International Journal of Advance and Applied Research
www.ijaar.co.in
ISSN – 2347-7075 Impact Factor – 7.328
Peer Reviewed Bi-Monthly
Vol.10 No.3 January – February 2023

METHODS FOR SENTIMENT ANALYSIS, USED FOR MINING DATA FROM TEXTS AND SOCIAL NETWORKS

Krutikaben Chandrakant Patel[1] & Dr. Shabnam Sharma[2]

[1]*Ph.D. Research Scholar, Department of Computer Science, Shri JJTU, Rajasthan, India*

[2]*Professor &Ph.D. Guide, Department of Computer Science, Shri JJTU, Rajasthan, India*

Corresponding Author – Krutikaben Chandrakant Patel

DOI - 10.5281/zenodo.7895398

Abstract:

Social networks (SNs) are well-established online communities in which individuals regularly exchange their thoughts and feelings with one another. Because of this, they have developed into an important source of big data relating to the opinion and sentiment arena. The goal of sentiment analysis, often known as SA, is to glean feelings, thoughts, or views from texts, which may be obtained from a variety of data sources such as SNs. This research offers an in-depth investigation of the procedures and primary instruments that are used in SA. In order to carry out the study, four criteria and a number of variables had to first be defined. The analysis then compared 24 tools based on objective criteria. In particular, the tools have been scrutinised and examined in order to validate their usability, flexibility of application, and other requirements relevant to the kind of study that was carried out. Positive, negative, and neutral polarity can all be detected by the vast majority of equipment, however only a small number of tools can only detect positive and negative polarity. In addition, seven of the tools were capable of identifying emotions, but only one of them offered a visual map for data that was georeferenced. The following 24 tools all provide their services via a web interface, with the exception of one. Finally, only nine tools offer both application programme interfaces and a client for common programming languages, making it possible for potential developer end-users to integrate a specific SA tool into their application. These tools are categorised as "tools that provide both application programme interfaces and a client for common programming languages." The report, in contrast to previous recent studies, offers and examines both methodologies and tools for evaluating texts and SN data sources in order to extract sentiment. Other surveys have focused just on one or the other. Additionally, it includes a full comparison with other recent polls that have been conducted. The comparative study of the tools that was carried out according to the objective criteria enables the highlighting of certain limitations on the primary tools that need to be addressed with the goal of improving the experience of the end user.

Keywords: Data mining, emotion detection, polarity detection, sentiment analysis, social networks

Introduction:

In recent years, there has been a tremendous increase in the usage of social networks (SNs), and users have been known to utilise SNs to share both their experiences and their views. The practise of publishing or disseminating written, aural, and visual information on a social networking platform is one that is gaining more and more adherents (Choi & Toma, 2014; Hidalgo, Tan, &Verlegh, 2015). As a result, a variety of information, either implicit or explicit, that may be relevant to the user's mental or physical health is posted, and this information may be potentially retrieved from such postings (Rosenquist, Fowler, & Christakis, 2011).

The users of social networking sites exhibit relationships among themselves, which may be either explicit or implicit, and this results in the establishment of communities. According to Coviello et al. (2014), users may be susceptible to phenomena such as emotional contagion among communities. This means that emotions may spread among users who are members of the same community. As a result, SN is an excellent data-source for the study of both individual and communal life and behaviour because of its high level of dynamic complexity. The interpretation of this data, also known as data mining, is becoming an increasingly popular subject of discussion among companies as well as scholars who are particularly interested in the study of feelings.

