Twitter Sentiment Analysis Approaches: A Survey

https://doi.org/10.3991/ijet.v15i15.14467

Omar Y. Adwan (✉), Marwan Al-Tawil, Ammar M. Huneiti,
Rawan A. Shahin, Abeer A. Abu Zayed, Razan H. Al-Dibsi,
University of Jordan, Amman, Jordan
adwanoy@ju.edu.jo

Abstract—Twitter is one of the most popular microblogging and social networking platforms where massive instant messages (i.e. tweets) are posted every day. Twitter sentiment analysis tackles the problem of analyzing users' tweets in terms of thoughts, interests and opinions in a variety of contexts and domains. Such analysis can be valuable for several researchers and applications that require understanding people views about a particular topic or event. The study carried out in this paper provides an overview of the algorithms and approaches that have been used for sentiment analysis in twitter. The reviewed articles are categories into four categories based on the approach they use. Furthermore, we discuss directions for future research on how twitter sentiment analysis approaches can utilize theories and technologies from other fields such cognitive science, semantic Web, big data and visualization.

Keywords—Data analysis, sentiment analysis, social media, twitter, machine learning, graph, survey.

1 Introduction

Nowadays, social media platforms like twitter or facebook have gained high importance for many readers as they allow people to share and express their opinions about topics and post messages across the world in a simple way [1]. Twitter is a microblogging and social networking platforms where massive instant messages (i.e. tweets) are posted every day. Tweets have limited number of characters (maximum of 140 characters for each tweet) and use of hashtags between words which facilitates processing and search, which has attracted researchers to analyze twitter data for different applications.

Analyzing twitter feeds for sentiment analysis has become a major research and business activity [2]. Twitter Sentiment Analysis (TSA) tackles the problem of analyzing the tweets in terms of the opinion they express [3]. However, analyzing sentiments is a challenging task due to the vast amount tweets with various topics. This encouraged researches in the field to develop approaches that can automatically detect and mine sentiments opinions within a huge amount of data [3]. There are numerous of articles focused on social media data analysis and, more recently, researchers increased their focus on TSA. Approaches for TSA vary from lexicon-based and ma-

chine learning to graph-based approaches. There has been some surveys focused on summarizing TSA approaches such as [2], [3]. Nevertheless, research on TSA has been growing rapidly in the past few years and these articles were published some time ago.

In this paper, we provide a survey of existing sentiment analysis approaches that have been used in multiple fields such as health, riots, stock sales, air pollution fields, and disaster management, etc. We categorize the approaches into four categories, namely machine learning, lexicon-based, hybrid (combines machine learning and lexicon-based approaches and graph-based approaches. We also review other approaches that cannot be roughly categories in these categories. For each reviewed article we describe the applied algorithm, size of the dataset used in the analysis and the results. Discussion of the TSA approaches and future directions of these approaches is provided. The rest of this paper is organized as follows: Section 2 provides a background with a brief description of TSA and sentiment classification stages. Section 3 presents a literature review of sentiment classification approaches. Section 4 provides discussion and conclusion.

2 Background

Twitter is a microblogging network service that started in 2006 by which users can share text messages, called tweets, and links to other content such as images, websites and articles. Every tweet has a maximum length of 140 characters, describing an event or peoples' opinion about an event or a person, and can have links to news articles, videos or images. A tweet can use hashtags to indicate relevant topics. Recent statistics show that there are over 320M active twitter accounts with 500M tweets sent every day1. Twitter sentiment analysis aims to classify opinion expressed tweets as positive or negative. Thus, it is a classification problem [3]. Next, we describe twitter sentiment classification phases.

2.1 Data ingestion phase

Involves ingesting data streams from twitter API and several resources [4]. Popular open-source options for ingesting data streams into analytics platforms or a data store, include MQTT, RabbitMQ, ActiveMQ, NSQ, ZeroMQ, NiFi, DistributedLog, and Kafka. Among these tools, Kafka is the most popular one [4]. It is a distributed streaming platform that uses a publish and subscribe method for streaming records and storing the streams for processing [5]. Kafka runs as a cluster on servers that span multiple data centers and storing streams of records in categories (i.e. topics). Each record consists of a key, a value and timestamp. Records that are ingested with data ingestion tools are stored in frameworks suitable for further analytics such as Hadoop Distributed File System (HDFS) and Cassandra. Hadoop is the most used framework.

[1] https://www.websitehostingrating.com/twitter-statistics/

It is an open-source software library written in JAVA programming language allows distributed processing of large amount of data and parallel processing of large datasets on cluster of nodes. Hadoop includes four main modules: (i) Hadoop Common which contains utilities used by other modules; (ii) Hadoop Distributed File System (HDFS) which provides storage capabilities by breaking large files into blocks and storing them in different nodes across a cluster; (iii) Hadoop Map-Reduce to process the large dataset in parallel by each map task work on a part of data input (the final output is processed further in the reduce phase); and (iv) Hadoop YARN which is a resource negotiator for scheduling cluster resources [6], [7].

