Domain dependence (Y/N)	Language
Kouloumpis et al. (2011)	Word level twitter sentiment analysis	N	English
Wang et al. (2012)	Sentence level text classification	N	English
Mensil et al. (2014)	Document level text classification	N	English
Duwairi et al. (2014)	Document level text categorization	N	Arabic
Nguyen et al. (2014)	Document level sentiment analysis	N	English
Tripathy et al. (2016)	Document level systematic text classification	N	English
Law et al. (2017)	Word level sentiment analysis	Y	English
Tiwari et al. (2017)	Document level content based analysis	N	English
Alomari et al. (2017)	Sentence level twitter sentiment analysis	N	Arabic
Mohamed et al. (2017)	Sentence level aspect based sentiment analysis	Y	English
Fenna Miedma (2018)	Sentence level sentiment analysis	N	English
Malik et al. (2018)	Word level aspect based sentiment analysis	Y	English
Ankit et al. (2018)	Sentence level sentiment analysis	N	English
Badr et al. (2020)	Sentence level data analytics	N	English

Table.2 Detailed summary of the most recent studies in sentiment analysis.

Researcher's name/ year	Algorithm and used features	Dataset	Dataset description	Highest obtained results
Duwairi et al. (2014)	Naive Bayes SVM Support Vector Machine K-Nearest Neighbour (KNN)	Arabic text corpus	1,000 documents that vary in length and writing style were collected. 10 predefined categories (sports,economic,Internet,art,animals,technology,religion,politics,plants and medicine) where every category contains 100 documents.	Precision SVM 75.25 % Recall KNN 69.04%
Kouloumpis et al. (2011)	Linguistic (n-grams and POS) and lexicons (MPQA and ILD)	HASH Edinburgh EMOT and iSieve	Collection of tweets with positive ':)' and negative ':(' emoticons. It contains 381,381 tweets, 230,811 positive and 150,570 negative.	75%
Wang et al. (2012)	NB, Multinomial Naive Bayes (MNB), SVM, N-gram, Bag of	RT-s: movie reviews CR:Customer review dataset	Short movie reviews dataset containing one sentence per review [6] CR: dataset is processed like in [20] Opinion polarity subtask of the MPQA dataset [21]	MNB-uni (Subj 92.6%) MNB-bi (RT-s 79%, MPQA 86.3%, Subj 93.6%) SVM-bi (MPQA 86.7%)

	Words (BoW)	MPQA Subj dataset RT-2k dataset IMDB	The subjectivity dataset with subjective reviews and objective plot summaries [6] The standard 2000 full-length movie review dataset [6] A large movie review dataset with 50k full-length reviews [7]	NBSVM-uni (RT-s 78.1%) NBSVM-bi (RT-s 79.4%, MPQA 86.3%, CR 81.8%, Subj 93.6%)
Mesnil et al. (2014)	Recurrent Neural Networks (RNNs), NB-SVM, Logistic Regression (LR) with N-grams, tf-idf and word embeddings as features	IMDB movie reviews	A large movie review dataset with 50k full-length reviews [7]	NB-SVM-3-grams (91.87%) N-grams (86.5%) All (92.57%)
Tripathy et al. (2016)	NB, Maximum Entropy ME, SVM, Stochastic Gradient Descent SGD with N-grams and tf-idf features	IMDB movie reviews	A large movie review dataset with 50k full-length reviews [7]	NB 86.23% ME 83.36 % SVM 70.16% SGD 83.36%
Law et al. (2017)	Logistic regression Neural Networks Decision trees 2 domain specific dictionaries (sparkle and smoke related terms) AFINN lexicon N-grams	Product reviews specific to dishwashers		
Nguyen et al. (2014)	SVM Rating-feature (regression model) N-grams	PL04 IMDB dataset	**PL04:** (movie reviews) 1000 pos, 1000 neg **IMDB:** (movie reviews) 50000 reviews.	PL04 91.6% IMDB 89.87%
Tiwari et al. (2017)	SVM Maximum Entropy Naive Bayes N-grams and TF-IDF	Rottentomatoes (reviews and rating)	Contains information about movies: synopsis, rating,genre,director,writer,theatre_date,DVD_date,currency,box_office, runtime,studio.	SVM-unigram 87.53% ME-unigram-bigram 89.64% NB-unigram-bigram 87.08%
Alomari et al. (2017)	SVM, NB with different weights (TF-IDF and TF), Stemming and N-grams	AJGT Arabic Jordanian Twitter corpus	1,800 tweets annotated as positive and negative. Modern Standard Arabic (MSA) or Jordanian dialect.	SVM-accuracy (88.72%, F-score 88.27%) NB-accuracy (83.61%, F-score 84.73)
Fenna Miedma (2018)	LSTM and RNN Word embeddings	IMDB dataset	A large movie review dataset with 50k full-length reviews [7]	86.74%
Malik et al. (2018)	Attribute scores and weights	E-commerce reviews Amazon (Fit-Bit)	(130M+ customer reviews)	
Badr et al. (2020)	RNN Long Short Term Memory (LSTM) Bidirectional Long Short-Term Memory (BiLSTM) Gated Recurrent Unit (GRU) with FastText as word embedding	Yelp Sentiment140	Yelp dataset is composed 6,685,900 classified reviews provided by 1,637,138 users for 192,609 businesses. Sentiment140: Twitter sentiment analysis data set, which is originated from Stanford University. This particular data set consists of 1,600,000 classified tweets. In this work, we have randomly selected 20,000 tweets as the original data set	**Sentiment140** (LSTM-accuracy 78.09%, F-score 78.43%) (BiLSTM-accuracy 87.76%, F-score 78.65%) (GRU-accuracy 78.88%, F-score 79.02%) **Yelp** (LSTM-accuracy 92.55%, F-score 92.69%) (BiLSTM-accuracy 92.91%, F-score 92.95%) (GRU-accuracy 93.28%, F-score 93.10%)
Mohamed et al. (2017)	Naive Bayes Part Of Speech tagging (POS), Chi-Square	SemEval 2014 Task 4 reviews dataset	(product reviews such as laptops/ restaurants)	F-score 78.12%

