Webology, Volume 16, Number 1, June, 2019

A Semantic Approach for Outlier Detection in Big Data Streams

Hussien Ahmad

Researcher, Faculty of Information Technology Engineering, Damascus University, Baramkeh, Damascus, Syria. ORCID: 0000-0003-4143-3833. E-mail: hussien.ahmad1@damascusuniversity.edu.sy

Salah Dowaji

Associated Professor, Faculty of Information Technology Engineering, Damascus University, Baramkeh, Damascus, Syria. ORCID: 0000-0003-0044-7040. E-mail: salah.dowaji@damascusuniversity.edu.sy

Received March 12, 2019; Accepted June 25, 2019

Abstract

In recent years, the world faced a big revolution in data generation and collection technologies. The volume, velocity and veracity of data have changed drastically and led to new types of challenges related to data analysis, modeling and prediction. One of the key challenges is related to the semantic analysis of textual data especially in big data streams settings. The existing solutions focus on either topic analysis or the sentiment analysis. Moreover, the semantic outlier detection over data streams as one of the key problems in data mining and data analysis fields has less focus. In this paper, we introduce a new concept of semantic outlier through which the topic of the textual data is considered as the primary content of the data stream while the sentiment is considered as the context in which the data has been generated and affected. Also, we propose a framework for semantic outlier detection in big data streams which incorporates the contextual detection concepts. The advantage of the proposed concept is that it incorporates both topic and sentiment analysis into one single process; while at the same time the framework enables the implementation of different algorithms and approaches for semantic analysis.

Keywords

Outlier detection; Big data; Big data stream; Distributed data streams; Graph data streams; Content-based outlier; Context-aware outlier

Introduction

In recent years, the technology revolution has enabled us to gather massive amounts of data from different sources like social networks, sensor data, scientific data, biological data, networked systems data, etc. Social networks and micro-blogging platforms have contributed to an important portion of the big datasets being produced. Such datasets have been an essential source for analysis for different purposes in many fields of research and industry.

Semantic analysis, topic extraction, sentiment analysis and opinion mining have been investigated deeply using social data. Many approaches and methods have been introduced to handle the problem of topic detection in online data or to detect the opinion of the crowd towards a specific issue. Twitter and other micro-blogging and social networks datasets formed a decent source for such research. However, less attention has been paid on mutual analysis of both topic extraction and sentiment analysis at the same time specially when coming to outlier detection in textual online datasets.

Outlier detection and analysis is an important data mining problem that aims to find anomaly points and behavior in data sets (Zhang, 2008). Although, this issue has highly been studied in a broad set of disciplines, there is a lack of focus on semantic analysis of the datasets in order to detect outliers. Some existing approaches focus mainly on sentiment analysis for outlier detection purposes, while some others adopt the concept of content analysis in order to detect outliers. This entails that existing approaches focus on the content of datasets during the semantic analysis process, while in outlier detection problem, the context and conditions under which the data has been generated have a great effect in decision making to consider one data point as outlier or not. Moreover, those methods focusing on semantic outlier detection were implemented in different settings and disciplines with little attention paid to Big Data streams settings. Also, the existing solutions do not handle both semantic analysis and the context at the same time while addressing the outlier detection problem.

In this paper, a new concept of semantic outliers is introduced, taking into consideration the content and context settings and using the semantic analysis methods related to both topic extraction and sentiment analysis. While topic extraction methods are used to detect outlier topics in the data stream content, the sentiment analysis methods are also applied to the data stream itself to detect the implicit context under which that content has been produced. Then, we build on the context-aware outlier detection framework in Big Data streams presented in our previous work to implement the new concept of semantic outliers with the ability to incorporate explicit contextual parameters (Ahmad & Dowaji, 2018). The design of the framework permits the implementation of different analysis methods for both topic and sentiment analysis.

The remaining sections of this paper are organized as follows: The "Related Work" section reviews relevant work, conducted in the field of semantic outlier detection and semantic analysis in general. Then "Our Approach" section describes the new concept of semantic outlier and how to implement it using the context-aware outlier detection framework. Experiments conducted and their results are presented in "Experiments". Finally, the paper concludes with conclusions and recommendations for future work.

