An Empirical Analysis of Articles on Sentiment Analysis

Vishal Vyas and V. Uma

Abstract Expression of a thought is not only important for an individual but there is a necessity for an automated system to get an opinion from it. Sentiment analysis (SA) or opinion mining (OM) is used to identify the sentiment/opinion of the speaker. Web 2.0 provides us various platforms such as Twitter, Facebook where we comment or post to express our happiness, anger, disbelief, sadness, etc. For SA of text, computationally it is required to know the concepts and technologies being used in the field of SA. This article gives brief knowledge about the techniques used in SA by categorizing various articles over the past four years. This article also explains the preprocessing steps, various application programmable interface (API), and available datasets for a better understanding of SA. This article is concluded with a future work which needs a separate attention of researchers to improve the performance of sentiment analysis.

Keywords Text mining · Sentiment analysis · Ontology · Machine learning

1 Introduction

For a human being mostly it is very easy to sense the sentiment in the text using their trained mind. Humans have trained their mind by learning through experiences. An inexperienced person always demands an opinion from others while trying new things. An automated system that can correctly identify the polarity is the need of the hour. Although there is a huge advancement in field of natural language processing (NLP) and machine learning (ML), automated systems still have not achieved 100% accuracy in dealing with sarcasm or finding the polarity of the text. With the advance-

V. Vyas (✉) · V. Uma
Department of Computer Science, Pondicherry University,
Puducherry 605014, India
e-mail: vyasvishaluni@gmail.com

V. Uma
e-mail: umabskr@gmail.com

P. K. Pattnaik et al. (eds.), *Progress in Computing, Analytics and Networking*,
Advances in Intelligent Systems and Computing 710,
https://doi.org/10.1007/978-981-10-7871-2_8

ment in the field of NLP and ML in the last decade, there is a growing need for sentiment analysis systems in order to help humans in getting accurate opinion which could subsequently help them in decision-making. Sentiment analysis (SA) is a field where we try to get the point of view, belief, and intensity toward entities such as organizations, manufactured items, creatures, occasion, and their elements by using ML [1], NLP and many other ontology-based mining algorithms. Sentiment analysis is a task that is becoming increasingly important for many companies because of the emergence of social media sites such as Facebook, Twitter, e-commerce Web sites, and the other trillions of them. Business organizations track tweets about their products to know about the people's demand and to modernize their impact over time, whereas politicians use them to track their campaign by looking around comments on the social media Web sites. With the comments and feedbacks on the Internet, it is necessary to have an automated system that can make sense out of them. Opinions are important both at personal and professional level. Either we ask for opinion or we get influenced by the advertisement that business organizations put on Internet after huge research. The existing approaches to deduce sentiments from the posts, reviews, tweets, and forums which are available in social media Web sites can be summarized in points stated below:

1. Text to analyze in the social media is unstructured in nature. While working in the noisy environment, the identification of the sentiment from the text is not an easy task. Forming a quintuple [2] reduces the noisy labels but in the presence of sarcasm and smaller sentences the formation of quintuple becomes difficult. The creation of emoticon vocabulary [3] is the other way to tackle the noise in the text. The vocabulary is used to train the ML classifiers such as support vector machine (SVM), Naive Bayes (NB), maximum entropy (MaxEnt). The trained SVM classifier has a high accuracy in analyzing the sentiment from the text.
2. To improve the sentiment classification accuracy, various approaches such as feature-based model and the tree kernel-based model, n-gram and lexicon features have been combined with machine learning methods [4].
3. Ontology can be defined as an explicit formal specification of concepts which is machine readable [5]. SA using ontology is done in two phases. (a) Creating a domain ontology which includes formal concepts analysis and ontology learning. (b) Analysis of the sentiment as per the concepts of ontology. The course of SA consists of classification process at different level, namely document-level sentiment analysis, sentence-level sentiment analysis, aspect-based sentiment analysis, and comparative-based sentiment analysis. In document-level sentiment analysis, whole document is considered as single information unit and classified as positive, negative, or neutral sentiment polarity. Sentence-level sentiment analysis considers each sentence as one information unit. In aspect-based sentiment analysis, classification of sentiment is with respect to particular aspect/entity and in comparative-based sentiment analysis, rather than having a direct opinion about a product, text have comparative opinions. The sole purpose of this survey is to cover the techniques for SA which comprises of machine learning techniques for automatic learning of patterns in data and ontology for better