Ankit et al. (2018)	Naive Bayes Random Forest SVM Logistic Regression Bag of Words (BoW)	Sentiment140 corpus HealthCare Reform (HCR) GOP Twitter sentiment analysis dataset	Sentiment140 corpus This is a Twitter sentiment analysis data set, which is originated from Stanford University. This particular data set consists of 1,600,000 classified tweets. In this work, we have randomly selected 20,000 tweets as the original data Set. HealthCare Reform (HCR) This dataset contains a set of tweets with the #hcr. The tweets with positive sentiment and negative sentiment are considered for the experiment. This dataset consists of 888 tweets (365 positive and 523 negative). GOP: This dataset consists tweets on the first GOP debate for the 2016 presidential nomination. It contains 13871 tweets with the positive, negative or neutral sentiment. Twitter sentiment analysis dataset This dataset has 99989 training tweets. Each tweet is either positive or negative. This dataset consists of 43532 negative and 56457 positive tweets. This dataset is available at Kaggle.	**Sentiment140** (accuracy 75.81%, F-score 75.79%) **HealthCare Reform** (accuracy 73.68%, F-score 70.28%) **GOP** (accuracy 85.83%, F-score 76.85%) **Twitter sentiment analysis** dataset (accuracy 74.67%, F-score 73.33%)

III. RESEARCH GAPS

After a comprehensive study of the research surveyed, several gaps manifested; on one hand, from the work of Kouloumpis [2], we concluded that POS (Part-of-Speech) features and features from existing sentiment lexicons were useful but not as the features found in Twitter messages. We also came with the conclusion that the training data will be of less benefit if they include micro-blogging features. On the other hand, since Twitter comments are mostly small in size, therefore, the proposed approaches in [2], [12], [18] may have some issues while considering these reviews. Another issue remains considering that different reviews or comments contain symbols known as emojis like (:), :(, thumbs up, thumbs down) which help in manifesting user's sentiment, but these being images were not taken into consideration in the surveyed research for further analysis.

As illustrated in Table 1. the work of [13], [14] was domain-dependent where lexicons were manually generated; one was specific to dishwasher's defects and the other was specific to Fit-bit evaluation aspects as shown in Table 2. but the main gap resides on the fact that using a dictionary-based approach has a major disadvantage which is the inability to find opinion words with domain and context-specific orientations. Besides, it is observed that better results are achieved when working on the domain-dependent corpus than working on the domain-independent corpus. Since creating the domain-specific corpus is more complex than using the domain-independent one, there is still a lack of research in the field of domain-dependent sentiment analysis which is known as context-based analysis. Last, to give stress on a word, it is observed that some users often repeat the last character of the word several times such as "greatttt, Fineee". These words do not have real meaning, but they may be examined and further processed to distinguish sentiment. However, this aspect was only considered in the work of [9].