The automated identification of sentiment and emotions derived from SNs data is often accomplished via the use of approaches and tools originating from the discipline of sentiment analysis (SA). The purpose of SA is to extract feelings, emotions, or views from texts. These texts may be obtained from a variety of textual data sources, ranging from plain text to many web contents such as news stories, public polls, blog entries, and of course, SNs data (Liu, 2015). Research in the area that applies SA to SNs is mostly concentrated on tweets for a number of reasons, the primary one being that tweets are easier to get than other types of SNs data. The first thing that industrial firms did with SA was utilise it to find out certain specifics, like how popular their brand was or how many people liked their goods. In more recent times, SA is also becoming more popular in a number of medical professions, including as psychiatry. Coppersmith, Harman, and Dredze (2014), for instance, utilised data taken from Twitter to diagnose certain diseases, such as posttraumatic stress disorder (PTSD). Tweets were also utilised for a variety of additional purposes, such

as health monitoring (Carchiolo, Longheu, & Malgeri, 2015), the detection of suicide notes (O'Dea et al., 2015), and the prediction of cardiac illness (Eichstaedt et al., 2015), amongst other things (Kim, Park, & Jo, 2014). Over the last several years, SA has attracted a lot of attention thanks in large part to the rising demand for public opinion monitoring tools among end users. For instance, a prospective end-user may prefer to keep an eye on his or her own brand by using his or her Twitter account. The increasing interest shown by the scientific community is shown by the publication of a large number of studies that conduct a survey of the usage of SA to extract sentiment and emotions both from written texts and from SNs. A detailed analysis of software tools for social networking media, wikis, blogs, and chats is provided by Batrinca and Treleaven (2015). However, this study does not go into depth about the SA approaches that have been used to SN; rather, it only gives scientists who are interested in using social media scraping and analytics an overview of the topic. This article by Serrano-Guerrero, Olivas, Romero, and Herrera-Viedma (2015) reviews and compares 15 free SA tools that are accessible as online services. The authors analyse the capability of each tool to conduct SA classification on three distinct datasets. The focus of the study is on the analytic capabilities of the various tools, and there is no mention of SA being applied to SNs anywhere in the document. The objective of this study is to provide an overview of some of the most recent and cutting-edge approaches that pertain to the SA process in general (both when SA is performed on a plain text dataset or data extracted from SNs). After then, a number of different SA tools were reviewed.

In order to identify SNs monitoring tools from generic SA tools, the first step in the classification process is to undertake a wide categorization. Tools that provide the functionality of extracting data from SNs based on some user research criteria (for example, username, keyword, or URL), and then performing some SA tasks (these tools belong to the SN tools class), are considered to be part of the first class. Tools that perform SA on plain text loaded by the user are considered to be part of the second class. After that, an in-depth investigation of each instrument is carried out, with careful attention paid to four significant criteria or orthogonal analytical aspects, namely technology, interoperability, visualisation, and analysis. This work is different from other surveys that have been done in the following ways: (a) both methodologies and tools for SA performed on both texts files and data extracted from SNs are considered; (b) in order to approach the

SA problem from the perspective of the end user, four major criteria or orthogonal analysis dimensions are defined, and these are as follows: technology, interoperability, visualisation, and analysis. On the other hand, each of these overarching dimensions is further subdivided into a large number of variables. Regarding the examination of existing SA comparison tools, this method seems to be original to the best of our knowledge. In addition, (c) the behaviour of multilingual tools is evaluated, and each tool is systematically evaluated according to the aforementioned dimensions; and (d) the majority of the tools have been tested in order to evaluate the variables for each dimension, as well as to make a general judgement regarding the quality of the analysis that was carried out. As a result, this comparative analysis enables the identification of several flaws that are presently present in SA products and that need to be addressed in order to improve the overall experience provided to end users.

Basic Tasks of SA:

The number of people using social networks (SNs) has increased at a fast rate over the last several years. Users of SNs have been known to share both their experiences and their views. The practise of publishing or disseminating written, audible, and visual information on a social networking platform is one that is gaining more and more adherents (Choi & Toma, 2014; Hidalgo, Tan, &Verlegh, 2015). As a result, many types of implicit or explicit information, maybe pertaining to the user's mental or physical health, isposted and might potentially be gleaned from such postings (Rosenquist, Fowler, & Christakis, 2011).

Communities are formed as a result of the relationships that are made between users of social networking sites, which may be either explicit or implicit. According to Coviello et al. (2014), individuals who are part of the same community might be susceptible to a phenomenon known as emotional contagion. This means that feelings can spread from one user to another within the same community. As a result, SN is a very dynamic data source that may be used to investigate both individual and communal forms of life and behaviour. The interpretation of this data, also known as data mining, is a subject that is becoming more popular among both companies and scholars that are interested in the study of attitudes in particular.