visualization of data to determine sentiment. Over the years, much research work have been carried out in the field of machine learning. In the age of Semantic Web, there is a need to explore ontology and machine learning together to analyze and classify the sentiments. This survey follows the pattern of [1] for the analysis of articles but the perspective here is an empirical categorization of articles by considering different techniques for SA. This survey is beneficial for researchers in the field of SA in various ways. Firstly, categorization of articles is done based on the approaches used for SA. Secondly, in-brief explanation of essential steps involved prior to SA is provided. Thirdly, the year-wise categorization of recent articles is presented on the basis of concepts and techniques, dataset, and data resources. Finally, the categorization of articles is analyzed with the help of bar charts and future work is discussed which needs a separate attention of researchers to improve sentiment analysis.

2 Sentiment Analysis Using Various Approaches

Approaches used for SA are machine learning and lexicon based. When two or more approaches are combined then it becomes a hybrid approach. Various algorithms which are also known as the types of machine learning are supervised and unsupervised machine learning algorithms. Lexicon-based approach contains dictionary-based approach (DBA) and corpus-based approach. Ontology building is a series of process and Web 2.0 empowers us with ontology creation for SA. Mukherjee and Joshi [6], Weichselbraun et al. [7], and Penalver-Martinez et al. [8] show that using ontologies it is possible to achieve tasks of SA. Table 1 categorizes various articles based on the approaches used in SA. Third column in this table specifies the task performed in the article. The table shows that the same task can be achieved using different approaches.

3 Preliminary Steps Involved in Sentiment Analysis

SA is performed in stages and it is better to call it as multifaceted problem [19]. The ample availability of heterogeneous online resources gives rise to the first step required for SA, which is data acquisition. The analysis approaches change with the various forms of data or multimedia data. Application programming interface (API) provided by microblogging sites such as Twitter, Facebook makes it easy for collecting public data, whereas few other sources are available which provide domain-oriented (movie, car, etc.) datasets. Available APIs and datasets for public use are discussed in Table 2.

Preprocessing is the second step before actual analysis starts. The data acquired from various data resources in the first step is in raw form which requires formatting. Various techniques involved in preprocessing are highlighted in Table 3.

Table 1 Approaches for sentiment analysis

References	Approach	Task
Blair-Goldensohn et al. [9]	Supervised machine learning	Binary classification of text
Lu et al. [10]	Supervised machine learning	5-star rating
Jakob and Gurevych [11]	Supervised machine learning	Opinion mining
Titov and McDonald [12]	Unsupervised machine learning	Aspect detection
Lakkaraju et al. [13]	Unsupervised machine learning	Aspect detection
Wang et al. [14]	Unsupervised machine learning	Aspect rating prediction
Popescu and Etzioni [15]	Hybrid method	Aspect detection
Raju et al. [16]	Hybrid method	Aspect detection
Mukherjee and Joshi [6]	Ontology	Binary classification of text
Weichselbraun et al. [7]	Ontology	Binary classification of text
Penalver-Martinez et al. [8]	Ontology	Binary classification of text
Moghaddam and Ester [17]	Dictionary-based approach (DBA)	Binary classification of text
Zhu et al. [18]	Dictionary-based approach (DBA)	Aspect sentiment analysis

Table 2 Articles using different API/Datasets for SA

References	Name	API/Datasets	Purpose
Kumar et al. [20]	Twitter REST	API	To extract profile information
Khan et al. [21], Kontopoulos et al. [22]	Twitter4J	API	Extract streaming tweets
Ortigosa et al. [23], Li and Xu [24]	Facebook Graph	API	Extract posts
Cruz et al. [25]	TBOD	DATASET	Reviews
Kouloumpis et al. [26]	EMOT	DATASET	Tweets and emoticons