Related Works

The issue of both semantic analysis and outlier detection has been heavily addressed in different disciplines of research and industry. However, little attention was paid to the problem of semantic outlier detection which is key problem in the Big Data era, where massive amounts of social data is being produced and need to be analyzed in real-time or near real-time settings.

Although, there are some attempts to tackle the semantic outlier detection problem, those attempts lacks to comprehensive approach to address this problem taking into consideration (i) the content of the dataset, (ii) the context under which the data has been produced and (iii) to run the methods in different settings specially in Big Data streams settings benefitting from the new Big Data tools and analysis methods.

In order to comprehensively address the literature related to semantic outlier detection we review both outlier detection methods with a component of semantic analysis and the semantic analysis methods related to topic detection and sentiment analysis. Such review will inform the future step of building a comprehensive approach for semantic outlier detection in Big Data streams.

Semantic outlier detection approaches handled the problem from different aspects. Mahapatra et al. (2012) propose an approach for outlier detection in large corpus of documents in order to minimize the false alarms based on the semantic similarity analysis of the content of the context parameters. This approach focuses only on the topic of the dataset; the semantic similarity analysis was implemented using two different methods: WordNet and the internet implementing the Normalized Google Distance. On the other hand, Cambria and Melfi propose an outlier detection method based on the sentiment analysis using concept-level sentiment analysis (Cambria & Melfi, 2015). Unlike the traditional statistical methods, Cambria and Melfi used sentic computing on a vector space model of affective common-sense knowledge to work with natural language at concept level and by applying the Least Absolute Deviations on Affective Space they were able to detect the semantic outliers. Also Golmohammadi (2016) uses the sentiment analysis of twitter dataset related to stock market of Oil industry sector. The stock market data was used as content and then the twitter dataset was used as a context for the stock market data. The sentiment analysis of twitter dataset enabled the detection of some false alarms.

None of the mentioned approaches has addressed the issue of semantic analysis of both the content and the context of the dataset or applied the two aspects of semantic analysis: sentiment and topic analysis on the full dataset; rather, they addressed part of semantic outlier detection issue. None of them has been implemented and tested against massive online datasets. However, the sentiment analysis method proposed by Cambria and Melfi forms a good basis for any comprehensive semantic outlier detection approach. Also, the adoption of semantic similarity analysis using WordNet has proved efficient results according to (Mahapatra, 2012).

While there are few researches addressing the semantic outlier detection, there is considerable research focus on the two parts of the problem separately: the outlier detection and the semantic analysis addressing different disciplines and settings. The outlier detection methods have been addressed in previous work handling the problem from different aspects and proposing a novel framework for outlier detection in Big Data streams based on both content and context parameters of the data stream (Ahmad & Dowaji, 2018).

Sentiment analysis has been addressed widely in different research papers. Saif et al. (2016) proposed SentiCircles which is a contextual semantics for sentiment analysis of Twitter using dynamic assignment of strength and polarity of words in sentiment lexicons according to the co-occurrence patterns of words in different contexts in the tweets. This is different from the typical lexicon-based approach which uses predefined static sentiment polarities. Wan and Gao (2015) used an ensemble sentiment classification strategy based on Majority Vote method of multiple classifiers including Naïve Bayes, SVM, Bayesian Networks, C4.5 Decision Tree and Random Forest algorithms. This method was applied on Twitter data for airline services analysis and tested on a static dataset of 12864 tweets. While such method enhances the accuracy of sentiment analysis since it incorporates five different methods, it might jeopardize the performance and efficiency in online data stream settings. However, it is still appropriate to use such methods for offline validation of the online analysis.

Predefined and manually constructed sentiment dictionaries were also used to enhance the sentiment classification process. Nagy and Stamberger used SentiWordNet 3.0 to detect the basic sentiment in Twitter data and complemented it by adding a comprehensive list of emoticons, a sentiment based dictionary, and a list of out-of-vocabulary words that are popular (Nagy & Stamberger, 2012). They applied this method to detect the crowd sentiment during disasters and crises. Also, Li et al. had manually constructed two sentiment dictionaries: Harvard IV sentiment dictionary and Loughran-McDonald financial sentiment dictionary in order to be used in sentiment analysis of news as to measure the impact of news on stock price.