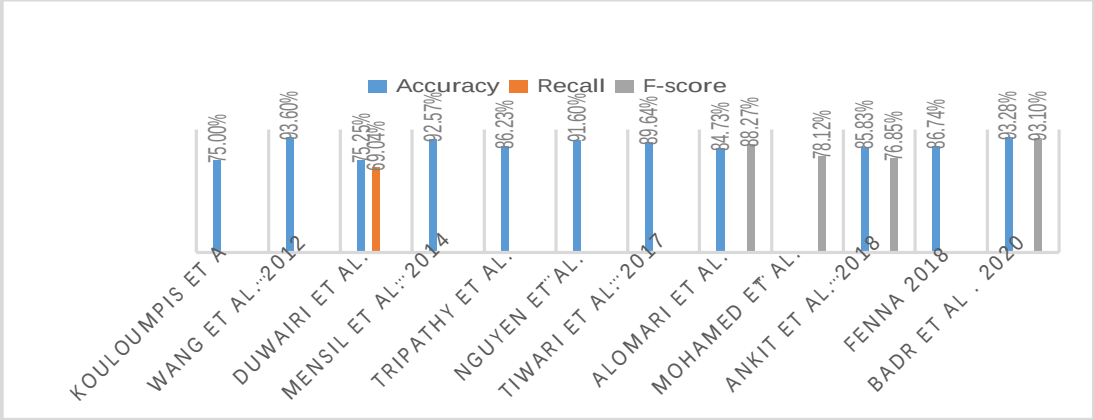
IV. EXPERIMENTAL RESULTS AND DISCUSSION

This review has reported a large selection of studies related to sentiment analysis using supervised learning as summarized in Table 1. After analyzing all these studies, it is established that sentiment analysis can be accomplished more efficiently and accurately by using deep learning methods. The main purpose of sentiment analysis task is to predict users' opinions and deep learning models are all about the prediction of the human mind, thus, these models yield more accuracy than superficial models. Deep learning networks are better than SVMs and standard neural networks because they have more hidden layers as compared to normal neural networks with only one or two layers. Deep learning networks are capable to provide training in both supervised/unsupervised ways. Deep learning models implement automatic feature extraction and do not involve human intervention, therefore, it can save time because feature engineering is not needed. Still, this method suffers from some limitations as well, as compared to previous models such as SVM. It requires large data sets and is tremendously costly to train. These complex models can obtain weeks to train by using machines equipped with expensive GPUs. The data used in sentiment analysis are mostly on product reviews as shown in Table 2. The other kinds of data are news articles or news feeds; web Blogs, social media, and others. We were interested too in seeing if the data used in the surveyed articles was domain-dependent or not Table 1. Sentiment analysis using other languages has engaged many researchers recently as shown in Table 1. These languages include Spanish, Italian, German, Far East languages (Chinese, Japanese); and Middle East languages (Arabic). But still, the English language is the main used language for sentiment analysis due to the following reason; data availability with resources such as lexica, corpora, and dictionaries. This represents a new challenge to

researchers to build lexica, corpora and dictionaries resources for other languages. Some additional insights were drawn while studying recent articles, we have discovered some points that could be considered open problems in research.

1) *The Data Problem*: It has been observed that there is a need for benchmark data sets in the SA field. It was expressed in [25] that a small number of most famous data sets are in the field of sentiment analysis. In Table 2., the most popular data sources and data sets used to accomplish the different tasks of sentiment analysis are presented. It can be noticed that Sentiment140, Health Care Reform (HCR), GOP, PL04, HASH, and iSieve are used in Twitter sentiment analysis articles. IMDB, Yelp, Amazon.com, and rottentomatoes are very famous data sources of review data (Table 2.). These data sets are used in sentiment analysis and sentiment classification tasks.
2) *The Language problem*: It was noticed in the articles presented in this survey that the English language is mostly used in the sentiment analysis field. Appropriately, many data sources are built for this language. There is still a lack of resources for the Middle East languages including the Arabic language. Even though the resources built for the Arabic language are not yet complete and not found easily as an open-source. This makes it a very good trend in research now.
3) *Natural language processing*: The natural language processing tools like Python's NLTK package and Vader lexicon [26] can be used to facilitate the sentiment analysis process. It gives better natural language understanding and thus can help obtain more accurate results. This provides a novel trend of research using NLP as a pre-processing phase before sentiment analysis.