Table 3 Steps involved in preprocessing of text for SA

Steps	Description
TOKENIZATION	Breaks sentence into meaningful tokens
STOPWORD REMOVAL	Removing stopword
STEMMING	Brings word to its root form
POS TAGGING	Recognizes different part of speech in the text
FEATURE EXTRACTION	Tackles the extreme noise in data captured

4 Literature Survey and Categorization of Articles

The survey is done with the aim to know the various concepts and technologies being used for SA. For this purpose, seventeen articles are summarized in Table 4. Table 4 contains five columns, where the first and second columns have the details regarding the survey papers and the year of publications, respectively. Third column shows the various concepts and technologies used in different articles. Fourth column specifies the domain of the data used for SA. Fifth column specifies the well-known datasets used in different articles.

Table 4 Categorization of articles

References	Year	Concept and technology	Type of data	Data set/Data source
Penalver-Martinez et al. [8]	2014	Ontology, DBA	Movie review	SWN
Bravo-Marquez et al. [27]	2014	Logistic regression	Microblog	Ten dictionaries
Mukherjee and Joshi [6]	2014	Ontology, DBA	Product review	SWN, GI, OL
Mukherjee and Joshi [30]	2014	DBA	Movie review	WN
Cambria et al. [31]	2014	DBA	Global domain	CN, DBPedia, WN
Poria et al. [32]	2014	Ontology, DBA	Global domain	AffectiveSpace
Poria et al. [33]	2014	Ontology, DBA	Global domain	SN 3, WNA
Weichselbraun et al. [7]	2014	Ontology, DBA	Product review	WN, CN
Krishnamoorthy [34]	2015	SVM	Product review	Amazon review datasets
Nakov et al. [35]	2016	NN	Tweets	Twitter API
Saif et al. [36]	2016	Lexicon-based senticircle	Tweets	STS-Gold
Wang and Cardie [37]	2016	SVM, RBF Kernal	Conversation	Wikipedia talk
Palomino et al. [38]	2016	Qualitative method	Nature health	Twitter API, Alchemy API
Poria et al. [29]	2016	MKL, CNN	Video	MOUD, ICT-MMMO
Poria et al. [39]	2017	Chi-square	Health	Health media collaboration
Ali et al. [40]	2017	Fuzzy ontology	Tweets, reviews	REST-API
Giatsoglou et al. [28]	2017	LBA	Tweets	Twitter API

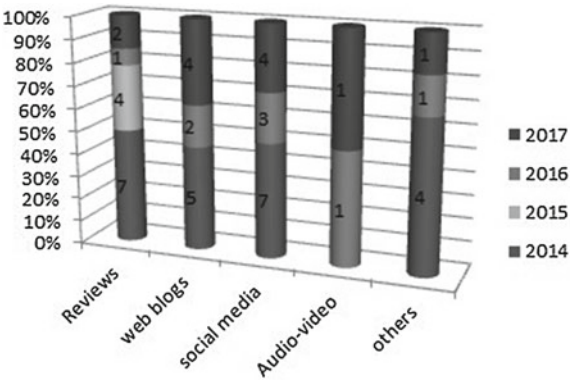

Fig. 1 Percentage and number of articles analyzing various text domains over the past 4 years

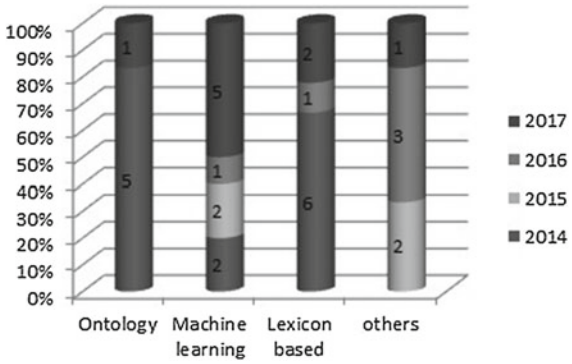

Fig. 2 Percentage and number of articles based on the approaches used over the past 4 years

Figure 1 shows that using product reviews and social media nearly 43% articles published in 2014, whereas [8, 27, 28] represent usage of microblogs for SA. Penalver-Martinez et al. [8] analyzed tweets using REST-API which is a publicly available dataset. Analysis of audio–video is done by Poria et al. [29] in 2016 using MKL and CNN.