2.2 Data analytics phase

Aims to process data analytics. Several engines can be used for distributed analytics such as: Spark and Flink [4]. The former is the most common used engine. Spark is an open-source clustering-computing framework with implicit data parallelism and fault tolerance features, SQL libraries and stream processing [8]. Spark runs as an independent process. The cluster manager in Spark assigns tasks to workers, one task per partition. Each task applies its unit of work in its dataset partition and the result saved to the disk. Spark framework supports four cluster managers: (i) spark standalone (i.e. simple cluster manager); (ii) apache mesos (i.e. general cluster manager which can run Hadoop application); (iii) apache Hadoop YARN (i.e. to split up functionalities of resource management and job scheduling), and (iv) kubernetes (i.e. for automated deployment, scaling, and management of containerized applications).

Sentiment classification phase: Extracts opinions and trends from data through four processes:

i. *Pre-processing:* Involves transforming raw data into an understandable format for machines through three steps: data cleaning, data transformation and data reduction [9]. In the data cleaning, all URLs, hashtag symbols and other special characters are removed. Stop words are also removed to save space and time (a stop word is a commonly used word such as - the, a, an, in - that a search engine can ignore them when retrieving them as a result) [9].
ii. *Feature extraction:* Aims to extract important features for training. The more extracted features the more accurate classification results [10]. Several features can be extracted, including the following:

- Sentiment features (SENF): related to the positivity and negativity of words and emotions (e.g. number of positive and negative words or emotions in a text).
- Syntax-based features (SYNF): related to question, exclamation, parentheses and quotation marks and their count in sentences (e.g. number of exclamation marks and Number of dots).
- Semantic features (SEMF): focus on the logic behind the sentences such as passive and active forms.
- Unigram-based features (UGF): include hypernyms (i.e. more generals) and hyponyms (i.e. more specific) features as seed words for a user-defined input.

- N-gram features (NGF): take an N number of sequential words as a group to specify a feature. If each two sequential words are used, it is called Bigram features (BGF) [11].
- Top words features (TWF): extract words with high number of occurrences in text.
- Pattern-based features (PTF): rely on Part-of-Speech tags (e.g. Positive and negative names, positive and negative verbs, positive and negative adjectives, pronouns, etc.) to extract patterns in sentiments.

iii. *Feature selection (filtering):* Used to reduce features' size in order to improve classification models speed and accuracy. The term Frequency-Inverse Document Frequency (TF-IDF) is the most widely feature selection method [12]. It calculates the most frequently used terms (TF) and how infrequently the term appears (IDF) in a text. The terms with the high TF-IDF (i.e. product of TF and IDE) scores indicate the most frequently terms with the most relevant information on a specific topic.

iv. *Classification:* The process of classifying text into several defined classes. Examples of classifiers include: Lexicon classifier [13]: a lexicon is a collection of words which has a predefined polarity; Naïve Bayes (NB) [14]: a simple probability classifier that calculates a set of probability by counting the frequency and combinations of values in a given data set; and (iii) Support Vector Machine (SVM) [15]: a tool for data mining tasks such as classification, regression and novelty detection. SVMs have been successfully applied in a number of applications such as particle identification, face identification and text categorization. According to [16], there are three main classification types:

- *Binary classification*: classification of sentiments to two basic polarities, positive (i.e. high positive scores) and negative (i.e. high negative scores).
- *Ternary classification*: classification of sentiments into three classes: positive, negative and neutral (i.e. don't include any positive or negative scores).

Multiclass classification: classification of sentiments into multiple predefined classes to extract not only positivity or negativity, but also to extract feelings and opinions. Furthermore. The classes may be defined to classify texts with happiness, enjoy, hate, etc.

3 Literature Review

In this section, several approaches used in twitter sentiment classification are reviewed. While earlier work has identified four categories of TSA approaches, namely machine learning, lexicon-based, hybrid-based and graph-based. We review recent research on TSA approaches, and we provide review of other approaches that cannot be roughly categorized in the above categories.

3.1 Machine Learning approaches

We reviewed recent TSA approaches which utilize machine learning classifiers. The work in [17] proposed an approach to automatically detect emotions on twitter messages that explores characteristics of the tweets and the writer's emotion using SVM LibLinear model. A total of 520K tweets were collected as raw data. TweetToSparseFeatureVector filter in Weka Affective tweets package was used to extract features. Each tweet in the dataset was annotated with the corresponding emotion based on the weightage computed using the extracted features. The results showed that the accuracy of the SVM classifier was 98%. In [18] the authors conducted a study to find if there is a difference between people's views from eight Western and Eastern countries on ISIS in terms of positive and negative words. 6853 tweets about ISIS from eight different countries. TF-IDF technique was used to conduct text sentiment analysis using R. By analysing the ratio of the negative and positive words from the eight countries, the results showed that the ratio of the positive and negative words ranged from (29% to 33% and 67% to 71%) respectively, out of the total number of the words. Authors in [1] used three machine learning classifiers: naïve Bayes (NB), Logistic Regression (LR) and Support Vector Machine (SVM) for extracting health-related opinions. The NB algorithm used involved Multinomial Naive Bayes and Bernoulli Naive Bayes, and the SVM used involved Support Vector Classification, Linear Support Vector Classification, Stochastic Gradient Descent and Nu-Support Vector Classification. A total of 2026 filtered tweets were used in the experimentation. Results showed that the accuracy ranged between 85% and 91% and the best classifiers were SVM using Linear Support Vector Classification and Stochastic Gradient Descent. The work in [19] used Least Squares Support Vector Regression (LSSVR) models to deal with multivariate regression data. Three types of data: tweets sentiment scores, stock market values, and hybrid data (i.e. contain both sentiment scores of tweets and stock market values), were used to forecast monthly total vehicle sales in USA. Seasonal factors were employed to deseasonalizing monthly total vehicle sales. 6M tweets were gathered using three keywords, namely buy car, buy truck, and buy vehicle. Results indicated that using hybrid data with deseasonalizing procedures by the LSSVR models obtained more accurate forecasting results that other models such as naïve method, the exponential smoothing. [20] proposed an analytical framework to process real-time twitter data. The framework used Kafka and Spark platforms. Kafka was connected to twitter streaming API for data ingestion. Spark was used for data processing. A total of 50K tweets were used in the analysis. Results showed that the proposed platform has the ability to process real-time data to view people's reaction to tragic or dangerous events. [11] used Bayesian Logistic Regression (BLR) method to find the correlation between twitter sentiment and events that have occurred. Data gathering was made up of two steps using twitter's Streaming API. The first was collecting the data to use as a training set to build the model (4162 tweets were collected and manually labelled positive or negative. The second step was collecting 30M tweets during the World Cup 2014 tournament. They compared between BLR and NB and found that BLR had better F-score for positive and negative tweets (74.8, 61.2 for positive tweets and 74.9, 70.1 for negative tweets) respectively.