Figure 1. Summary of best achieved performance in terms of accuracy, recall and F-score



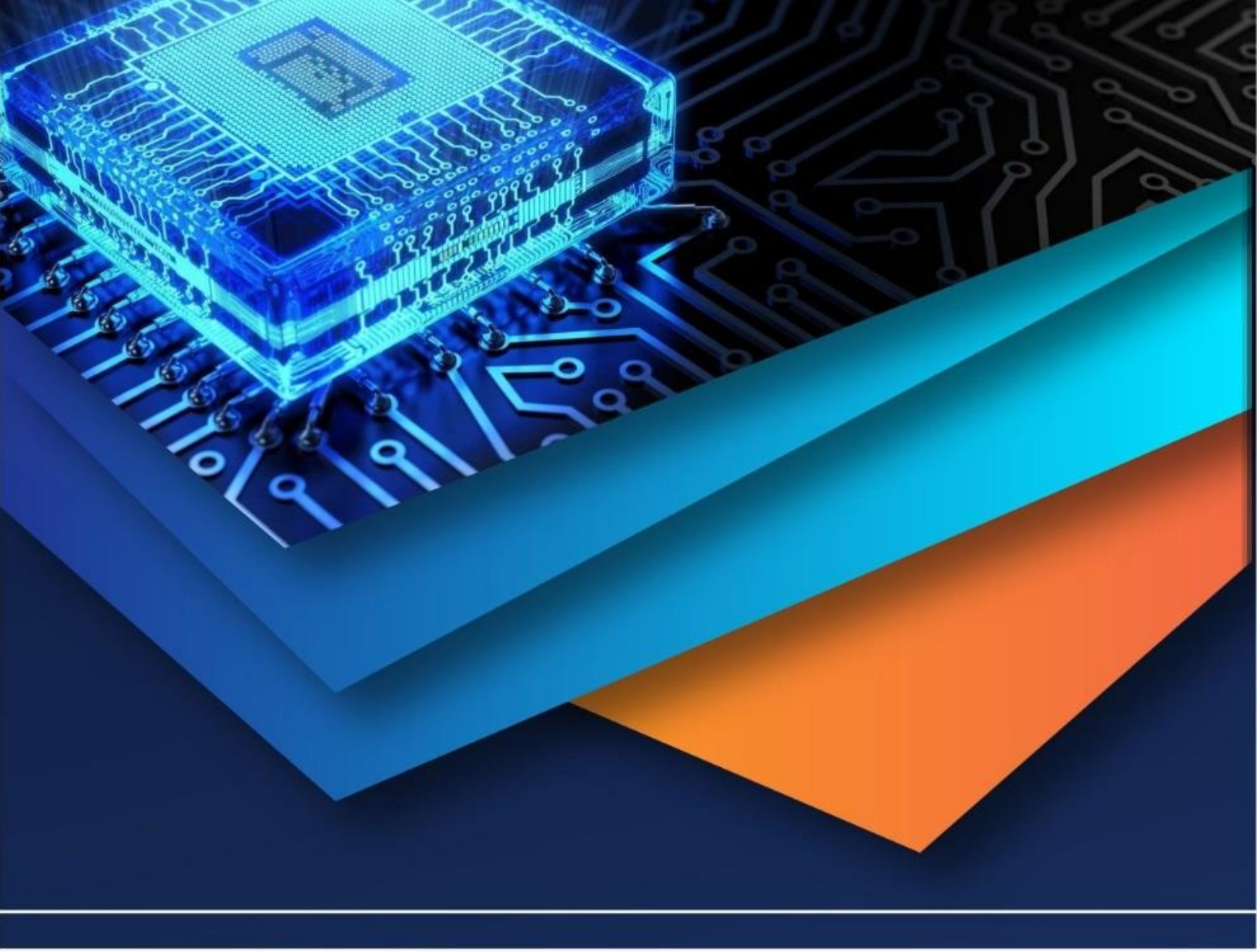
V. CONCLUSION

In this paper, we provide both a survey and comparative study of recent techniques for sentiment analysis including machine learning and lexicon-based approaches. For our study, we considered both cross-domain and cross-lingual methods and some evaluation metrics such as accuracy, recall, and F-score. Research results show that machine learning methods, such as SVM and Naive Bayes have the highest accuracy and can be regarded as the baseline learning methods, while lexicon-based methods are very effective in some cases, which require few efforts in a human-labeled document. We also studied the effects of various features on the classifier. We can conclude that cleaner data, more accurate results can be obtained. The use of the bi-gram model provides better sentiment accuracy as compared to other models. Future work includes an extensive comparison of different text mining and sentiment analysis approaches on different data sets acquired from multiple resources and in multiple languages. We will also work towards finding the most computationally inexpensive algorithms for various tasks and sub-tasks. Various prediction applications will also be studied.