The automated detection of sentiment and emotions gleaned from SNs data is often accomplished via the use of approaches and tools originating from the discipline of sentiment analysis (SA). The

purpose of SA is to extract feelings, emotions, or views from texts that have been made accessible from a variety of textual data sources ranging from plain text to many web contents such as for example news stories, public polls, blog entries, and of course, SNs data (Liu, 2015). The area of study that applies SA to SNs focuses almost entirely on tweets, primarily because tweets can be obtained more readily than other types of SNs data. In the beginning, industrial businesses employed SA to find out certain specifics, such how well-known their brand was or whether or not consumers liked their goods. In recent years, SA has also been gaining popularity in a number of medical sectors, including psychiatry. Coppersmith, Harman, and Dredze (2014), for instance, utilised data taken from Twitter in order to diagnose certain illnesses, such as posttraumatic stress disorder (PTSD). Tweets were also utilised, for instance, for health monitoring (Carchiolo, Longheu, & Malgeri, 2015), for the detection of suicide notes (O'Dea et al., 2015), for the prediction of cardiac illness (Eichstaedt et al., 2015), and for a variety of other issues (Kim, Park, & Jo, 2014). In recent years, SA has acquired a lot of attention also because end users have been asking for more public opinion monitoring solutions. This is one of the reasons why. As an example, a prospective end-user could want to keep tabs on his or her own brand through Twitter. The increasing interest from the academic community is shown by the publication of a large number of studies that conduct a study of the usage of SA to extract sentiment and emotions from SNs as well as from written texts. Batrinca and Treleaven (2015) provide an in-depth analysis of software tools for social networking media such as wikis, blogs, and chat rooms. However, this paper does not give a detailed discussion on how SA approaches might be used to SN; rather, it merely offers an overview to scientists who are interested in using social media scraping and analytics. This article by Serrano-Guerrero, Olivas, Romero, and Herrera-Viedma (2015) reviews and compares 15 free SA tools that are accessible as online services. The authors analyse the capability of each tool to conduct SA classification on three distinct datasets. In the poll, the focus is on the analytic capabilities of the chosen tools, but there is no mention of SA being applied to SNs. This article's goal is to provide readers with an overview of the most recent and cutting-edge approaches that pertain to a broad SA process (both when SA is performed on a plain text dataset or data extracted from SNs). After that, a variety of SA instruments were dissected and examined.

To identify SNs monitoring tools from generic security analysis tools, a first broad classification is carried out. Tools that provide the functionality of extracting data from SNs based on some user research criteria (for example, username, keyword, or URL), and then performing some SA tasks (these tools belong to the SN tools class), are considered part of the first class. Tools that perform SA on plain text loaded by the user are considered part of the second class. Following that, an in-depth investigation of each instrument is carried out, with careful attention paid to four significant criteria or orthogonal analytical aspects, namely technology, interoperability, visualisation, and analysis. This work is different from other surveys that have been done in the following ways: (a) both methodologies and tools for SA performed on both text files and data extracted from SNs are considered; (b) in order to tackle the SA problem from the perspective of the end user, four major criteria or orthogonal analysis dimensions are defined, and these are as follows: technology, interoperability, visualisation, and analysis. On the other hand, each of these broad dimensions is further subdivided into a large number of variables. This method is unheard of in terms of our understanding of the SA comparative tools evaluation, to the best of our knowledge. In addition, (c) the behaviour of multilingual tools is evaluated, and (d) the majority of the tools have been tested in order to evaluate the variables for each dimension, as well as to make a general judgement on the quality of the analysis that was performed. (c) Each tool is systematically examined by those dimensions. Therefore, this comparative analysis enables the identification of specific shortcomings that are presently present in SA tools and that need to be addressed in order to improve the overall experience that end users have.