The work in [4] proposed a scalable framework for multilevel streaming analytics of social media data by leveraging distributed open-source tools and deep learning architectures. It combined Spark streaming for real time text processing, the Long Short Term Memory (LSTM) deep learning model for higher level sentiment analysis, and other tools for SQL-based analytical processing to provide a scalable solution for multilevel streaming text analytics. 1.6M random tweets were labelled as expressing either positive or negative sentiments. Accuracy values of positive and negative tweets were 82.1% and 79.9% respectively. [21] conducted an experimental study by applying a computational methodology using NLP techniques to identify potentially significant fragments. The authors developed a frame-based method and Virtual Research Environment (VRE) by matching retweets to their sources allowing up to 30 individual character differences between the original and its retweet (cut-off). A total of 2.6M tweets were collected from 700K distinct user accounts. The analysis confirmed that the proposed approach can be used to detect crises using tweets. [12] proposed an approach that uses NB and LR for classifying sentiments as positive or negative using Hadoop platform. They authors propose to use TF-IDF as a feature selection method instead of using Part-Of-Speech (POS) labels for classification. The work considered each term in sentiments claiming that POS doesn't give a good accuracy due to the difference of word grammar in different contexts. Results showed that LR had better accuracy value than NB (67.76%, 66.66%) respectively. Table 1 summarizes the machine learning approaches for TSA.

Table 1. Summary of machine learning approaches

Article	Algorithm	#of tweets	Results
[17]	SVM LibLinear model	520K	98% (Accuracy)
[18]	*TF-IDF*. R dictionary	6853	ratio of the positive and negative words ranged from (29% to 33% and 67% to 71%) respectively
[1]	NB, SVM, LR	2026	91.87% (Accuracy)
[19]	LSSVR	6M	LSSVR had better accuracy that other models
[20]	Kafka, Spark platforms	50K	framework is able to accurately offer accurate location and time based processing
[11]	BLR, NB	4162 & 30M	BLR has better F-score than NB for positive and negative tweets (74.8, 61.2 for positive tweets and 74.9, 70.1 for negative tweets) respectively
[4]	LSTM, Spark	1.6M	Accuracy values of positive and negative tweets were 82.1% and 79.9% respectively
[21]	VRE	2.6M	TSA can be used to identify crises situations
[12]	NB, LR	6MB size	Results showed that LR had better accuracy value than NB (67.76% , 66.66%) respectively

3.2 Lexicon-based approaches

Lexicon-based approaches leverage a list of words annotated by polarity or polarity score to determine the opinion score of a given text. Such approaches make use of lexicon dictionary that consists of list of positive, negative and neutral words. The work in [22] developed a framework that classifies movie reviews into positives,

negatives and neutral polarity using the lexicon published in [23] which has 2195 positive words and 4972 negative words. 100 tweets were used for evaluation. All stop words and noises were removed. Separated words were matched with the positive and negative words from [23]. Results showed that the proposed lexicon-based method was able to classify sentiment with 52% accuracy. In [9] the authors proposed a scoring model incorporating language and non-language features to find the sentiment polarity of twitter messages. The language features comprises of the text which describes a subject either in a positive, negative or neutral way. The non-language features consist of the symbols used by the users of twitter like emoticons and shortened words. A total of 750 tweets were collected and manually classified as positive, negative and neutral. Accuracy of polarity to tweets reached (84%). The work in [24] developed methods for a statistical comparison of algorithms which does not rely on human annotation or on known class labels. Sentiment was assigned to the output by the use of three separate lexicons, OpLex, SentiLex, and LIWC. A total of 1144 tweets were analyzed. Results showed that the three lexicons had different statistical-based values. In [25] the authors explored the effectiveness of performing Real-Time Sentiment Analysis approaches on small twitter dataset. Twitter sentiment has been performed by using RNN components in Standford Core NLP, which is a standard natural language software used for extracting various form of sentiments from large set of texts. The authors used 56K tweets for experimentation. The results were promising and showed that the proposed model can be used to forecast movements of individual stock prices. Table 2 summarizes the articles that employed lexicon-based approaches for TSA.

Table 2. Summary of lexicon-based approaches

Article	Algorithm	#of tweets	Results
[22]	LB	*100*	52% (accuracy)
[9]	Sentiment Scoring (SS)	*750*	*84% (Accuracy)*
[24]	*OpLex, SentiLex, LIWC lexcons*	*1144*	*Variation of statistical values*
[25]	*RNN components in Standford Core NLP*	*56000*	*High accurate prediction rates.*

3.3 Hybrid-based approaches

A number of studies have combined two or more approaches for TSA such as combining machine-learning and lexicon-based approaches [3]. The work in [26] proposed a hybrid method by discussing a real-time sentiment analysis using Apache Spark's machine learning library, Hadoop distributed file system and streaming engine for sentiment prediction. The sentiment classification performance of the proposed system for offline and real-time modes were 86.77% and 80.93%, respectively. Authors in [13] proposed a framework for topic classification and sentiment analysis of twitter data. The framework used Apache Flume within Hadoop platform to extract twitter data in real-time environments. For topic classification, a bag of words algorithm was used where each category vector contains related keywords and each tweet is classified to a category based on a count variable. A hybrid algorithm of lexicon

classifier and Naive Bayesian classifier (HL-NBC) was developed for sentiment analysis. The HL-NBC classifier was compared to naive and lexicon classifiers showing the best accuracy. Overall of 1M tweets were used in the analysis. Results showed that the model had 82% accuracy. The work in [27] developed a hybrid approach which consists of the hierarchical combination of SVM and RF. A total of 10500 tweets were used in the analysis. A portion of 3,000 and 10,500 of the stemmed data with equal distribution from each class has been identified as the first dataset and second dataset to be used in the testing phase. The developed hybrid approach achieved an accuracy of up to 86.4% and 82.8% on the first and second datasets, respectively. The work in [28] proposed a hybrid approach that involves machine learning and lexicon-based approaches that pre-process and re-label tweets using weight-based classification. The proposed approach was tested using 40,000 tweets. Experimental results showed that that pre-processing and drift detection techniques significantly improve the classification accuracy (over 70%). Table 3 summarizes the hybrid-based approaches for TSA.

Table 3. Summary of hybrid-based approaches

Article	Algorithm	#of tweets	Results
[26]	*NB model from Sparks MLlib*	*2K*	*Performances of the system for offline and real-time modes are 86.77% and 80.93%, respectively*
[13]	*Lexicon ,NB, Hybrid Lexicon-NB (Best)*	*1 M*	*82% (Accuracy)*
[27]	*Hybrid of support vector machines and random forest algorithms*	*10500*	*Hybrid approach had better accuracy (86.4%)*
[28]	*ML and Lexicon*	*40K*	Pre-processing and drift detection techniques significantly improve the classification accuracy (over 70%)

3.4 Graph-based approaches

A graph can be represent as a set of vertices (i.e. nodes) interconnected via directed linked (i.e. edges). In [29] the authors used Social Network Analysis (SNA) to study the community of twitter users disseminating information during the crisis caused by the Australian floods in the period 2010-2011 to reveal interesting patterns and features. Using SNA, users were represented as nodes and responses between users about a particular tweet were represented as edges. Ego analysis was applied by analyzing centrality of nodes which is measured by the degree of the various nodes in the graph with degree representing the number of other nodes to which a node is adjacent. A total of 7520 tweets were used in the analysis. The finding showed that SNA can be used to identify influential members of the online communities. The work in [30] used graph-based optimization to enhance the performance of SVM classifier by taking the related tweets into consideration. A graph was constructed using three types of relations between tweets were used (retweets, tweets containing the target and published by the same person, and tweets replying to or replied by the classified tweet). Based on these three tweet relations, a graph was constructed using the input tweet collection

of a given target. Total of 1993 tweets were used in the analysis (459 positive, 268 negative and 1,212 neutral). Accuracy of the proposed graph-based optimisation approach was 68.3%. Authors in [31] proposed a hashtag graph model which incorporate the co-occurrence information of hashtags. Edges of the graph are the links between hashtags, and each edge represents an undirected link between two hashtags, which co-occur in at least one tweet. After creating the hashtag graph, SVM algorithm was used to determine the sentiment polarity of tweets. Experimental results on a real-life data set consisting of 29195 tweets and 2181 hashtags showed the effectiveness of the proposed hashtag graph model. Table 4 summarizes graph-based approaches for TSA.