Stanford NLP research group has conducted deep research on sentiment analysis as part of natural language processing effort and provided a toolkit ready to use and implement for different natural language processing tasks including sentiment analysis (Manning et al., 2014). Other research articles explained some aspects of the implemented sentiment analysis approach (Hamilton, 2016; Socher, 2013; Wang & Manning, 2012). Further, Jurafsky et al. (2014) used Stanford NLP sentiment analysis to detect customer sentiment in online restaurant reviews.

Medhat et al. (2014) conducted a survey on sentiment analysis algorithms and applications, also Bravo-Marquez et al. presented a comprehensive review on sentiment analysis providing a meta-level sentiment models for Big Social Data analysis (Bravo-Marquez et al., 2014).

Besides the research on sentiment analysis, the topic extraction has been widely explored. Yoon et al. used natural language processing model for topic extraction related to diseases, and provided a web-based system that monitors and extracts relevant information to certain diseases (Yoon et al., 2018). Similarly, Nakanishi et al. proposed a method to extract the most important topics in the flow of conversation in meetings (Nakanishi et al., 2017); also Zhang and He proposed a method based on Reinforced Knowledge LDA to extract topics of social media events (Zhang & He, 2018). On the other hand, Patil et al. (2017) used data mining algorithms such as Naïve Bayes and Apriori to detect the rare topics in a corpus of tweets; a preprocessing step is mandatory to remove the stop words from the tweet and also to remove the duplicated words between the series of tweets of the same user then it comes to the session identification step followed by Sequential Topic Patterns (STP) and finally detecting the user rate sequential topic patterns. The performance of this method is relatively weak since it uses the data mining algorithms that require multi scan of the dataset which is not applicable in online and stream settings.

Topic extraction and semantic similarity measures have been made available through Stanford NLP and WordNet libraries providing different similarity measures such as WUP, JCN, LCH, LIN, RES, PATH, LESK and HSO (Miller, 1995).

While outlier detection and semantic analysis have a plenty of methods and approaches in different domains, the big question is still on how to use semantic analysis in a comprehensive outlier detection method that is capable of running in different domains and settings especially for Big Data streams. The second question is how to inject the semantic analysis into the outlier detection framework regardless of the implemented semantic analysis methods in order to provide the ability to choose the semantic analysis method by the data analyst based on the nature of data and the problem under investigation.

Research Approach and Methods

Semantic Outlier Concept

The primary goal is to detect outliers and abnormal data points in Big Data streams that consist mainly of textual data such as Twitter stream, online reviews, and other social media streams. Outlier detection process should consider both stream content and context at the same time (Ahmad & Dowaji, 2018). In textual data stream, it is obvious that the text itself represents the content of the data stream, while the context becomes more complex to consider. There are two types of contextual parameters: (i) explicit parameters which are the other data parameters connected to the stream itself like the timestamp, the weather status, location, etc. and (ii) implicit parameters which represent the condition and status of the content creator at the creation time; those parameters can be inferred from the text itself using the sentiment analysis methods in order to detect the emotional conditions under which the creator has produced the content.

Definition (1): Semantic Outlier: is a data point that lays far from other data points in the same set based on the analysis and measures calculated using semantic analysis techniques on both the content and context of that point.

The definition of semantic outlier implies conducting semantic analysis of the content and context of each data point. The content analysis is to detect what a data point is referring to regardless any other factors; thus, the content semantic analysis can be defined as:

Definition (2): Content Semantic Analysis: is the detection of what a text is referring to using the semantic relation and meaning of the words composing the text. In other words, it is the topic that the text is referring to.

While the context semantic analysis should detect the conditions that are related to the text itself. The semantic analysis should be applied on both the implicit and explicit context parameters at the same time; furthermore, it should be applied on the semantic of the parameter itself rather than on the values. This advance semantic analysis requires robust semantic spaces provide besides the basic semantic measures the ability to incorporate the experiment settings with the absolute semantic meaning; for instance, the concept temperature differs when used in twitter stream talking about tourism from when used in the context of sensor readings in spaceship and thus the high temperature concept differs between the two cases. In this paper, we limit the context semantic analysis to the implicit context which represents the emotional status of the text creator at the creation time. Thus, the implicit context semantic analysis can be defined as:

Definition (3): Implicit Context Semantic Analysis: is the detection of the emotional conditions that are related to the data point. In other words, it is the sentiment analysis of the text to detect the emotional status reflected in the text.

According to the previous definition, the concept of semantic outliers can be summarized as one data point is considered outlier semantically when it lays far from other data points using the semantic measures of both topic and sentiment which represent the content and implicit context respectively. The concept of semantic outliers can be implemented using the context-aware outlier detection framework and use the outlier scoring mechanism (Ahmad & Dowaji, 2018) and thus it can incorporate the explicit context parameters.

Semantic Outlier Detection

The semantic outlier detection process should incorporate both content and context semantic analysis to calculate the distance between data points themselves and between the training dataset and the data point under processing. In previous work, we have presented a comprehensive framework (Figure 1) for outlier detection in Big Data streams that incorporates both content and context parameters in the calculation of the anomaly score of each data point based on local and global measures.

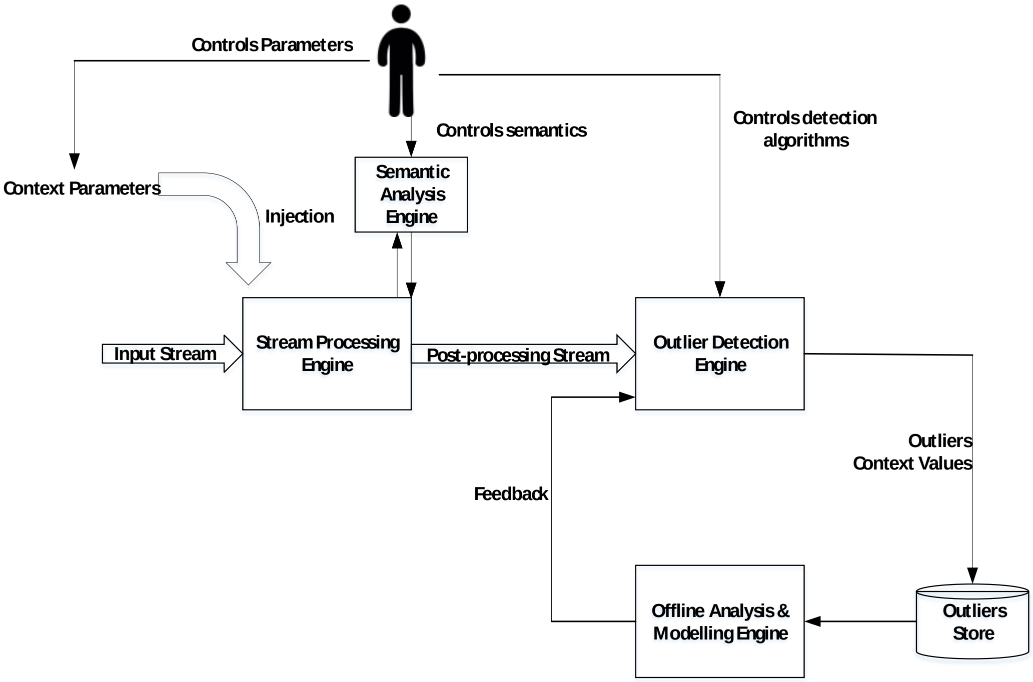

Figure 1. Conceptual framework for outlier detection in Big Data streams (Ahmad & Dowaji, 2018)

The framework has a semantic analysis engine that participates in the data stream pre-processing. The semantic analysis engine has two main objectives:

1- Topic extraction: The engine has to implement a topic extraction method and run a calculation measure against the topics of the training dataset. The training dataset consists of representative sample of keywords which represents all normal topics. In the current implementation of the framework, we use WordNet similarity measures WUP, JCN, LCH, LIN, RES, PATH, LESK and HSO. Other measures can be used for different implementations and under different settings. The output of topic extraction functionality is the semantic distance between the data point and the reference dataset.

2- Sentiment Analysis: The engine has to implement a sentiment analysis method to calculate the sentiment polarity of the data point which represents the implicit context parameter. In the current implementation of the framework, we use StanfordNLP sentiment analysis engine which returns a polarity measure of five scales.

Both the topic extraction and sentiment analysis components use a reference dataset which is the milestone of the analysis process and is used as well as reference for the training dataset. The purpose of reference dataset is to define a reference point acting as zero point for all measures. The selection of the reference dataset depends on the nature of the problem under investigation.

Figure 2 shows the implementation of the semantic analysis engine as part of the stream pre-processing stage. Both topic extraction and sentiment analysis can be run in parallel in order to enhance the performance.

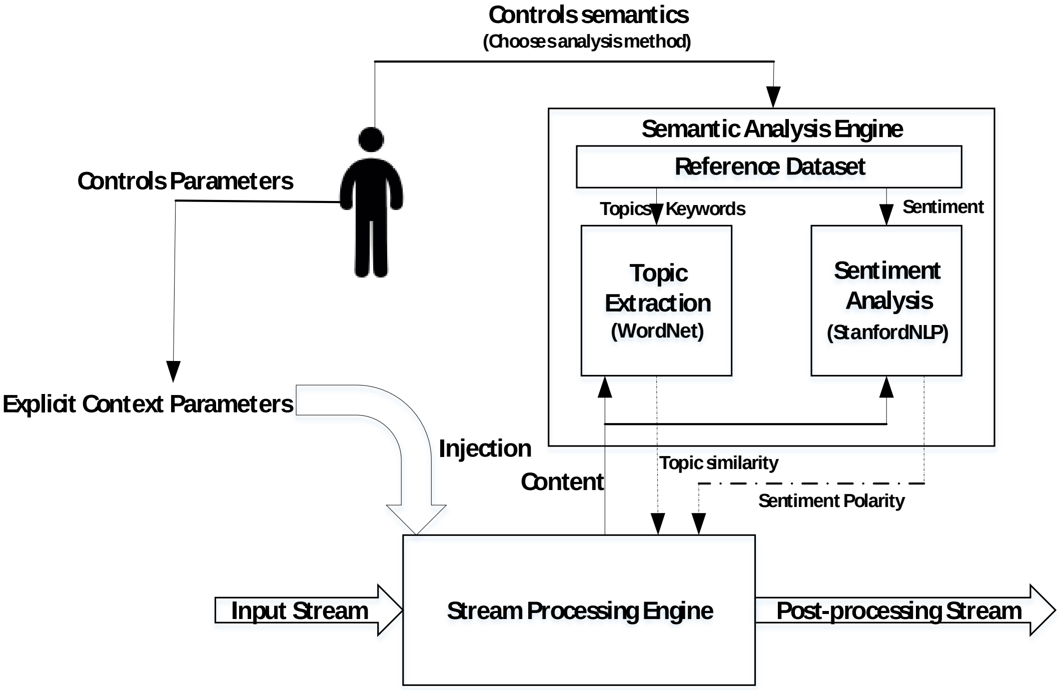

Figure 2. Semantic Analysis Engine

The usage of WordNet similarity measures and StanfordNLP sentiment analysis engine for the current implantation of the framework is due to the wide adoption of both libraries in academic and industry. The design of the framework allows the implementation of other semantic analysis engines according to the need and the settings of dataset and experiment.