Sentiment Analysis Pipeline:

The illustration provides a broad overview of the SA process's workflow. Standard approaches for natural language processing (NLP) and text mining are used during the preprocessing phase, after which the input data are transformed into text. The analysis stage is the most important part of SA, and it may be carried out in two different ways: either by making use of an algorithm for machine learning or by using a lexicon-based approach. The second method involves the extraction of sentiment via the annotation of sentiment derived from language and lexical resources. As was mentioned in the previous section, the output can be binary (for instance, positive or negative in polarity detection) or multiclass. For

instance, a positive/negative/neutral polarity detection, a polarity detection task with scoring, or an emotion recognition task are all examples of multiclass outputs. The document analysis can be performed at a different level (see below), and the output can be binary (for instance, positive or negative in polarity detection). In specifically, we may differentiate between the following degrees of analysis:

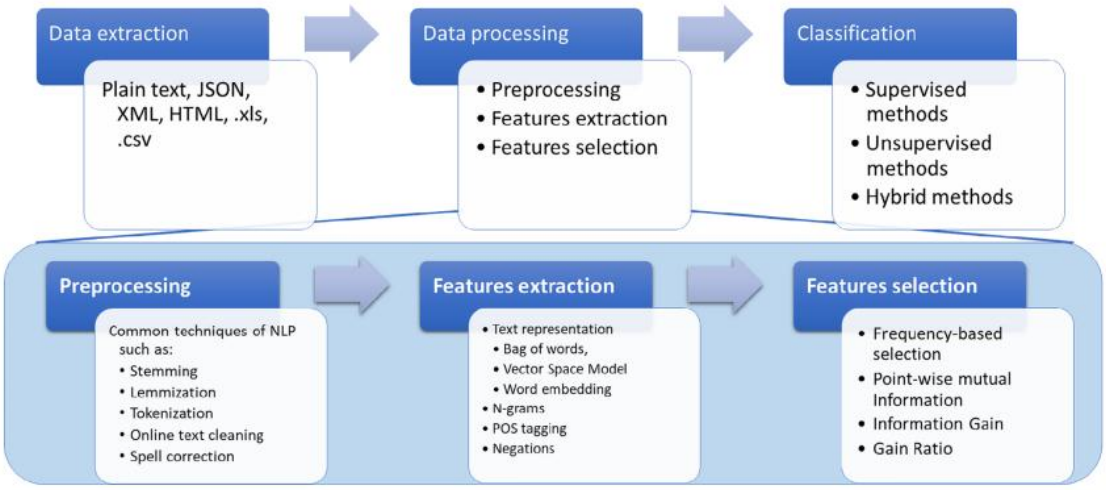

Figure 1 The Sentiment Analysis Process Pipeline

Data Preprocessing:

When texts are studied, there are numerous words that do not have an influence on the direction of the sentiments they convey. Since each word in the text is considered to represent a single dimension, for instance, question words such as "what," "how," and "when" do not make a contribution to the polarity of the text; hence, it is possible to eliminate them in order to minimise the dimensionality of the issue. In addition, particularly in cases where the messages originated from online SNs, they are in their raw form and often include a great deal of background noise in addition to data sections that are either missing or inconsistent. As a result, a step known as data preprocessing, which is the procedure that is carried out to clean and prepare the text for classification, becomes essential. The preprocessing stage in SA is quite comparable to the text preprocessing stage traditionally used in text mining.

Literature Surveys Comparison:

The purpose of this section is to provide a summary of these surveys in order to highlight the ways in which they vary from the current work. We took into consideration 33 SA reviews and categorised them according to the four macrocategories that follow: general sentiment analysis (GSA) review, domain-oriented (DO) sentiment analysis review, specific task or specific technology

oriented (STO) sentiment analysis review, and tool's classification (TC) review. The bulk of the surveys that were evaluated had the objective of providing a general overview of the primary tasks, important obstacles, and SA approaches. These surveys are classified under the more general GSA category. Generic Sentiment Analysis (GSA) Class Was Further Divided Into Generic Sentiment Analysis Overview (GSA-O) And Generic Sentiment Analysis Comparative (GSA-C) Review The GSA class was further divided into generic sentiment analysis overview (GSA-O), which main focus is to give a critical picture of SA framework, without deeply going into details with respect to algorithms or approaches or tasks, and generic sentiment analysis comparative (GSA-C) review, which compares some existing works generally.

A quantitative investigation of South African publications was carried out by Mntyl et al. (2016), who made use of the conventional methods and instruments of bibliometric research, in addition to clustering and text mining strategies. The authors of this research found that there were over 7,000 papers pertaining to SA, practically all of which were published after the year 2004. According to these findings, SA is a subject of interest in a number of different study domains, and it also represents one of the areas where research is expanding at a quicker rate.

It is shown once again in Mntyl et al. (2016) that four of the twenty publications that have received the highest citations from Google Scholar and Scopus are literature reviews. In specifically, the research conducted by Pang and Lee was found to be the publication that received the most citations in both Google Scholar and Scopus (2008). The purpose of this article is to provide an introduction to several applications of SA, key core principles, and primary tasks, as well as an overview of fundamental technologies and methodologies. In their study, Medhat et al. (2014) looked at 54 publications that were published between 2010 and 2013 to determine the most significant advancements that have been made in terms of SA applications and algorithms. The authors identify a total of five distinct categories, which are as follows: GSA applications, sentiment classification, feature selection, emotion detection, building resources (also known as annotated lexica corpora or dictionaries), and transfer learning (also known as domain adaption) approaches. For each of the primary methods, a discussion is given, and also, a classification of the 54 publications depending on how the core SA tackled the issue is offered.

An exhaustive literature review of more than 160 papers is presented in Ravi and Ravi (2015). The authors outline seven main dimensions, five of which define different SA tasks, and other different subcategories dedicated to differentiating further reviewed articles by subtasks, approaches, techniques, or applications.

Conclusion:

An overview of the approaches and software tools connected to the SA process were offered in this article. In the existing SA system, 24 tools have been analysed, evaluated, and a comparative and systematic review of these tools has been carried out by defining four orthogonal coordinates and, for each size, by identifying the most significant variables. This review was carried out using the tools that are currently in use. Due to the studies that were carried out, we were able to determine that the majority of the tools that were taken into consideration had the capability to detect positive, negative, and neutral polarity, but just a few instruments can only identify positive and negative polarity. In addition, seven of the tools that were considered were able to identify emotions, but just one of the tools offered a visual map that could be used for geo-referenced data. The remaining 23 tools, with the exception of Opinion Finder, provide their services through a web-based interface. Finally, only nine of the remaining tools provide clients and APIs for the most popular programming languages. This leaves a small group of five tools that do not provide the exporting of functions via application programming interfaces (APIs). As was mentioned in the preceding sections, the primary contributions of this review are as follows: (a) an analysis methodology that is based on new several orthogonal dimensions of analysis; (b) a classification of 20 tools according to such variables; and (c) an extensive evaluation of the tools that is performed through several experiments on real data that is extracted primarily from Twitter. Each of these contributions is described in more detail below. In addition, the current research, in contrast to other studies that have been previously published, offers a full overview of approaches for SA pipeline as well as a number of software tools for online and social network texts.

References:

[1]. Abbasi, A., Hassan, A., & Dhar, M. (2014). Benchmarking Twitter sentiment analysis tools. In LREC (Vol. 14, pp. 26–31). Reykjavik, Iceland: European Language Resources Association (ELRA).

[2]. Abdulla, N. A., Ahmed, N. A., Shehab, M. A., Al-Ayyoub, M.,

Al-Kabi, M. N., & Al-rifai, S. (2014). Towards improving the lexicon-based approach for Arabic sentiment analysis. International Journal of Information Technology and Web Engineering, 9(3), 55–71.

[3]. Abirami, A. M., & Gayathri, V. (2017). A survey on sentiment analysis methods and approach. In 2016 eighth international conference on advanced computing (ICOAC), Chennai, India (pp. 72–76). Piscataway, NJ: IEEE. https://doi.org/10.1109/ICoAC.2017.7951748

[4]. Agerri, R., &Garcıa-Serrano, A. (2010). Q-wordnet: Extracting polarity from wordnet senses. In LREC. Valletta, Malta: European Language Resources Association (ELRA).