Table 4. Summary of graph-based approaches

Article	Algorithm	#of tweets	Results
[29]	*SNA and EGO Analysis*	*7520*	*SNA can be used to identify influential members*
[30]	graph-based optimization model	*1993*	Accuracy of the proposed graph-based optimisation approach was 68.3%
[31]	*SVM, LBP, RL,ICA*	*29195*	*LBP had the best accuracy value of 77.72%*

3.5 Other approaches

In the TSA literature, there are some approaches that cannot be roughly categorized in the above categories. For instance, utilizing interactive visualization tool to TSA was proposed by [32]. In this work, the authors proposed Plexus, a system that identifies and visualizes people's emotions on any two related topics by streaming and processing data from twitter. The effectiveness of Plexus was evaluated and demonstrated by a feasibility study with 14 participants. The results showed that proposed approach was significant to understand people's reactions to certain topics. Another category of approaches applied cognitive science theories for TSA. For example, the work in [33] applied a social cognitive theory, a learning theory stating that people learn by observing and imitating others and by positive reinforcement, to analyze social network analysis conversations among people. In [34], the authors used cognitive science to build a comprehensive cognition-driven opinion-mining engine. In this work, the authors proposed SenticNet, a publicly available semantic and affective resource for concept-level sentiment analysis. It builds on the energy-based COGBASE common sense knowledge formalism to provide semantics for 30,000 multi word expressions, enabling a deeper and more multi-faceted analysis of natural language opinions. Furthermore, the work in [35] proposed a cognition based attention (CBA) model for sentiment analysis. The proposed model learns from cognition grounded eye-tracking data. The authors build a regression model to map syntax, and context features of a word to its reading time based on eye-tracking data. Then, they apply the model to sentiment analysis text to obtain the estimated reading time of words at the sentence level. Evaluation on benchmarking datasets validates the effectiveness of the proposed method. Some researchers have focused on analyzing twitter data using SNA metrics and similarity-based algorithms. For instance, the work in [36] analyzed event mentions in microblogs of social media, like twitter, for quantify-

ing user's interests using similarity-based region network. Regional user interests are obtained for each topic by applying latent Dirichlet allocation to region-specific collections of tweets, and then compute pairwise similarities among regions. Social similarity based on user socially important locations was also quantified using Levenshtein distance and evaluated on a real-life twitter dataset [37]. Another group of studies on TSA utilized big data platforms such as Hadoop to analyze large number of tweets [38]. The authors in [39] proposed the use of cloud environment for big data analytics by utilizing Hadoop platform to perform TSA. The work in [40] analyzed sentiment data by using Hadoop platform, in particular Hadoop Distributed File System (HDFS) and MapReduce modules. In [41] the authors proposed a method to perform real time sentimental analysis on the tweets that are extracted from the twitter and provide time based analytics to the user. They used NLP, machine learning and unigram Naïve techniques for classification, and extracted tweets are loaded into Hadoop platform. The work in [42] developed a large scale architecture by combining Storm and Hadoop to process social media data and facilitate their integration into the traditional data warehouse. Recent researches started working on semantic driven approaches for TSA. For instance, the work in [43] proposed an approach to determine domain-based social influencers by means of a framework that incorporates semantic analysis and machine learning modules to measure and predict users' credibility from twitter data. The authors in [44] introduced an approach of adding semantics as additional features into the training set for sentiment analysis. For each extracted entity from tweets, a semantic concept was added as an additional feature to that concept. In [45] the authors proposed a cloud based system for real time targeted advertising based on TSA and Apache Spark for implementation. The work used twitter streaming data as data source. Results showed the usability of the proposed system.

4 Discussion and Conclusion

From the literature, we noticed that TSA is an open field for research and there are emerging TSA approaches that combine machine learning and lexicon-based approaches with theories and technologies from cognitive science, semantic Web and big data.

Approaches that use big data platforms: The term Big Data is globally used for collection of datasets that are huge and complex, which makes it difficult to process by adopting traditional way of data processing techniques. The challenges related to big data provide a chance to understand the data patterns and helps in prediction of events and results. Hence, there is a growing demand for tools which can process and analyze big data [7]. In this regards, twitter produces humungous amount of data in a daily basis. This abundant data is mainly unstructured or structured and is termed as big data. Accordingly, this requires advanced technologies that can handle such large amounts of data efficiently. This high volume of data leads to some challenges like processing of large data sets, extraction of useful information from online generated data sets etc. For this, many emerging TSA approaches are using Hadoop platform to

provide best solution to analyze and process large data sets. In particular, they use Hadoop platform such as HDFS and MapReduce to analyze output data from machine learning and lexicon-based approaches. We believe that this approach is required for TSA, especially that we are in the big data era and there are many real-time applications which require fast sophisticated analysis of large amount of data.

Approaches that use semantic Web technologies: The Semantic Web vision started in 2001 in order to evolve the conventional Web from a global Web graph of Webpages linked via Hyperlinks into a global data space where both documents and data entities are semantically linked in a structured way. Recent TSA started using semantic technologies to generate ontologies representing concepts of a domain. These concepts can be used as features or instances to enrich datasets. In this regards, semantic relations can be extracted from user generated content in social media platforms such as twitter to generate an ontology representing a domain [46]. Furthermore, semantic Web applications such as semantic browsers, which exploits ontologies, can be used to visualize links and associations between different tweets. We believe that more work in the coming years will focus of utilizing semantics and knowledge graphs to facilitate analysis of twitter data.

Approaches that are based on cognitive science theories: Another group of approaches focused on applying cognitive science theories for TSA. Cognitive science studies the human mind and its processes. For instance, cognitive learning theories argues that the human mind is structured to different level of abstraction and there is a level called the basic level object where most familiar concepts exists [47]. This theory has been applied over knowledge graphs to develop algorithms to identify familiar concepts in a domain [48]. These algorithms can be applied over social networks (e.g. twitter) to identify familiar tweets of persons.

Approaches that use SNA metrics: Social Network Analysis (SNA) enables analysis of the social network such as twitter based on some metrics. For instance, node centrality (e.g. a node could represent a particular twitter user) can be used to measure a node's importance in the network. There are several centrality-based algorithms in SNA that can be used to measure importance of a node such as degree centrality (i.e. nodes with higher connections are more important), closeness centrality (i.e. nodes which are reachable at shorter path lengths are more important) and betweenness centrality (i.e. a node's importance is based on the number of the shortest paths between pairs of other nodes that go through that node)[49]. Similarly, we noticed that some works on TSA have used similarity metrics to identify similar topics, news headlines and user personality. Hence, and since we are dealing with social networks we believe that there is good potential to embed SNA and similarity metrics with current TSA approaches, especially machine learning approaches.

Visualization-driven approaches. Visualization provides an important tool for exploration that leverages the human perception and analytical abilities to offer exploration trajectories for the users. Visualization-based applications use visual or graphic structures, such as images, maps or graphs (individually and in combinations) to represent associations between tweets and or users. Researches can apply famous visualization theories such as using Shneiderman's [50] (overview first, zoom and filter, then

details-on-demand) seminal visual information-seeking mantra as a guiding principle in evaluating the usability and utility of visualization-driven approaches.

In this survey we have presented an overview of TSA approaches. Over 40 articles of recent research on TSA were briefly reviewed and categorized. From the discussion we concluded that TSA will be an active research area in the coming years. We also have discussed future directions and enhancements for TSA approaches.

5 References

[1] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "Arabic language sentiment analysis on health services," pp. 114–118, 2017. https://doi.org/10.1109/asar.2017.8067771

[2] B. Batrinca and P. C. Treleaven, "Social media analytics: a survey of techniques, tools and platforms," AI Soc., vol. 30, no. 1, pp. 89–116, 2015. https://doi.org/10.1007/s00146-014-0549-4

[3] A. Giachanou and F. Crestani, "Like It or Not: A Survey of Twitter Sentiment Analysis Methods," ACM Comput. Surv., vol. 49, no. 2, Jun. 2016. https://doi.org/10.1145/2938640

[4] S. Ge, H. Isah, F. Zulkernine, and S. Khan, "A scalable framework for multilevel streaming data analytics using deep learning," Proc. - Int. Comput. Softw. Appl. Conf., vol. 2, pp. 189–194, 2019. https://doi.org/10.1109/compsac.2019.10205

[5] G. Wang et al., "Building a replicated logging system with apache kafka," Proc. VLDB Endow., vol. 8, no. 12, pp. 1654–1655, 2015. https://doi.org/10.14778/2824032.2824063

[6] "Learn Hadoop - Big Data Analysis Framework." [Online]. Available: https://www.tutorialspoint.com/about/index.htm.

[7] A. S. Hashmi and T. Ahmad, "Big Data Mining: Tools & Algorithms," Int. J. Recent Contrib. from Eng. Sci. IT, vol. 4, no. 1, pp. 36–40, 2016.

[8] "Spark 101: What Is It, What It Does, and Why It Matters | MapR.".

[9] J. Akilandeswari and G. Jothi, "Sentiment classification of tweets with non-language features," Procedia Comput. Sci., vol. 143, pp. 426–433, 2018.

[10] M. Bouazizi and T. Ohtsuki, "Sentiment analysis in twitter: From classification to quantification of sentiments within tweets," 2016 IEEE Glob. Commun. Conf. GLOBECOM 2016 - Proc., 2016. https://doi.org/10.1109/glocom.2016.7842262

[11] P. Barnaghi, P. Ghaffari, and J. G. Breslin, "Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment," Proc. - 2016 IEEE 2nd Int. Conf. Big Data Comput. Serv. Appl. BigDataService 2016, pp. 52–57, 2016. https://doi.org/10.1109/bigdataservice.2016.36

[12] A. Prabhat and V. Khullar, "Sentiment classification on big data using Naïve bayes and logistic regression," 2017 Int. Conf. Comput. Commun. Informatics, ICCCI 2017, 2017. https://doi.org/10.1109/iccci.2017.8117734

[13] A. P. Rodrigues and N. N. Chiplunkar, "A new big data approach for topic classification and sentiment analysis of Twitter data," Evol. Intell., no. 0123456789, 2019. https://doi.org/10.1007/s12065-019-00236-3

[14] F. Paquin, J. Rivnay, A. Salleo, N. Stingelin, and C. Silva, "Multi-phase semicrystalline microstructures drive exciton dissociation in neat plastic semiconductors," J. Mater. Chem. C, vol. 3, pp. 10715–10722, 2015. https://doi.org/10.1039/c5tc02043c

[15] K. P. Bennett and C. Campbell, "Support Vector Machines: Hype or Hallelujah?," SIGKDD Explor. Newsl., vol. 2, no. 2, pp. 1–13, Dec. 2000. https://doi.org/10.1145/380995.380999

[16] M. Byrkjeland, F. Gørvell de Lichtenberg, and B. Gambäck, "Ternary Twitter Sentiment Classification with Distant Supervision and Sentiment-Specific Word Embeddings," pp. 97–106, 2019. https://doi.org/10.18653/v1/w18-6215
[17] J. Ranganathan and A. Tzacheva, "Emotion mining in social media data," Procedia Comput. Sci., vol. 159, pp. 58–66, 2019. https://doi.org/10.1016/j.procs.2019.09.160
[18] S. Mansour, "Social media analysis of user's responses to terrorism using sentiment analysis and text mining," Procedia Comput. Sci., vol. 140, pp. 95–103, 2018. https://doi.org/10.1016/j.procs.2018.10.297
[19] P. F. Pai and C. H. Liu, "Predicting vehicle sales by sentiment analysis of twitter data and stock market values," IEEE Access, vol. 6, no. c, pp. 57655–57662, 2018. https://doi.org/10.1109/access.2018.2873730
[20] B. Yadranjiaghdam, S. Yasrobi, and N. Tabrizi, "Developing a Real-Time Data Analytics Framework for Twitter Streaming Data," Proc. - 2017 IEEE 6th Int. Congr. Big Data, BigData Congr. 2017, pp. 329–336, 2017. https://doi.org/10.1109/bigdatacongress.2017.49
[21] R. Procter, F. Vis, and A. Voss, "Reading the riots on Twitter: Methodological innovation for the analysis of big data," Int. J. Soc. Res. Methodol., vol. 16, no. 3, pp. 197–214, 2013.
[22] A. Azizan, N. N. S. A. Jamal, M. N. Abdullah, M. Mohamad, and N. Khairudin, "Lexicon-Based Sentiment Analysis for Movie Review Tweets," in 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS), 2019, pp. 132–136. https://doi.org/10.1109/aidas47888.2019.8970722
[23] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 168–177. https://doi.org/10.1145/1014052.1014073
[24] M. T. Machado, E. Ruiz, and K. J. Abraham, "A New Statistical Approach for Comparing Algorithms for Lexicon Based Sentiment Analysis," CoRR, vol. abs/1906.0, 2019.
[25] S. Das, R. K. Behera, M. Kumar, and S. K. Rath, "Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction," Procedia Comput. Sci., vol. 132, no. Iccids, pp. 956–964, 2018. https://doi.org/10.1016/j.procs.2018.05.111
[26] D. Kılınç, "A spark-based big data analysis framework for real-time sentiment prediction on streaming data," Softw. - Pract. Exp., vol. 49, no. 9, pp. 1352–1364, 2019. https://doi.org/10.1002/spe.2724
[27] H. A. Shehu and S. Tokat, "A Hybrid Approach for the Sentiment Analysis of Turkish Twitter Data," in Artificial Intelligence and Applied Mathematics in Engineering Problems, 2020, pp. 182–190. https://doi.org/10.1007/978-3-030-36178-5_15
[28] L. A. Deshpande and M. R. Narasingarao, "Addressing social popularity in twitter data using drift detection technique," J. Eng. Sci. Technol., vol. 14, no. 2, pp. 922–934, 2019.
[29] F. Cheong and C. Cheong, "Social media data mining: A social network analysis of tweets during the Australian 2010-2011 floods," PACIS 2011 - 15th Pacific Asia Conf. Inf. Syst. Qual. Res. Pacific, 2011.
[30] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter sentiment classification," ACL-HLT 2011 - Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol., vol. 1, no. January, pp. 151–160, 2011.
[31] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic Sentiment Analysis in Twitter: A Graph-Based Hashtag Sentiment Classification Approach," in Proceedings of the 20th ACM International Conference on Information and Knowledge Management, 2011, pp. 1031–1040. https://doi.org/10.1145/2063576.2063726
[32] X. Wu, L. Bartram, and C. Shaw, "Plexus: An Interactive Visualization Tool for Analyzing Public Emotions from Twitter Data," CoRR, vol. abs/1701.0, 2017.

[33] H. J. Yoon and G. Tourassi, "Analysis of online social networks to understand information sharing behaviors through social cognitiv. Theory," Proc. 2014 Biomed. Sci. Eng. Conf. - 5th Annu. ORNL Biomed. Sci. Eng. Conf. Collab. Biomed. Innov. - Multi-Scale Brain Spanning Mol. Cell. Syst. Cogn. Behavi, 2014. https://doi.org/10.1109/bsec.2014.6867744

[34] E. Cambria, D. Olsher, and D. Rajagopal, "SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis," Proc. Natl. Conf. Artif. Intell., vol. 2, pp. 1515–1521, 2014.

[35] Y. Long, Q. Lu, R. Xiang, M. Li, and C. R. Huang, "A cognition based attention model for sentiment analysis," EMNLP 2017 - Conf. Empir. Methods Nat. Lang. Process. Proc., pp. 462–471, 2017.

[36] G. Jang and S.-H. Myaeng, "Predicting event mentions based on a semantic analysis of microblogs for inter-region relationships," J. Inf. Sci., vol. 44, no. 6, pp. 818–829, Mar. 2018. https://doi.org/10.1177/0165551518761012

[37] M. Celik and A. S. Dokuz, "Discovering socially similar users in social media datasets based on their socially important locations," Inf. Process. Manag., vol. 54, no. 6, pp. 1154–1168, 2018. https://doi.org/10.1016/j.ipm.2018.08.004

[38] A. Kotwal, P. Fulari, D. Jadhav, and R. Kad, "Improvement in Sentiment Analysis of Twitter Data Using Hadoop," Imp. J. Interdiscip. Res., vol. 2, no. 7, pp. 2454–1362, 2016.