REFERENCES

[1] D. Tang and M. Zhang, 2018. Deep Learning in Sentiment Analysis. In: Deng L., Liu Y. (eds) Deep Learning in Natural Language Processing. Springer, Singapore, 219-253
[2] Kouloumpis, E., Wilson and Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg!. Icwsm, 11 (538-541), 164.
[3] Wang, S. and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification, In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2 (pp. 90-94).
[4] Duwairi. M. and Islam Qarqaz (2014), Sentiment Analysis in Arabic Tweets, In Information and Communication Systems (ICICS), 5th International Conference on IEEE
[5] Nguyen, D. Q., Nguyen, D. Q., Vu, T., & Pham, S. B. (2014). Sentiment classification on polarity reviews: an empirical study using rating-based features. In Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis (pp. 128-135).

[6] Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), pp. 271–278
[7] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis.In Proceedings of the 49th Annual Meeting of theAssociation for Computational Linguistics: Human Language Technologies, Vol 1, pages 142–150.
[8] Mesnil, G., Mikolov, T., Ranzato, M. A. and Bengio, Y. 2014, Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews. arXiv preprint arXiv:1412.5335.
[9] Tripathy, A., Agrawal, A., and Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach, in Expert Systems with Applications, 57, 117-126.
[10] Tiwari, P., Mishra, B. K., Kumar, S., & Kumar, V, 2017. Implementation of n-gram methodology for rotten tomatoes review dataset sentiment analysis., in International Journal of Knowledge Discovery in Bioinformatics (IJKDB), 7 (1), 30-41.
[11] Alomari, K. M., ElSherif, H. M., and Shaalan, K., 2017. Arabic Tweets Sentimental Analysis Using Machine Learning, In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 602-610). Springer, Cham.
[12] Mohamad Syahrul Mubarok, Adiwijaya, and Muhammad Dwi Aldhi, 2017 . Aspect-based sentiment analysis to review products using Naïve Bayes, in AIP Conference Proceedings 1867, 020060, ; https://doi.org/10.1063/1.4994463, Published Online: 01 August 2017
[13] Law, D., Gruss, R., and Abrahams, A. S. (2017). " Automated defect discovery for dishwasher appliances from online consumer reviews. In Expert Systems with Applications, 67, 84-94
[14] Monica Malik, Sharib Habiba and Parul Agarwal, 2018. A Novel Approach to Web-Based Review Analysis Using Opinion Mining, in International Conference on Computational Intelligence and Data Science (ICCIDS 2018), Procedia Computer Science 132, 1202–1209
[15] Haider, S. Z. (2012). An Ontology-Based Sentiment Analysis: A case study (Dissertation).: 1-103
[16] Yaakub, M. R., Li, Y., Algarni, A., & Peng, B. 2012. Integration of opinion into customer analysis model. in Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology: 164-168.
[17] Fenna Miedema 2018, research paper in Business analytics Sentiment Analysis with Long Short-Term Memory networks.
[18] Ankit and Nabizath Saleena 2018, An Ensemble Classification System for Twitter Sentiment Analysis, in International Conference on Computational Intelligence and Data Science (ICCIDS 2018), Procedia Computer Science 132 937–946
[19] Badr Ait Hammou, Ayoub Ait Lahcen, Salma Mouline 2020, Towards a real-time processing framework based on improved distributed recurrent neural network variants with fastText for social big data analytics, in Information Processing & Management, Volume 57, Issue 1, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2019.102122.
[20] Nakagawa, Tetsuji & Inui, Kentaro & Kurohashi, Sadao. 2010. Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables.. NAACL-HLT. 786-794.
[21] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, pages 347–354.
[22] Salloum, S. A., AlHamad, A. Q., Al-Emran, M., & Shaalan, K. 2018. A Survey of Arabic Text Mining. In Intelligent Natural Language Processing: Trends and Applications pp. 417-431, Springer, Cham.
[23] Vishal Kharde and Sheetal Sonawane 2016, Sentiment Analysis of Twitter Data: A Survey of Techniques, In International Journal of Computer Applications, Volume 139 – No.11, pp. 0975 – 8887
[24] NLTK 3.4.5 documentation, http://www.nltk.org/#, last retrieved on 6th November, 2019
[25] Tsytsarau, M. & Palpanas, T. Data Min Knowl Disc (2012) 24: 478. https://doi.org/10.1007/s10618-011-0238-6
[26] Hutto, C.J. & Gilbert, Eric. (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY