ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 12, December 2018

Twitter Sentiment Analysis: Techniques and Applications

[1] Veerappa B. Pagi, [2] Ramesh S. Wadawadagi, [3] Soumya Patil, [4] Neha Kulkarni
[1][2][3][4] Basaveshwar Engineering College, Bagalkot, Karnataka, India

***Abstract*—With the advent of web 2.0 and related technologies, the volume of user-generated social media content (UGSMC) is rapidly growing and likely to increase even more in the near future. Social networking apps such as Twitter, Facebook and Google+ are gaining more popularity as they allow people to share and express their views about happenings, have discussion with different communities, or post messages across the world. Twitter sentiment analysis (TSA) extends any organization's ability to capture and study public sentiment towards the social events and commodities related to them in real time. This paper provides a comprehensive survey on techniques and applications of TSA available in the literature. The survey focuses on issues such as pre-processing techniques, feature selection methods, learning models, and performance of each method as a criterion. The survey reveals some of the traditional machine learning (ML) algorithms have been efficiently used to work on Twitter data. In conclusion, the paper cites many promising issues for further research in this domain.**

***Keywords* – Twitter Sentiment analysis, Twitter data, Machine learning, Natural language processing**

I. OVER VIEW

Social media applications such as Facebook, WhatsApp, and Twitter are becoming increasingly important and offer valuable user-generated content (UGC) by publishing and sharing information. In particular, Twitter alone generates a large volume of sentiment rich data in the form of tweets, status updates, blog posts, comments, reviews, etc. Identifying interesting and potentially useful contents from large collection of tweets is a crucial issue in social media because many users struggle with information overload [1]. Sentiment analysis (SA) is the process of automatically detecting whether a text segment contains emotional or opinionated content, and it can furthermore determine the text's polarity [2]. Twitter sentiment classification aims to classify the sentiment polarity of tweets as positive, negative or neutral. However, tweets are generally composed of incomplete, noisy and unstructured sentences, irregular expressions, ungrammatical words and non-dictionary terms. In addition, it is hard to identify correlations between tweets due to the diversity of issues, and makes the classification tasks more challenging [3]. Ultimately, a real-time classification system needs to be constructed in order to process large volume of tweets in a very short time. Knowing the public emotions is useful in many contexts including marketing, politics, online shopping, and many more [4]. TSA is usually conducted at different levels varying from coarse-grained level to fine-grained level. Coarse-grained TSA deals with determining the sentiment of an entire document and fine-grained deals with attribute level sentiment analysis. However, utilizing the right technology, tools and applying it to key business drivers TSA is a powerful tool for steering companies and their individual business units to successful outcomes. The general framework of TSA is depicted in Figure 1

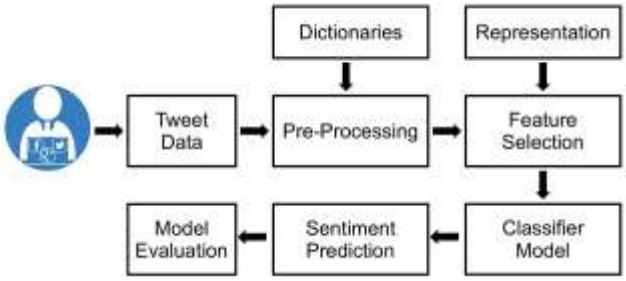

Fig 1: Twitter sentiment analysis framework

It is apparent from the current state of affairs that the framework of TSA is slowly emerging from a disparate set of tools and technologies to a unified model. Essentially, the framework consists of sub-tasks pipeline, in which the first task is primarily composed of various sources and data formats of tweets content. During this phase, several criteria (such as sources and number of tweets to download) on which the data to be collected can be identified and delineated. After identifying various sources, tweet data can either be stored directly into memory for rapid evaluation of unstructured data (in-memory processing) or can be archived to disk (in-database processing) as messages, files, or any machine-generated content. The pre-processing and feature selection phase is a data processing pipeline that covers steps from the knowledge discovery in textual

databases (KDTD) [5]. To facilitate the process of identifying document relevancy, the documents need to be transformed from the full text version to a document vector which describes the content of the documents. Many irregular and implicitly structured representations are transformed into an explicitly structured representation. Basically, there are two types of representation techniques: feature based representation and relational representation. Perhaps, the most commonly used document representation is vector space model (VSM) [6]. Based on the idea of VSM, some other representation techniques have been proposed such as N-Gram, key-phrase and hypernym representations. A central problem in statistical text analysis is the high dimensionality of feature space. Feature selection attempts to remove non-informative words from documents in order to improve the effectiveness of mining task and reduce computational complexity. In model building phase, a generic model is being constructed to describe discoverable patterns. As TSA is highly application-driven domain, it incorporates many techniques from other domains such as information retrieval, machine learning, pattern recognition, natural language processing, ontology, text mining, web intelligence, high-performance computing, visualization, and many application domains. The interdisciplinary nature of TSA research and development contributes significantly to its success and extensive applications. Finally, with this unified architecture, the business analysts or researchers can rely on richer and high quality data patterns discovered. The data and analysis flow would be seamless as they navigate through various data and information sets, test hypothesis, analyze patterns, and make informed decisions. The research conducts a survey on various contemporary TSA techniques and compares them in three dimensions: pre-processing techniques, learning models used, and applications of TSA as the criteria. The remainder of the paper is organized as follows: Section 2 presents an overview of TSA system and related concepts. Section 3 covers the survey methodology and comparative analysis of various research articles studied in the literature. Finally, in Section 4 the article concludes quoting some open research issues in the area of TSA.

II. COMPARATIVE ANALYSIS OF TSA SYSTEMS

Ideally, all the research articles cited in the literature survey are chosen to necessarily satisfy the following two criteria. First, the research work presented ought to exhibit a specific contribution to the area of TSA by increased ability to construct a stable TSA system. This criterion defines a necessary condition for selecting papers for comparative study. Second, the models proposed in the articles are consequently used for automatic analysis of tweets generated through Tweeter. The study also highlights some advantages and disadvantages of the models reported in the literature. However, it is very hard to make a practical comparison of the models cited. This is mainly due to the fact that the research works carried out are problem-specific and are typically based on domain related datasets, which are usually not available openly.

2.1 Pre-processing techniques

Data pre-processing is an essential step of tweet analysis system that identifies the basic units of coding which further interacts with some additional components such as morphological analyzers, word sense disambiguation (WSD) [7], and part-of-speech taggers. Furthermore, tweets are generally composed of incomplete, noisy and unstructured sentences, irregular expressions, ungrammatical words and non-dictionary terms. Before feature selection, a series of pre-processing (e.g., removing stop words, removing URLs, replacing negations) are applied to reduce the amount of noise in the tweets.

Many scholarly approaches have been reported in the literature. The study reveals that the most widely accepted techniques for Twitter data pre-processing includes lexicon-based and linguistic pre-processing. Lexicon-based techniques exploit the features of dictionaries to model the data suitable for learning algorithms. The commonly used strategy for sentence level cleaning is removal of stopwords [8]. Stopwords consists of illatives, pronouns and others words that do not contain much semantic value individually. The basic idea of removing stopwords is to eliminate words that convey little or no content information, like articles, conjunctions, prepositions, etc. Similarly, morphological and inflexional endings from words in English can be removed using a simple pre-processing technique called stemming. Stemming reconstruct the words to its base form, e.g., remove the plural ‘s’ from nouns, the ‘ing’ from verbs, or other affixes. A stem is a natural group of words with identical meaning. Each word is represented by its base word after stemming process. In [9], Porter originally proposed a set of production rules that iteratively transforms English words into their stems called Porter stemmer. A process called lemmatization [10] is used to extract term scores from SentiWordNet [11] to convert each token to its dictionary equivalent. Further, lemmatization maps each verb form to several tenses and nouns to a singular form. Key-phrase extraction [12] algorithms play a critical role in automatic extraction of topical words and phrases from opinion content [13]. Key-phrases provide a concise description of a document’s content and are useful in many TSA tasks. Techniques based on linguistic pre-processing may be used to characterize the data, which increases the learning capabilities of algorithms. Following approaches are frequently applied to web-based content. WSD can be used in TSA systems to address the problem of selecting the

most appropriate sense for a word with respect to the given context. The technique follows the semantic model of the text content for selecting the most appropriate meaning for each word. Yet another pre-processing technique called, named-entity recognition (NER) [14] can be employed as a pre-processing technique to locate and classify the information units from given text content into predefined categories. It labels sequences of words in a sentence that are the names of things, such as person and company names, or city and country. For free text, NLP techniques such as part-of-speech tagging (POST) and word-net semantic classes are used as augmented features [15]. Given each word in a sentence, POST determines the part-of-speech tag, e.g., noun, verb, adjective, etc. In [16], a method known as negation phrases identification (NPI) is used to predict negated parts from communication data over social media. A method called divergence analysis (DA) is utilized to extract relevant information from links and randomly chosen blog posts [17]. Nevertheless, the above list is not exhaustive and also includes many other techniques such as bi-term extraction, URL-pretreatment, removal of replicated characters from fancy words, replacement of emoticons, substitute abbreviations by their full names and so forth.

2.2 Feature selection techniques

Given a set of candidate features the task of feature selection is to choose a subset that performs best under some learning model. In feature selection, it is assumed to be enough relevant features in the original feature set to discriminate clearly between categories, and that some irrelevant features can be eliminated to improve categorization efficiency and reduce computational complexity. In this section, we discuss some prominent feature selection and extraction techniques that are effectively used in TSA systems. A simple technique based on statistics known as information gain (IG) is used to measure the number of bits of information obtained for a given category prediction by knowing the presence or absence of a word in a document [18]. It measures how much information the presence or absence of a term contributes to make the correct classification decision for a given class. Another popular feature selection method 'chi-square' statistics is effectively used to measure the lack of independence between each term and a class. If there exist a significant dependency, then the occurrence of the term makes the occurrence of the class more likely, so it should be helpful as a feature. In [8], researchers used intrinsic annotation refinement to filter out the irrelevant noisy words from vlogs. In [19], Liu et al. proposed a coherence-HMM model to extract coherence features and rank text content from blogs. Bi-normal separation (BNS) is another feature selection metric found to be excellent in ranking words for feature selection filtering [20]. Feature extraction or re-parameterization is the process that generates new features based on transformation or combinations of the original feature set. Numerous feature extraction techniques that are commonly used in various TSA systems are discussed. LSI is one such approach based on the fact that there exist some underlying or hidden structure in the pattern of word usage across documents, and that statistical techniques can be used to estimate the structure [21]. A technique related to eigen-decomposition and factor analysis known as SVD provides the bases for LSI. Yet another popularly used feature extraction technique called PCA produces a new features set from the original feature set, where each feature generated is an orthogonal linear combination of the original features [22]. Unlike PCA, independent component analysis (ICA) aims to determine those features that minimize MI between the new features set [23]. Another supervised feature extraction method called latent Dirichlet allocation (LDA) searches for those vectors in the underlying space that best discriminates among classes.

2.3 Learning models

In this section, various TSA systems are compared and evaluated from the perspective of underlying techniques used. Following the research trend, it is possible to classify TSA systems into three major categories: lexicon-based approaches, ML techniques and hybrid models.

Lexicon-based systems: Lexicon-based systems offer a massive lexicon, thesaurus and semantic linkage between most of the English words. These systems are often used to determine the semantic linkage that interrelates the terms with predefined relationships present in the document or in the sentence [24]. Much of the lexicon-based systems focused on using adjectives as indicators of the semantic orientation of text. Numerous works are reported in the literature for lexicon-based systems applied for varieties of TSA applications. Many opinion terms are used in TSA tasks, e.g., to express favorable states, positive opinion words are used, while negative opinion terms are used to express unfavorable states. To support this opinion lexicons are employed that consist of opinion terms, phrases and idioms.

Machine learning: Supervised learning acquired a great degree of success in solving TSA problems. Supervised learning is also known as classification or inductive learning in ML context. This type of learning is analogous to human learning from previous experiences to gain new knowledge, in order to improve our ability to perform real-world tasks. Numerous techniques have been proposed in the literature for supervised learning. For instance, rule-based systems define a set of rules that use some assertions to make decisions on which rules need to be executed upon those assertions. The probabilistic topic models or simply topic models are suite of algorithms whose aim is to discover the

hidden semantic structure in large archives of documents. Given an observation of an input data, a probabilistic classifier predicts the probability distribution over the classes instead of predicting the most likely class. The naive Bayes (NB) classifier is the simplest probabilistic classifier based on applying Baye's theorem with strong (naive) independence assumptions between the features [25]. This model computes the posterior probability of a class, based on the distribution of the terms in the document. Furthermore, a decision tree classifier decomposes the data hierarchically based on the condition of attribute values. Decomposition of data space is recursively performed until minimum number of records is left in the leaf nodes that are used for the purpose of classification. Yet another kind of classifier called linear classifier performs classification based on the values of a linear combination of the feature values typically presented in a vector form called a feature vector. Example, given a set of labelled training data, support vector machines (SVM) produces an optimal hyperplane which categorizes new samples, making it a non-probabilistic binary linear classifier. A hyperplane is a line selected to best separate the points in the input variable space by their class, either class 0 or class 1. Artificial neural networks (ANN) on the other hand, composed of artificial neurons, which imitate biological neurons of human brain. The neurons are connected by links and they interact with each other. In contrast with supervised learning, unsupervised learning does not have class attributes for data. However, the analyst wants to explore the data to find some intrinsic structures within them. Hence, unsupervised learning methods are beneficial to deal with the problem of identifying hidden structure in unlabeled data.

Hybrid techniques: Hybrid approaches are the class of learning models that combines a supervised learning algorithm, which provides a base model trained with a labelled corpus, with a rule-based expert system that improves the results. Many ML researchers have found that labelled data, when used in conjunction with a rule-based expert systems can produce considerable improvement in learning accuracy. Reinforcement learning is another technique that allows learning of a suitable policy for negation classification directly through trial-and-error experience. In contrast to explicit teaching using supervised learning methods, the fundamental idea of the reinforcement learning approach is to learn a so called agent from the outcome of its actions on the basis of experience. This method tries to replicate human-like learning and thus appears well suited for NLP.

III. APPLICATIONS

Investigation and analysis of tweets is conceivably significant for many ongoing topics of interest such as understanding how topics evolve together with the underlying social interaction between participants and distinguish vital members who have great influence in various topics of discussions. Considering the proposed methodology of comparative analysis, in the following section we present several application regions of TSA techniques.

Brand sentiment analysis: Twitter messages are increasingly used to determine consumer sentiment towards a brand. Twitter offers a unique dataset in the world of brand sentiment. Public figures and brands receive sentiment messages directly from consumers in real time in a public forum. Both the targeted and competing brands have the opportunity to dissect these messages to determine changes in consumer sentiment [26]. Taking advantage of this data, however, requires researchers to deal with analyzing an immense amount of data produced by Twitter each day, referred to as the Twitter fire hose. For instance, TripAdvisor is an American travel website company providing reviews from travelers about their experiences in hotels, restaurants, and monuments. Stephen Kaufer and Langley Steinert, along with others, founded TripAdvisor in February 2000 as a site listing information from guidebooks, newspapers, and magazines.

Forecasting political trend: Today Twitter stands out as one of the most popular micro-blogging services, where information propagates in no time, and words and actions trigger immediate responses from users. Such an environment is ideal for advertising political views, especially during the heat of election campaigns [27, 28]. In order to predict how support or opposition toward a candidate would spread, or "diffuse," on Twitter, we learned statistical models from the Twitter data. Such models are called sentiment diffusion models, and we used these models to forecast the eventual winner of these elections.

Emotion detection: Online social network platforms, with their large-scale repositories of user-generated content, can provide unique opportunities to gain insights into the emotional "pulse of the nation", and indeed the global community [29]. Twitter (San Francisco, USA) offers the opportunity for the analysis of expressed mood, and previous studies have shown that geographical, diurnal, weekly and seasonal patterns of positive and negative affect can be observed. Studies have also examined the sentiment expressed on Twitter during specific sporting events, and to measure levels of happiness of cities within specific countries.

Implicit sentiment in financial news: The application of sentiment analysis to financial newspaper text enables

researchers in the field of finance to identify positive and negative company news in an automatic way [30]. Consequently, more data can be processed in less time, which could lead to new insights into the correlations between news (media) and the stock markets.

Measuring public health concerns: An important task of public health officials is to keep track of health issues, such as spreading epidemics [31]. We explore the potential of mining social network data, such as tweets, to provide a tool for public health specialists and government decision makers to gauge the measure of concern (MOC) expressed by Twitter users under the impact of diseases.

Sarcasm Detection: Sarcasm is a sophisticated form of irony widely used in social networks and micro-blogging websites. It is usually used to convey implicit information within the message a person transmits. Sarcasm might be used for different purposes, such as criticism or mockery [32]. However, it is hard even for humans to recognize. Therefore, recognizing sarcastic statements can be very useful to improve automatic sentiment analysis of data collected from micro-blogging websites or social networks. Hate speech refers to the use of aggressive, violent or offensive language, targeting a specific group of people sharing a common property, whether this property is their gender (i.e., sexism), their ethnic group or race (i.e., racism) or their believes and religion. While most of the online social networks and micro-blogging websites forbid the use of hate speech, the size of these networks and websites makes it almost impossible to control all of their content.

Topic Detection: The process of extracting and summarizing trending issues in the form of useful information is called topic detection [3]. Conventional approaches are not applicable to the documents produced by Twitter, which consist of extremely short sentences containing misspelled words. Moreover, since tweets often include not only sentences but also special characters or URLs, the ratio of tweets including useful information is very low in comparison with the volume of all tweets. In addition, it is difficult to identify correlations between tweets due to the diversity of issues, and this makes it hard to distinguish topics.

IV. CONCLUSION

This survey article presents the state-of-the-art techniques and applications in the field of TSA. The cited articles were categorized and summarized according to their significance. The review additionally shed light on various applications that asserts the importance of automated TSA. In spite of the fact that, the continuous evolution of TSA, still there are ample opportunities and challenges for researchers. Interest in UGSMC for regional languages other than English is growing as there is still a lack of tools and technologies concerning these languages. Lexicons similar to WordNet which supports many regional languages other than English need to be developed. In many cases, opinions are very much dependent on the context. Hence, it is important to consider the context of the opinion and research more on context-based TSA systems.

REFERENCES

[1] Min-Chul Yang and Hae-Chang Rim, 'Identifying interesting Twitter contents using topical analysis', Expert Systems with Applications, Expert Systems with Applications, 41 (2014) 4330–4336.

[2] Zhao Jianqiang and Gui Xiaolin, Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis, IEEE Access, 5 (2017), 2870-2879.

[3] Hyeok-Jun Choi and Cheong Hee Park, Emerging Topic Detection in Twitter Stream based on High Utility Pattern Mining, Expert Systems With Applications, (2018), doi:10.1016/j.eswa.2018.07.051.

[4] Ortigosa, A., et al. Sentiment analysis in Facebook and its application to e-learning. Computers in Human Behavior (2013), http://dx.doi.org/10.1016/j.chb.2013.05.024.

[5] Feldman, R. and Dagan, I. (1995) 'Knowledge discovery in textual databases', 1st International Conference on Knowledge Discovery and Data Mining, pp.112–117.

[6] Salton, G. and Yang, C.S. (1975) 'A vector space model for automatic indexing', Communications of the ACM, Vol. 18, No. 11, pp.613–620.

[7] Tsatsaronis, G., Varlamis, I. and Nørvg, K. (2010) 'An experimental study on unsupervised graph-based word sense disambiguation', Computational Linguistics and Intelligent Text Processing, LNCS, Vol. 6008, pp.184–198, Springer.

[8] Zhang, X., Xu, C., Cheng, J., Lu, H. and Ma, S. (2009) 'Effective annotation and search for video blogs with integration of context and content analysis', IEEE Transactions on Multimedia Special Issue on Integration of Context and Content, Vol. 11, No. 2.

[9] Porter, M.F. (2006) 'An algorithm for suffix stripping', Electronic Library and Electronic Systems, Vol. 40, pp.211–218

[10] Pappas, N., Katsimpras, G. and Stamatatos, E. (2012) 'Extracting informative textual parts from web pages containing user-generated content', Proceedings of 12th International Conference on Knowledge Management and Knowledge Technologies, No. 4, pp.1–8.

[11] Baccianella, S., Esuli, A. and Sebastiani, F. (2010) 'Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining', Proceedings of

the Annual Conference on Language Resources and Evaluation, pp.2200–2204.

[12] Turney, P. (2000) 'Learning algorithms for keyphrase extraction', Information Retrieval, Vol. 2, pp.303–336.

[13] Fang, X. and Zhan, J. (2015) 'Sentiment analysis using product review data', Journal of Big Data, Vol. 2, No. 5, Springer.

[14] Nadeau, D. and Sekine, S. (2007) 'A survey of named entity recognition and classification', Lingvisticae Investigationes, pp.3–26.

[15] Pang, L., Zhu, S. and Ngo, C-W. (2015) 'Deep multimodal learning for affective analysis and retrieval', IEEE Transactions on Multimedia, Vol. 17, No. 11, pp.2008–2020.

[16] Pröllochs, N., Feuerriegel, S. and Neumann, D. (2016) 'Negation scope detection in sentiment analysis: decision support for news-driven trading', Decision Support Systems, Vol. 88, pp.67–75, Elsevier.

[17] Araujo, L. and Martinez-Romo, J. (2010) 'Web spam detection new classification features based on qualified link analysis and language models', IEEE Transactions on Information Forensics and Security, Vol. 5, No. 3.

[18] Aas, K. and Eikvil, L. (1999) Text Categorization: A Survey, Technical report, Norwegian Computing Center.

[19] Liu, C-L., Hsaio, W-H., Lee, C-H. and Chi, H-C. (2013) 'An HMM based algorithm for content ranking and coherence feature extraction', IEEE Transactions on Systems, Man, and Cybernetics: Systems, Vol. 43, No. 2, pp.440–450.

[20] Forman, G. (2003) 'An extensive empirical study of feature selection metrics for text classification', J. Mach. Learn. Res., Vol. 3, pp.1289–1305.

[21] Prieto, V.M., Álvarez, M., López-García, R. and Cacheda, F. (2012) 'Analysis and detection of web spam by means of web content', IRFC'12 Proceedings of the 5th conference on Multidisciplinary Information Retrieval, pp.43–57.

[22] Young, T.Y. (1971) 'The reliability of linear feature extractors', IEEE Transactions on Computers, Vol. 20, No. 9, pp.967–971.

[23] Chimphlee, S., Salim, N., Ngadiman, M.S.B., Chimphlee, W. and Srinoy, S. (2006) 'Independent component analysis and rough fuzzy based approach to web usage mining', Proceedings of the 24th International conference on Artificial intelligence and applications (IASTED), ACTA Press, Anaheim, pp.422–427.

[24] Taboada, M. et al. (2011) 'Lexicon-based methods for sentiment analysis', Computational linguistics, Vol. 37, No. 2, pp.267–307.

[25] Russell, S. and Norvig, P. (2003[1995]) Artificial Intelligence: A Modern Approach, 2nd ed., Prentice Hall.

[26] M. Ghiassi a, J. Skinner b and D. Zimbra, Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network, Expert Systems with Applications 40 (2013) 6266–6282.

[27] Aibek Makazhanov, Davood Rafiei and Muhammad Waqar, Predicting political preference of Twitter users, Soc. Netw. Anal. Min. (2014) 4:193.

[28] Vadim Kagan and Andrew Stevens and V.S. Subrahmanian, Using Twitter Sentiment to Forecast the 2013 Pakistani Election and the 2014 Indian Election, IEEE Intelligent Systems (2015), 1-5.

[29] Mark E. Larsen, We Feel: Mapping emotion on Twitter, IEEE Journal of Biomedical and Health Informatics, (2015) 1246-1252.

[30] Van de Kauter, M., et al. Fine-grained analysis of explicit and implicit sentiment in financial news articles. Expert Systems, with Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.02.007

[31] Xiang Ji et al., Twitter sentiment classification for measuring public health concerns, Soc. Netw. Anal. Min. (2015) 5:13.

[32] Mondher Bouazizi and Tomoaki Otsuki, A Pattern-Based Approach for Sarcasm Detection on Twitter, 4, 2016, 5477-5488.