An Efficient Model for Mining Outlier Opinions

Neama Hassan[1], Laila A. Abd-Elmegid[2], Yehia K. Helmy[3]
Information Systems Department
Faculty of Commerce and Business Administration
Helwan University, Helwan, Egypt

***Abstract*—In the internet era, opinion mining became a critical technique used in many applications. The internet offers a featured chance for users to express and share their views and experiences anywhere and at any time through various methods as online reviews, personal blogs, Facebook, Twitter and companies' websites. Such treasure of online data generated by users play an essential role in decision-making process and have the ability to make radical changes in several fields. Although the opinionated text can provide significantly invaluable information for the wide community either are individuals, business, or government, the outlier or anomaly opinions could have the same impact but in opposite manner which harm these fields. Consequently, there is an urge to develop techniques to detect the outlier opinions and avoid their negative impacts on several application domains which rely on opinion mining. In this paper, an efficient model for mining outlier opinions has been proposed. The proposed MOoM model, stands for Mining Outlier Opinion Model, offers for the first time the ability to mine outlier opinions from product's free-text reviews. Accordingly, it can help the decision makers to improve the overall sentiment analysis process and perform further analysis on the outlier opinions to get better understanding for them and avoid their negative impact. The proposed model consists of three modules; Data preprocessing module, Opinion mining module and outlier opinions detection module. The proposed model utilizes the lexicon-based approach to extract sentiment polarity from each review in the dataset. Also, it uses the Distance-based outlier detection algorithm to produce a graded list of review holders with outlier opinions. Experimental study is presented to evaluate the proposed model and the results proved the model's ability to detect outlier opinions in the product reviews effectively. The model is adaptable to be used in other fields rather than product's reviews by customizing its modules' layers.**

Keywords*—Opinion mining; sentiment analysis; anomaly detection; outliers; reviews; text analysis; natural language processing; rapidminer

I. Introduction

Data mining targets extracting hidden and implicit information, known as knowledge, from data. To achieve its objective, data mining utilizes various techniques as association rules, clustering, anomaly detection, sequential analysis, and classification. The output of such data mining techniques is used in strategic decision making. A featured technique of data mining known as opinion mining has been issued recently. It can be considered as a special variation of Text Mining where the core content is a set of opinions described through subjective statements [1]. The main objective of opinion mining is to sentimentally analyze opinions, wishes, evaluations, and emotions written by the users in natural language. Accordingly, it utilizes both human and electronic intelligences. An alternative name of Opinion Mining is Sentiment Analysis [2]. There are five components of the sentiment: 1) The target entity it relates to, 2) The specific target entity aspect to which the opinion refers, 3) The opinion holder, 4) The time of expressed opinion, and 5) The polarity of the opinion which is related to the target entity aspect. For example, a hotel review posted on May 1, 2019 says: "as a business traveller, I found the hotel's location to be great," includes the 1) target entity "hotel", 2) the aspect "location", 3) the opinion holder "I", who travels for business, 4) the opinion posting time "May 1, 2019", and 5) the sentiment "great", which reflects a positive polarity. Not all opinions have these five components, according to [3-5]. There are two main approaches used for sentiment analysis: lexicon-based approach and machine learning approach. Each approach has its own advantages and limitations. A Hybrid technique of both approaches can be used to overcome the shortcomings of the individual techniques. The main difference between Lexicon-based approach and machine learning approach lies on the method of developing the lexicon. Lexicon-based approach works with pre-developed lexicon while as machine learning approach develops its own lexicon dynamically through continuous learning from the data [6]. According to the complexity of the derived knowledge from mining opinions there are three levels of opinion mining; Document level, Sentence level, and Aspect level.

Anomaly Detection (AD) which is also known as Outliers Detection (OD) is an important technique of data mining that is mainly utilized in critical applications like Credit Card fraud detection. In such technique the focus is on up-normal data, which do not conform to normal expected behavior, rather than normal data [7]. Highlighting such up-normal data, known as outliers or anomalies, help the decisions makers to accurately assess the working information regardless to such malicious and vexatious data. However, not all the outliers are harmful attacks as they can just represent data with previously unknown and surprising behavior [8]. The techniques of anomaly detection can be classified in three main categories; Unsupervised, Supervised, and Semi-supervised techniques. The unsupervised outliers detetction techniques use unlabled test data and assume that most of the data instances are normal by searching for such instances seem to leastly fit to the reminder of the data set. On the contrary, supervised outliers detection techniques work with lableled date set and utlizes a classifier to learn from the data. The semi-supervised outliers detection techniqueswork with a developed model of normal data set that it uses to test other instances of test data set. Anomalies could be one of three types; *Point anomalies*: where a data point exists too far from the other data points;

Contextual anomalies: the event is anomalous regarding to specific context and it is common type used in time-series data set; *Collective anomalies*: represent a group of anomalous regarding to the whole data set [9]. This paper proposes a model to sentimentally analyze opinions and extract outlier opinions. The proposed model uses dictionary-based opinion mining approach. Such approach consumes fewer resources and can be adapted easily to different types of opinionated text domains. The proposed model also applies the distance-based outlier detection algorithm for anomaly detection.

The research paper is organized as follows: Section two summarizes the related work in both opinion mining and anomaly detection fields. Section Three describes in details the proposed Mining Outlier opinions Model (MOoM). Section four shows our experimental study. The efficiency of the proposed model is tested and measured in Section five. Finally, the conclusion and future work are highlighted in Section six.

II. Related Works

This section reviews the recent presented research work regarding opinion mining and anomaly detection.

A. Opinion Mining

Opinion mining, or sentiment analysis, has been extensively studied recently with different and novel approaches. The research of such field aims to improve its results especially with the rapid growth of using web technology in different aspects of life. In the tourism field, the classification problem of published tourists' comments about their various experiences have been studied as in [10]. The authors have developed and compared various classifiers of different deep learning techniques. The main objective is to help tourists to get benefit from online published comments in developing their own trip plan efficiently. In the education field, a sentiment analysis lexicon-based approach for automatic analysis of students' comments to predict the expected performance level of teachers have been proposed in [11]. The approach gave an attention to the terms of intensifier words and blind negation words in the analysis process. A hybrid approach on mobile reviews have been applied as in [12]. They proved that the accuracy of the hybrid approach is greater than the accuracy of the individual techniques of Naïve Bayes and KNN. A five Twitter data sets used for applying a convolution algorithm to train the deep neural network in [10]. The results have proved that the proposed GloVe-DCNN model has a better accuracy and faster analysis speed.

B. Anomaly Detection

Deferent anomaly detection techniques have been successfully applied across several domains. Generally, the anomaly detection techniques can be classified as machine learning-based, distance-based, and density-based techniques. In [13] a hybrid approach of both clustering-based and distance-based techniques is used to mine efficiently the existing outliers. The experimental study proved the efficiency of the proposed approach over distance-based approach by consuming less computational resources. Various outlier detection algorithms have been used such as K-nearest neighbor in [13] and [14]. Histogram-based outlier score algorithm is used by [15]. an Angle-based anomaly detection method applied for high-dimensional datasets in [16] and many other algorithms which work efficiently on high dimensional datasets.

A framework for the argumentation process to effectively identify outlier opinions of stakeholders have been applied in an argumentation systems in [17]. The proposed framework detects the outlier opinions of stakeholders from both individual and collective viewpoint.

The previous exploration of the most recent research presented for opinion mining and outliers detection highlights the urgent need to propose a model that combines anomaly detection and opinion mining techniques in order to mine outlier opinions. Outlier opinion is the opinion which deviate from the general opinion. Such new knowledge output helps the decision makers to take decisions with minimal effect of outlier opinions.

III. Proposed Model (MOoM)

An Efficient Mining Outlier Opinions Model (MOoM) is proposed to mine outlier opinions from free-text product reviews. A rare research work is dedicated for such point. The proposed model integrates both the sentiment analysis domain and the anomaly detection domain to achieve its objective. Fig. 1 clarifies with an example the various modules of the proposed MOoM model starting with manipulating the available product's reviews and ending by a list of review holders which have outlier opinions.

MOoM model consists of three major modules; data pre-processing module, opinion mining module, and outliers' detection module. MOoM architecture appears in Fig. 2.

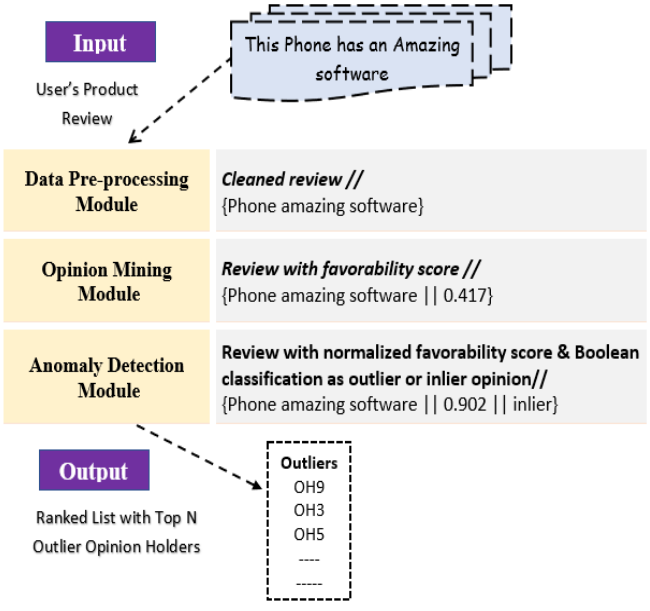


Fig. 1. An Overview of the MOoM Model.

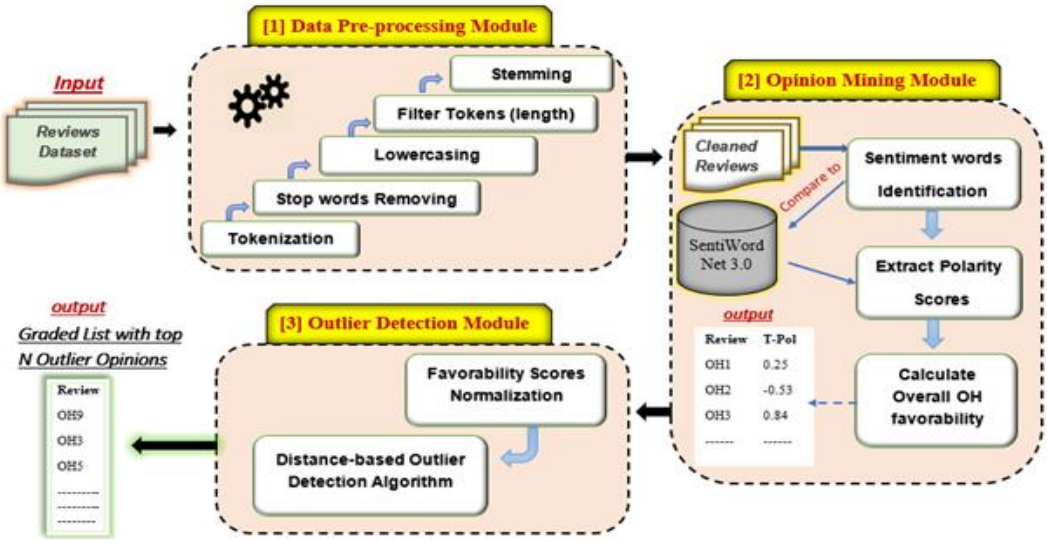


Fig. 2. The Proposed Mining Outlier Opinions Model (MOoM).

A. Data Pre-Processing Module

In this module, the opinion reviews are transformed to cleaned reviews to be suitable for processing in the next module, opinion mining. To pre-processing of the opinion reviews passes through five NLP tasks as described in the following sections.

1) Tokenization: In tokenization step, the large strings of text are divided into tokens which are a small set of words. Bigger chunks of text can be divided into sentences, sentences can be divided into words, etc. Tokenization is done by locating the word boundaries [18]. The sample sentence below shows an example of Tokenization.

This mobile phone is very good.

After tokenization, the output text will be:

This mobile phone is very good advanced processing is commonly conducted after a text reviews has been suitably tokenized.

2) Lower casing: Text transformation into the lower case is a simple and effective way to pre-process the text reviews. It is appropriate for almost problems related to text mining and NLP. lower casing is beneficial when the dataset isn't so big and extremely helps to make data consistency [19]. The lower casing is important to make sure that the word matched to respective feature, for example "AMAZING" & "AmAzing" --- should be converted to "amazing".

3) Stopwords removing: Stop words are a group of words which are commonly used in a specific language. For example, in English a, the, is, her, are, on, of, with, about, what, when, where, that, this, by, be. and etc. are considered as a stopword. The reason behind removing these stop words is that they are valueless and removing them from reviews enables the model to concentrate on the other words which are most important and consequently, achieving a high accurate classification [20]. For example, the next review is acceptable if the stop words are removed:

This mobile phone has an amazing software.

4) Filtering: Performing more cleansing for data is done by removing non-English words and filtering words by their length, where words with length less than minimum length will be removed [21].

5) Stemming: Stemming is the process which used to eliminate the word affixes (circumfixes, prefixes, suffixed, infixes) with the aim of getting root form a word stem. Stemming techniques put word variations like "great", "greatly" and "greatest" to concept of "great" [22].

B. Opinion Mining Module

The sentiment extraction method presented in this paper is depending on a dictionary-based approach for document level sentiment classification task. The cleaned reviews obtained from the previous Data pre-processing module are passed through three steps; identify sentiment words, extract the polarity scores of sentiment words, and finally calculate the overall favorability for each opinion holder. A detailed explanation of each step is given below.

1) Identify sentiment words: The identification of sentiment words is essential to understand the expressed opinions in user reviews. Words which are usually used by the people to express their positive or negative feelings are known as Opinion/sentiment words. Example for positive sentiment words (nice, amazing, and wonderful) and for negative sentiment words (horrible, bad, and terrible). Thus, A part-of-speech (POS) patterns are useful to extract opinionated words. Part of speech tagging is "the process of mapping a word in the text to its corresponding tag [23]. The main aim of doing POS tagging is that adjectives and adverbs would be strong indicators of the opinion of the review, so they help to perform opinion mining because of that the most used opinion words are adjectives and adverbs.

Adjectives are indicated as JJ; Adverbs are indicated as RB.

2) Extract the polarity scores: After the identification of sentiment words in each review document, the next step is to the polarity strength of each sentiment word. For this purpose, 'SentiWordNet' which is a lexical resource for sentiment analysis has been used. SentiWordNet 'SWN' is an opinion lexicon derived from the WordNet database and it is commonly available for the research purposes. In SWN, each word is related to numerical scores which refer to positive and negative opinion information, scores are ranged between -1 and 1. If a word in review document agreed with the word in Wordnet, then its score from SentiWordNet can be used to find its polarity. SentiWordNet is constructed based on a semi-automated process, and simply could be upgraded for incoming versions of WordNet, and for other different languages [24]. In the SentiWordNet database, the terms are categorized based on the parts of speech coming from WordNet, so that be able to apply scores to terms, a part of speech tagging was needed to extract adjectives and adverbs as opinion words then catching their sentiment strengths as the polarity score for each term.

3) Calculate the overall OH favorability: After the identification of sentiment words in a review document and extracting their sentiment polarity scores based on WordNet and SentiWordNet lexicon as explained in the two previous sections, now it is needed to calculate the total score of sentiment polarity for each review in the dataset which represents the opinion holder favorability about the product. The overall polarity score of a document is calculated by the Summation of the polarity of all words in a document divided by the total number of words in the document as shown in "equation (1)".

$$sentscore = \frac{\sum_{i=0}^{n}(Word\_polarity)}{No.of.words} \quad (1)$$

Word polarity // The polarity value returned by SentiWordNet.

No. of. words // The total number of words in the document.

C. Anomaly Detection Module

In the anomaly detection module, the list of opinion holders' favorability scores which obtained from the previous opinion mining module is normalized using appropriate normalization technique to be ready for applying a distance-based outlier detection algorithm which extracts the top K opinion holders with outlier opinions as explained in the following sections.

1) Favorability scores normalization: In an unsupervised anomaly detection, the Normalization process has a private significance because of various attributes in the dataset may have distinct units for measurement [25]. The normalized data produces output with higher accuracy and performance in opposite to the unnormalized data which produces output with lower accuracy and performance according to [12]. Normalization technique is used to scale values in a determined range because all data should lie on similar range to achieve fairly comparison. The proposes MOoM model utilizes the Min-Max normalization technique to normalize the OH favorability scores "equation (2)". The max represents the largest value, the min represents the smallest value and the other values are ranged between (0,1).

$$Xnew = \frac{X - Xmin}{X\,max - X\,min} \quad (2)$$

X // The original data point.

X_{new} // The normalized data point.

X_{max}, X_{min} // The Maximum and minimum data point respectively.

2) Distance based outlier detection algorithm: The Distance based outlier detection algorithm as shown in Algorithm (1) is used to measure dissimilarities of OH favorability cores. This algorithm measures the distance between a data point and its k nearest neighbor. Every data point is ranged based on the distance to its k-th nearest neighbor and the top N data points in this ranked list are defined to be outliers. The K value represents the number of neighbors and the N value represents the number of outliers. Different distance functions can be used to measure the distance between two data points, in our research the Euclidian distance function has been used "equation (3)". The data point which is numerically differ from the other data is considered as an outlier.

Algorithm 1/Distance based outlier detection algorithm

Input Normalized list of OH favorability scores

Output Top N OH with outlier options

Step1 Assign K and N value

K-The number of neighbors

N-The number of outlier parameters

Step2 Compute the Euclidian distance [Eq3] for each point on the basic of its distance to its *k-th* nearest neighbor.

Stpe3 Each point is ranked on the basic of measured distance and the top N points in this ranking are declared to be outliers.

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{n}(\boldsymbol{x_i} - \boldsymbol{y_i})^2} \quad (3)$$

The ranked list with top N outlier opinions helps decision-makers to get a better understanding for the outlier opinions and apply further analysis on these anomalies and their OHs, and this would lead to a noticeable improvement in the overall sentiment analysis process and consequently decision making. The value of N is chosen by decision maker to define and filter the top outlier opinions, so it could be varied in different cases.

IV. Experimental Study

In this section, an experimental study is explained to outline the methodology used for evaluating the proposed model.

1) Tools and dataset: A free educational version of RapidMiner studio (Version 9.5.1) is used. Rapid Miner is a powerful software with open source data science platform which provides an integrated environment for data mining, text mining and machine learning processes [26]. The experiment is conducted on a pc with Microsoft Windows 10 operating system with Intel® Core™ i5- 4200U CPU @ 1.60 GHz with 4.00 GB RAM. To validate the proposed model, the experiment is performed on sample dataset about 100 customer reviews, which is taken from amazon mobile phone reviews dataset in (https://www.kaggle.com/PromptCloudHQ /amazon-reviews-unlocked-mobile-phones) that is one of the popular dataset sites.

2) Data pre-processing module: As illustrated in Fig. 3. *Read Excel* operator is used to input the sample dataset reviews then nominal to text operator used to convert the comment review from nominal to string attribute. To prepare the tested dataset for opinion mining module, the *process document from data* operator which works as a container operator is used. *Tokenize* operator divides the text of a review into a sequence of tokens. The non-letter character mode is used which result in tokens with one single word. *Transform Cases* operator is used to convert all characters in a review to lower case. Then, the noisy words that do not affect the classification task removed from document by using *Filter Stopwords (English)* operator which deletes every token matches a stopword from the built-in stopword list. *Filter Token (by length***)** operator filters tokens based on the number of characters they contain. For the proposed model Minimum number of characters is chosen to be two.

3) Opinion mining module: The experiment has been applied a dictionary-based sentiment analysis approach. *Open WordNet Dictionary* operator is responsible for connecting RapidMiner with WordNet-3.0 dictionary which is stored in a specified directory in pc while defining the wordlist. *Stem (WordNet)* operator used to reduce the length of the words to the minimum length by applying the Porter stemming algorithm and the rule-based replacement for word suffixes. Stem Wordnet uses Wordnet dictionary to define the stem rule. Extract Sentiment operator plays the major role in this module. This operator uses a WordNet 3.0 and a SentiWordNet 3.0.0 database which are connected by Synset IDs to extract sentiment of an input review. This operator allows us to identify the opinionated words by selecting the type of words to be used for calculating sentiment value. In the running experiment, adjectives and adverbs are used as sentiment words. The operator calculates the sentiment of each word to get the total sentiment of a document, where the first meaning of a word has the most influence on a sentiment and each next meaning has less influence on a sentiment. Overall favorability for each opinion holder is then calculated as the average value of all word sentiments as shown in Equation (1). The value of sentiment is in the range (-1.0 to 1.0]) where -1.0 means very negative and 1.0 means very positive. Fig. 4 show the output resulting from opinion mining module.

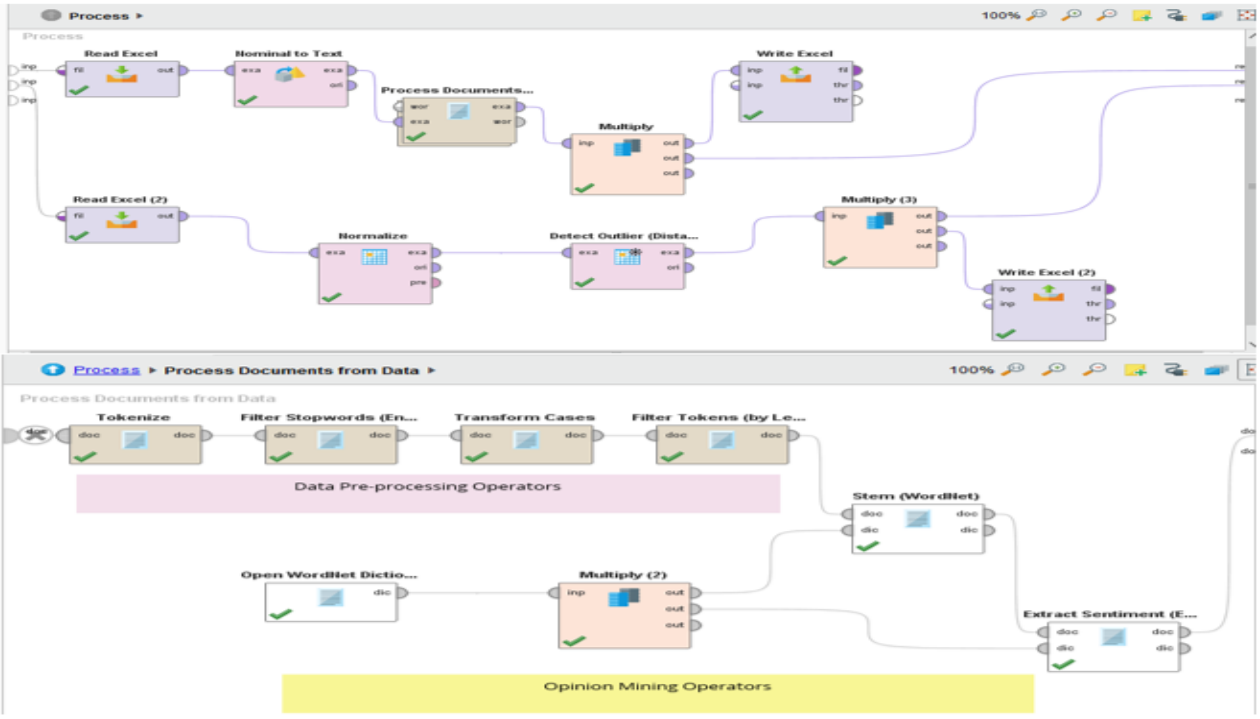


Fig. 3. RapidMiner: Complete Process for MOoM Workflow.

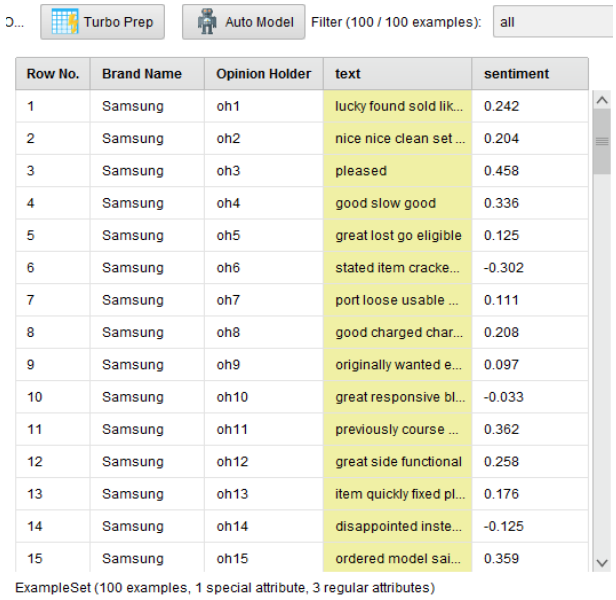
4) Outlier detection module: After getting the overall favorability for each opinion holder in the sample dataset, the *Normalize* operator is used to scale these values so they fit in a certain range. Normalization is very significant when working with attributes of different scales for a fair comparison. This Operator performs normalization with four different methods which are Z-normalization, range transformation, proportion transformation and interquartile range. For the running experiment the simplest range transformation method is used to scale all favorability scores between 0 and 1 as illustrated in Equation (2) for Min-Max normalization technique. Then the Detect Outlier (Distances) operator is used to find opinion holders who have an outlier favorability scores in our sample dataset.

As illustrated in algorithm (1) Distance-based outlier detection algorithm, the operator enables us to identify N outliers based on the distance to their k nearest neighbors. Firstly, the variables N and k are specified through parameters as 7 and 5, respectively. Secondly, measure the distance from each point to its 5 nearest neighbors. The operator provides different functions for distance measurement which are Euclidean distance, Squared distance, Cosine distance, Inverted Cosine distance and Angle.

For the running experiment the Euclidean distance is used as shown in Equation (3).

Thirdly, each data point is ranked based on its measured distance to its 5-th nearest neighbor and the top 7 points in this ranking are specified to be outliers. The operator adds a new Boolean attribute called 'outlier'. If the value of this attribute is true that example is an outlier and vice versa.

Fig. 5 shows the output resulting from outlier detection module. In the outlier attribute the true value means that data point is an outlier and the false value means that data point is an inlier.



Row No.	Brand Name	Opinion Holder	text	sentiment
1	Samsung	oh1	lucky found sold lik...	0.242
2	Samsung	oh2	nice nice clean set ...	0.204
3	Samsung	oh3	pleased	0.458
4	Samsung	oh4	good slow good	0.336
5	Samsung	oh5	great lost go eligible	0.125
6	Samsung	oh6	stated item cracke...	-0.302
7	Samsung	oh7	port loose usable ...	0.111
8	Samsung	oh8	good charged char...	0.208
9	Samsung	oh9	originally wanted e...	0.097
10	Samsung	oh10	great responsive bl...	-0.033
11	Samsung	oh11	previously course ...	0.362
12	Samsung	oh12	great side functional	0.258
13	Samsung	oh13	item quickly fixed pl...	0.176
14	Samsung	oh14	disappointed inste...	-0.125
15	Samsung	oh15	ordered model sai...	0.359

ExampleSet (100 examples, 1 special attribute, 3 regular attributes)

Fig. 4. The Final Output of Opinion Mining Process.

Row No.	Brand Name	Opinion Hold...	text	sentiment	outlier
37	Samsung	oh37	great	0.486	true
56	Nokia	oh56	excellent	1	true
65	Nokia	oh65	excellent	1	true
93	Nokia	oh93	good	0.613	true
98	Nokia	oh98	friendly said amazingly c...	0.342	true
99	Nokia	oh99	fine easy heavy	0.470	true
100	Nokia	oh100	lots found pretty good	0.475	true
1	Samsung	oh1	lucky found sold liked ol...	0.474	false
2	Samsung	oh2	nice nice clean set easy ...	0.470	false
3	Samsung	oh3	pleased	0.624	false
4	Samsung	oh4	good slow good	0.490	false
5	Samsung	oh5	great lost go eligible	0.393	false
6	Samsung	oh6	stated item cracked side...	0.097	false
7	Samsung	oh7	port loose usable sold	0.367	false
8	Samsung	oh8	good charged charged l...	0.391	false

Fig. 5. The Final Output of Outlier Detection Module.

V. Result and Discussion

This section discusses the results obtained from the previous runs of the conducted experiment. This paper has performed the experimental evaluation on a moderate sized dataset as a sample of Amazon mobile phone review dataset. Fig. 6 depicts a 3D scatter plot after the proposed model has been applied. The data points in green represent the top 7 outlier opinion from ranked list and the data points in blue represent the inlier (normal) opinion.

Table I presents a ranked list of opinion holders who have an outlier favorability score. Oh37 ranked 1 in the outlier ranked list. A decision maker can also change the value of top N outlier opinion based on their domains and objectives.

Evaluation of the proposed model is done using a standard evaluation metrics of Recall, Precision and F-measure. Precision and recall are defined in terms of true positive (TP), false positive (FP) and false negative (FN) as shown in the following equations:

$$\text{Precision(p)} = \frac{TP}{TP+FP} \quad (4)$$

$$\text{Recall(R)} = \frac{TP}{TP+FN} \quad (5)$$

$$\text{F} - \text{Measure} = \frac{2\,PR}{P+R} \quad (6)$$

In this paper, we conduct our experiment on sample dataset from Kaggle repository (Amazon-mobile phone reviews dataset). We applied a common data preprocessing steps to obtain a cleaned data ready for further analysis in the opinion mining module. Then we used a dictionary-based approach to perform the sentiment analysis process depending on the SentiWordNet dictionary to obtain the overall favorability score for each opinion holder in our dataset. The results from opinion mining module has been normalized

using a Min-Max normalization technique in the anomaly detection module. After that, the normalized data used as input for the distance-based outlier detection algorithm to produce a ranked list with the top N opinion holders (OH) which have an outlier opinion. The evaluation metrics like precision, recall, and F-measure have been used to measure our models' performance with 86% for precision, 85.7% for recall and F-measure.

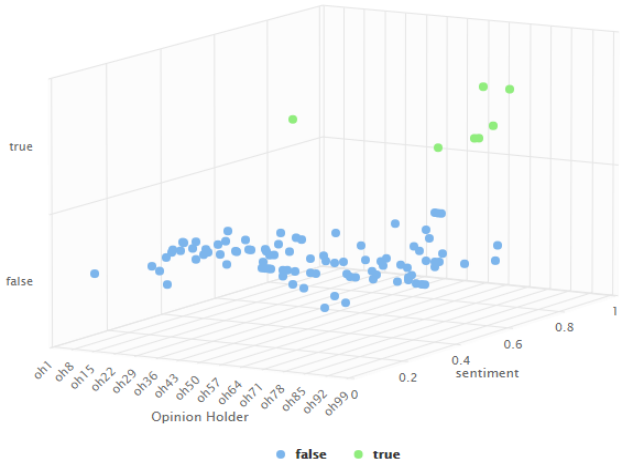


Fig. 6. The 3D Scatter Plot for Outlier Detection Module.

TABLE I. RANKED LIST OF OUTLIER OPINIONS BASED ON A DISTANCE-BASED OUTLIER DETECTION ALGORITHM

Top N opinion holders with outlier opinions
Oh37
Oh56
Oh65
Oh93
Oh98
Oh99
Oh100

TABLE II. SHOWS THE PERFORMANCE'S METRICS OF THE APPLIED MODEL

Performance Indicator	***Result***
Precision (%)	86%
Recall	0.857
F-Measure	0.857

VI. CONCLUSION AND FUTURE WORK

The processing of free–text users' opinions are now being giving more attention according to their critical impact. As such opinions have the ability to make radical changes in most if not all fields, especially with the increasing of the people's ability to share and publish their opinions in an easy manner at anytime and anyplace. The importance of outlier detection is coming from the fact that outliers in data are translated into significant (and often critical) actionable information in a wide variety of application domains. As the opinionated comments give individuals, companies and government very useful informative data which is used in sentiment analysis process for decision making and future forecasts of users' behavior. The anomalies in these opinionated comments could lead to high errors in data analysis and decision-making process. Thus, different application domains which use the opinion mining should be interested with the outlier opinions since they could have a negative impact on their domains. In this paper, we proposed an efficient model for mining outlier opinions (MOoM) to preprocess the opinionated comments and extract the overall favorability score for each opinion holder, then apply a distance-based outlier detection algorithm to generate the top N outlier opinions. To validate the proposed model, a sample dataset from Amazon mobile phone reviews is used. The precision, recall and F-measure of the proposed MOoM model have been evaluated using 100 document reviews. The results of the proposed model can help decision makers to not only analyze the results but also to make more informed decision and decide the best using of this information.

As future work, the following points are considered:

- Various experiments to be conducted using large data sets of different domains.
- Implementation of the hybrid approach for opinion mining process rather than the used dictionary-based approach.
- Exploring different techniques in anomaly detection for mining the outlier opinions.

REFERENCES

[1] Wilson, T., J. Wiebe, and P. Hoffmann, Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational linguistics, 2009. 35(3): p. 399-433.

[2] Kaur, A. and V. Gupta, A survey on sentiment analysis and opinion mining techniques. Journal of Emerging Technologies in Web Intelligence, 2013. 5(4): p. 367-371.

[3] Liu, B., Sentiment analysis and subjectivity. Handbook of natural language processing, 2010. 2: p. 627-666.

[4] Liu, B., Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 2012. 5(1): p. 1-167.

[5] Liu, B., Sentiment analysis: Mining opinions, sentiments, and emotions. 2015: Cambridge University Press.

[6] Paltoglou, G. and M. Thelwall, Sensing social media: A range of approaches for sentiment analysis, in Cyberemotions. 2017, Springer. p. 97-117.

[7] Chandola, V., A. Banerjee, and V. Kumar, Anomaly detection: A survey. ACM computing surveys (CSUR), 2009. 41(3): p. 1-58.

[8] Agrawal, S. and J. Agrawal, Survey on anomaly detection using data mining techniques. Procedia Computer Science, 2015. 60: p. 708-713.

[9] Ahmed, M., A.N. Mahmood, and J. Hu, A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 2016. 60: p. 19-31.

[10] Martín, C., et al., Using Deep Learning to Predict Sentiments: Case Study in Tourism. Complexity, 2018. 2018.

[11] Aung, K.Z. and N.N. Myo. Sentiment analysis of students' comment using lexicon based approach. in Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference on. 2017. IEEE.

[12] Campos, G.O., et al., On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data Mining and Knowledge Discovery, 2016. 30(4): p. 891-927.

[13] Pachgade, M.S. and M.S. Dhande, Outlier detection over data set using cluster-based and distance-based approach. International Journal of Advanced Research in Computer Science and Software Engineering, 2012. 2(6).

[14] Su, M.-Y., Real-time anomaly detection systems for Denial-of-Service attacks by weighted k-nearest-neighbor classifiers. Expert Systems with Applications, 2011. 38(4): p. 3492-3498.

[15] Goldstein, M. and A. Dengel, Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. KI-2012: Poster and Demo Track, 2012: p. 59-63.

[16] Zhang, L., J. Lin, and R. Karim, An angle-based subspace anomaly detection approach to high-dimensional data: With an application to industrial fault detection. Reliability Engineering & System Safety, 2015. 142: p. 482-497.

[17] Arvapally, R.S., et al., Identifying outlier opinions in an online intelligent argumentation system. Concurrency and Computation: Practice and Experience, 2017: p. e4107.

[18] Mayo, M., A general approach to preprocessing text data. KDnuggets, https://www. kdnuggets. com/2017/12/general-approach-preprocessing-text-data. html, 2017.

[19] Vijayarani, S., M.J. Ilamathi, and M. Nithya, Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks, 2015. 5(1): p. 7-16.

[20] Riloff, E. Little words can make a big difference for text classification. in Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. 1995.

[21] Baldwin, T., et al. How noisy social media text, how diffrnt social media sources? in Proceedings of the Sixth International Joint Conference on Natural Language Processing. 2013.

[22] Asian, J., Effective techniques for Indonesian text retrieval. 2007.

[23] Saharia, N., et al. Part of speech tagger for Assamese text. in Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. 2009. Association for Computational Linguistics.

[24] Ohana, B. and B. Tierney. Sentiment classification of reviews using SentiWordNet. in 9th. it & t conference. 2009.

[25] Kandanaarachchi, S., et al., On normalization and algorithm selection for unsupervised outlier detection. Data Mining and Knowledge Discovery, 2020. 34(2): p. 309-354.

[26] Hofmann, M. and R. Klinkenberg, RapidMiner: Data mining use cases and business analytics applications. 2016: CRC Press.