Discussion

The semantic analysis engine works as part of the context-aware outlier detection framework proposed in (Ahmad & Dowaji, 2018) and more specifically under the stream pre-processing stage. The primary role of the semantic analysis engine is to transform the input textual data into two numeric values: (i) the semantic similarity distance between the input data and the training data representing the content parameter of the stream and (ii) the sentiment polarity of the input data representing the implicit context parameter. As shown in figure 2, the stream processing engine provides the input to semantic analysis engine which return two numeric values; then the stream processing engine incorporates these values representing the content and one context parameter along with other explicit context parameters into the micro batch that represents the sliding window to be delivered to the outlier detection engine.

The output of the semantic analysis engine represents both the deviation of the input content from the training dataset and the sentiment polarity of input content. The possible variations of the output can be summarized in Table 1.

Table 1. Output variations of Semantic Analysis Engine

Content: Input deviation from the training dataset	Context: Input sentiment polarity	Description
Content is similar to the training dataset	sentiment polarity is similar to training dataset sentiment polarity	Same topic with similar sentiment polarity; this represents a normal data point according to both content and context **[False-False]**
Content is similar to the training dataset	sentiment polarity is different from training dataset sentiment polarity	Same topic with different sentiment polarity; this represents a potential hidden outlier **[False-True]**
Content is different from the training dataset	sentiment polarity is similar to training dataset sentiment polarity	New topic with similar sentiment polarity; this represents an explicit outlier **[True-False]**
Content is different from the training dataset	sentiment polarity is different from training dataset sentiment polarity	New topic with different sentiment polarity; this represents a potential false alarm since it appears as an outlier according to the content parameter while if putting it in the context it might be a normal point **[True-True]**

The proposed architecture guarantees full integration of any semantic analysis method in the semantic analysis engine and thus the outlier detection process does not depend on the semantic analysis engine itself; rather it uses the output of the semantic analysis methods and then applies a fully independent outlier detection method.

The selection of the semantic analysis methods for both topic extraction and sentiment analysis depends on the nature of the problem under investigation and preferences of the data analyst who controls the detection process.

Experiments

The semantic analysis engine was implemented using WordNet 3.0 for topic extraction and semantic similarity measures and StanfordNLP for sentiment analysis. The engine was implemented as part of the context-aware outlier detection framework which was implemented using Kafka 2.11 and Spark 2.3.0. The code has been written in Java and compiled using jdk1.8.0. All tests were run on 2.30GHz core I7, 12G RAM running Windows 10. The tests were run in a single processing node configuration for both Spark and Kafka.

The test was conducted using a specific purpose dataset which was designed to demonstrate the features and capabilities of the proposed model and framework. The “Student Behaviour” dataset has 100 instances consisting of sentences describing good student behaviour at school; 5 sentences described oil prices in negative and positive ways used to demonstrate different topic and 5 sentences described the student behaviour in negative way to demonstrate different sentiment. The training dataset consisted of 25 instances out of the 100 instance of the dataset while the test dataset had 75 instances including the 10 potential outliers.

Other several tests were conducted using predefined datasets from UCI machine learning repository. Two datasets were used for the test, the “Eco-Hotel” with 400 instances and “Health News in Twitter” using the BBC dataset with 3929 instances. Both data were designed for different purposes; thus, there is no pre-judgment on the semantic outliers in these datasets, however, the test was run to demonstrate the full functionality of the framework while the “Student Behaviour” dataset was used to demonstrate the accuracy of the model.

The test using the “Student Behaviour” dataset showed the ability of the model to detect all outlier from different types, however, the dataset was designed for the research purpose and the results cannot be generalized for real world datasets. The accuracy of the model depends mainly on the used semantic analysis techniques which are in the current situation WordNet and StanfordNLP. The ability of using different semantic analysis techniques enables the model to form a basis for any future implementation of semantic outlier methodology. In addition, being part of the context-aware outlier detection framework which was tested heavily in previous work guarantees the efficiency in outlier detection using the two phases' outlier detection method.

The other tests were conducted to demonstrate the ability of the framework to handle real world datasets. The performance of the model depends as well on the used semantic analysis techniques in addition to the size of the instance. Running the tests on described machine showed average performance of the model especially with “Eco-Hotel” dataset which has lengthy data points. In order to enhance the model performance, the used semantic technique should be parallelized and run on a cluster of Kafka nodes.

Conclusion

In this paper, we propose a new concept of semantic outliers that incorporates both content and context parameters into the outlier detection process. The new concept has two main aspects: the semantic content analysis based on the topic extraction and related similarity measures and the semantic context analysis based on the sentiment analysis of the content which reflects the condition and emotional status of the content creator. The implantation of this new concept in outlier detection in Big Data streams has been done using the context-aware outlier detection framework and based on the WordNet similarity measures and StanfordNLP sentiment analysis for content and context analysis respectively. The design of the framework allows the implantation of different semantic analysis methods.

A major challenge faced the test of the proposed model was the lack of well-designed datasets for the semantic outlier detection purposes or to have pre-judgment on existing datasets. However, the model was tested using a specific purpose dataset which was designed for this research and other tests were conducted to demonstrate the full functionalities of the framework.

In future, we plan to implement more semantic analysis methods that fit with different problems in order to provide a wide range and comprehensive semantic analysis engine. Also, the semantic analysis of the explicit context and the semantic of the context parameter in the experiment settings should be studied and included within the semantic outlier detection.

References

Ahmad, H., & Dowaji, S. (2018). A novel framework for context-aware outlier detection in big data streams. *Journal of Digital Information Management*, *16*(5), 213-222.

Bravo-Marquez, F., Mendoza, M., & Poblete, B. (2014). Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems*, *69*, 86-99.

Cambria, E., & Melfi, G. (2015, April). Semantic Outlier Detection for Affective Common-Sense Reasoning and Concept-Level Sentiment Analysis. In *FLAIRS Conference* (pp. 276-281).

Golmohammadi, S. K. (2016). *Time series contextual anomaly detection for detecting stock market manipulation*. Doctoral dissertation, University of Alberta.

Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* (Vol. 2016, p. 595). NIH Public Access.

Jurafsky, D., Chahuneau, V., Routledge, B. R., & Smith, N. A. (2014). Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*, *19*(4).

Mahapatra, A., Srivastava, N., & Srivastava, J. (2012). Contextual anomaly detection in text data. *Algorithms*, *5*(4), 469-489.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55-60).

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093-1113.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Nagy, A., & Stamberger, J. (2012, April). Crowd sentiment detection during disasters and crises. In *Proceedings of the 9th International ISCRAM Conference* (pp. 1-9).

Nakanishi, T., Okada, R., Tanaka, Y., Ogasawara, Y., & Ohashi, K. (2017, July). A topic extraction method on the flow of conversation in meetings. In *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)* (pp. 351-356). IEEE.

Patil, B., Takmare, S., Mirajkar, R., & Kharade, P. (2017). Mining users rare sequential topic patterns from tweets based on topic extraction. *International Research Journal of Engineering and Technology*, 4(9), 680-683.

Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management*, *52*(1), 5-19.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631-1642).

Wan, Y., & Gao, Q. (2015, November). An ensemble sentiment classification system of twitter data for airline services analysis. In *2015 IEEE international conference on data mining workshop (ICDMW)* (pp. 1318-1325). IEEE.

Wang, S., & Manning, C. D. (2012, July). 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). Association for Computational Linguistics.

Yoon, J., Kim, J. W., & Jang, B. (2018). DiTeX: Disease-related topic extraction system through internet-based sources. *PloS one*, 13(8), e0201933.

Zhang, J. (2008). *Towards outlier detection for high-dimensional data streams using projected outlier analysis strategy* .Doctoral dissertation, Dalhousie University Halifax.

Zhang, X., & He, R. (2018, August). Topic Extraction of Events on Social Media Using Reinforced Knowledge. In *International Conference on Knowledge Science, Engineering and Management* (pp. 465-476). Springer, Cham.

Bibliographic information of this paper for citing:

Ahmad, Hussien, & Dowaji, Salah (2019). "A semantic approach for outlier detection in big data streams." *Webology*, 16(1), Article 186. Available at:
http://www.webology.org/2019/v16n1/a186.pdf

Copyright © 2019, Hussien Ahmad and Salah Dowaji.