[39] M. Kumar and A. Bala, "Analyzing Twitter sentiments through big data," Proc. 10th INDIACom; 2016 3rd Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2016, pp. 2628–2631, 2016.

[40] M. T. Cs, "Sentiment Analysis of Big Data Applications using Twitter Data with the Help ofHADOOP Framework," vol. V, pp. 251–255, 2016.

[41] M. Trupthi, S. Pabboju, and G. Narasimha, "Sentiment analysis on twitter using streaming API," Proc. - 7th IEEE Int. Adv. Comput. Conf. IACC 2017, pp. 915–919, 2017. https://doi.org/10.1109/iacc.2017.0186

[42] F. Jenhani, M. S. Gouider, and L. Ben Said, "Streaming social media data analysis for events extraction and warehousing using hadoop and storm: Drug abuse case study," Procedia Comput. Sci., vol. 159, pp. 1459–1467, 2019. https://doi.org/10.1016/j.procs.2019.09.316

[43] B. Abu-Salih et al., "Time-aware domain-based social influence prediction," J. Big Data, vol. 7, no. 1, p. 10, 2020.

[44] H. Saif, Y. He, and H. Alani, "Semantic Sentiment Analysis of Twitter," pp. 508–524, 2012.

[45] L. R. Nair, S. D. Shetty, and S. D. Shetty, "Streaming big data analysis for real-time sentiment based targeted advertising," Int. J. Electr. Comput. Eng., vol. 7, no. 1, pp. 402–407, 2017. https://doi.org/10.11591/ijece.v7i1.pp402-407

[46] A. Al-Abri, Z. Al-Khanjari, Y. Jamoussi, and N. Kraiem, "Mining the students' chat conversations in a personalized e-learning environment," Int. J. Emerg. Technol. Learn., vol. 14, no. 23, pp. 98–124, 2019. https://doi.org/10.3991/ijet.v14i23.11031

[47] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P.Boyes-Braem, "Basic Objects in Neutral categories," Cogn. Psychol., vol. 8, pp. 382–439, 1976. https://doi.org/10.1016/0010-0285(76)90013-x

[48] M. Al-Tawil, V. Dimitrova, and D. Thakker, "Using knowledge anchors to facilitate user exploration of data graphs," Semant. Web, vol. 11, pp. 205–234, 2020. https://doi.org/10.3233/sw-190347

[49] A. Landherr, B. Friedl, and J. Heidemann, "A Critical Review of Centrality Measures in Social Networks," Bus. {&} Inf. Syst. Eng., vol. 2, no. 6, pp. 371–385, Dec. 2010. https://doi.org/10.1007/s12599-010-0127-3

[50] B. Shneiderman, "The eyes have it: a task by data type taxonomy for information visualizations," in Proceedings 1996 IEEE Symposium on Visual Languages, 1996, pp. 336--343. https://doi.org/10.1109/vl.1996.545307

6 Authors

Omar Y. Adwan is currently an Associate Professor with the University of Jordan, King Abdullah II School for Information Technology, Computer Information Systems Department. Dr. Omar holds a B.Sc in Computer Science (Eastern Michigan University, 1987) and a M.Sc. in Computer Science (The George Washington University, 1998), and a Ph.D in Computer Science (The George Washington University, 2008). He served as a chairman to CIS Dept. of KASIT during 2012-2016. His current areas of interest include Software Engineering, System Engineering Tools, and Databases.

Marwan Al-Tawil is currently an Assistant Professor with the University of Jordan, King Abdullah II School for Information Technology, Computer Information Systems Department. Dr. Marwan holds a B.Sc in Computer Information Systems (Al-Hussein Bin Talal University, 2006) and a M.Sc. in Information Systems (The University of Jordan, 2011), and a Ph.D in Computer Science (The University of Leeds, 2018). He is currently the Dean Assistant for Automated exams at the University of Jordan. His current areas of interest include: Knowledge Graphs, Data Exploration, and Databases. (e-mail: m.altawil@ju.edu.jo).

Ammar M. Huneiti is currently a Professor with the University of Jordan, King Abdullah II School for Information Technology, Computer Information Systems Department. Prof. Ammar holds a B.Sc in Computer Science (University of Wales College of Cardiff, 1991) and a M.Sc. in Information Systems Technologies (The University of Wales College of Cardiff, 1992), and a Ph.D in Intelligent Information Systems (Cardiff University, 2004). He served as Vice Dean of KASIT during 2015-2016. His current areas of interest include: Intelligent Information Systems, Data Mining, Performance Support Systems, Multimedia, Geographic Information Systems, Spatial Databases, Adaptive Hypermedia.

Razan H. Al-Dibsi, Rawan A. Shahin and Abeer A. Abu Zayed are M.Sc. students at The University of Jordan, King Abdullah II School for Information Technology, Computer Science.
(razancl@gmail.com, rshahin81@gmail.com, abeerabuzayd@gmail.com).

Article submitted 2020-04-06. Resubmitted 2020-05-23. Final acceptance 2020-05-25. Final version published as submitted by the authors.