[5]. Appel, O., Chiclana, F., Carter, J., & Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level. Knowledge-Based Systems, 108, 110–124.

[6]. Arnold, M. (1960). Emotion and personality. New york, NY: Columbia University Press.

[7]. Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In LREC (Vol. 10, pp. 2200–2204). Valletta, Malta: European Language Resources Association (ELRA).

[8]. Balazs, J. A., & Velásquez, J. D. (2016). Opinion mining and information fusion: A survey. Information Fusion, 27, 95–110.

[9]. Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: A survey of techniques, tools and platforms. AI & SOCIETY, 30(1), 89–116.

[10]. Boiy, E., & Moens, M.-F. (2009). A machine learning approach to sentiment analysis in multilingual web texts. Information Retrieval, 12(5), 526–558.

[11]. Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In proceedings of the Biennial German Society for Computational Linguistics Conference (GSCL) (pp. 31–41).

[12]. Brody, S., &Diakopoulos, N. (2011). Co: using word lengthening to detect sentiment in microblogs. In Proceedings of the conference on empirical methods in natural language processing, Edinburgh, Scotland (pp. 562–570). Stroudsburg, PA: Association for Computational Linguistics.

[13]. Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, 28(2), 15–21.

[14]. Carchiolo, V., Longheu, A., & Malgeri, M. (2015). Using Twitter data and sentiment analysis to study diseases dynamics. In Information technology in bio-and medical informatics (pp. 16–24). Berlin: Springer.

[15]. Chaturvedi, I., Cambria, E., Welsch, R. E., & Herrera, F. (2018). Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. Information Fusion, 44, 65–77.

[16]. Chaudhari, P., &Chandankhede, C. (2017, March). Literature survey of sarcasm detection. In 2017 international conference on wireless communications, signal processing and networking (WiSPNET), Chennai, India (pp. 2041–2046). Piscataway, NJ: IEEE. https://doi.org/10.1109/ WiSPNET.2017.8300120

[17]. Choi, M., & Toma, C. L. (2014). Social sharing through interpersonal media: Patterns and effects on emotional well-being. Computers in Human Behavior, 36, 530–541.

[18]. Chopra, F. K., & Bhatia, R. (2016). A critical review of sentiment analysis. International Journal of Computer Applications, 149(10), 37–40.

[19]. Coppersmith, G., Harman, C., &Dredze, M. (2014). Measuring post traumatic stress disorder in Twitter. In ICWSM. Palo Alto, CA: Association for the Advancement of Artificial Intelligence.

[20]. Coviello, L., Sohn, Y., Kramer, A. D., Marlow, C., Franceschetti, M., Christakis, N. A., & Fowler, J. H. (2014). Detecting emotional contagion in massive social networks. PLoS One, 9(3), e90315. da Silva, N. F. F., Coletta, L. F., Hruschka, E. R., & Hruschka, E. R., Jr. (2016). Using unsupervised information to improve semi-supervised tweet sentiment classification. Information Sciences, 355, 348–365.

[21]. Das, S. R., & Chen, M. Y. (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web. Management Science, 53(9), 1375–1388. Desai, M., & Mehta, M. A. (2016). A hybrid classification algorithm to classify engineering students' problems and

perks. arXiv preprint arXiv: 1604.02358.

[22]. Devika, M., Sunitha, C., & Ganesh, A. (2016). Sentiment analysis: A comparative study on different approaches. Procedia Computer Science, 87, 44–49.

[23]. Drake, A., Ringger, E., & Ventura, D. (2008). Sentiment regression: Using real-valued scores to summarize overall document sentiment. In 2008 IEEE international conference on semantic computing, Santa Clara, CA (pp. 152–157). Washington, D.C.: IEEE Computer Society.

[24]. Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., … Seligman, M. E. (2015). Psychological language on Twitter predicts county-level heart disease mortality. Psychological Science, 26(2), 159–169.

[25]. Ekman, P., & Wallace, V. (2003). Unmasking the face. Cambridge, MA: Malor Book.