Figure 2 shows nearly 80% of articles published in 2014 considered ontology approach for SA. Machine learning algorithm remains the best approach for SA in past 4 years. 60% of articles in 2014 and 20% in 2017 considered lexicon-based approach as a best way for SA.

5 Conclusion

Over the years, researchers have focused on the binary classification of the text by collecting posts from social networking Web sites, reviews from e-commerce Web sites, etc. Problem arises when it comes to conditional sentence. It is hard to identify the sentiment expressed by conditional sentence such as “If I find his address, I will

send her an invitation." Narayanan et al. [41] proposed an approach for the above-said problem. Using SVM classifier, they have achieved 75.6% and 66% accuracy for binary and ternary classification, respectively. High accuracy in the case of conditional sentences is to be achieved. Our future work will include comparing semantic and hybrid approaches to identify which will perform better when analyzing conditional sentences to identify the technique that will perform better.

References

1. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal 5(4) (2014) 1093–1113
2. Liu, B.: Web data mining: exploring hyperlinks, contents, and usage data. Springer Science & Business Media (2007)
3. Read, J.: Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: Proceedings of the ACL student research workshop, Association for Computational Linguistics (2005) 43–48
4. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics (2010) 36–44
5. Studer, R., Benjamins, V.R., Fensel, D.: Knowledge engineering: principles and methods. Data & knowledge engineering **25**(1–2) (1998) 161–197
6. Mukherjee, S., Joshi, S.: Sentiment aggregation using conceptnet ontology. In: IJCNLP. (2013) 570–578
7. Weichselbraun, A., Gindl, S., Scharl, A.: Enriching semantic knowledge bases for opinion mining in big data applications. Knowledge-Based Systems **69** (2014) 78–85
8. Penalver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Rodriguez-Garcia, M.A., Moreno, V., Fraga, A., Sanchez-Cervantes, J.L.: Feature-based opinion mining through ontologies. Expert Systems with Applications **41**(13) (2014) 5995–6008
9. Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G.A., Reynar, J.: Building a sentiment summarizer for local service reviews. In: WWW workshop on NLP in the information explosion era. Volume 14. (2008) 339–348
10. Lu, B., Ott, M., Cardie, C., Tsou, B.K.: Multi-aspect sentiment analysis with topic models. In: Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, IEEE (2011) 81–88
11. Jakob, N., Gurevych, I.: Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In: Proceedings of the 2010 conference on empirical methods in natural language processing, Association for Computational Linguistics (2010) 1035–1045
12. Titov, I., McDonald, R.: Modeling online reviews with multi-grain topic models. In: Proceedings of the 17th international conference on World Wide Web, ACM (2008) 111–120
13. Lakkaraju, H., Bhattacharyya, C., Bhattacharya, I., Merugu, S.: Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments. In: Proceedings of the 2011 SIAM international conference on data mining, SIAM (2011) 498–509
14. Wang, H., Lu, Y., Zhai, C.: Latent aspect rating analysis without aspect keyword supervision. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2011) 618–626
15. Popescu, A.M., Etzioni, O.: Extracting product features and opinions from reviews. In: Natural language processing and text mining. Springer (2007) 9–28
16. Raju, S., Pingali, P., Varma, V.: An unsupervised approach to product attribute extraction. In: European Conference on Information Retrieval, Springer (2009) 796–800

17. Moghaddam, S., Ester, M.: Opinion digger: an unsupervised opinion miner from unstructured product reviews. In: Proceedings of the 19th ACM international conference on Information and knowledge management, ACM (2010) 1825–1828
18. Zhu, J., Wang, H., Tsou, B.K., Zhu, M.: Multi-aspect opinion polling from textual reviews. In: Proceedings of the 18th ACM conference on Information and knowledge management, ACM (2009) 1799–1802
19. Liu, B.: Sentiment analysis: A multifaceted problem. IEEE Intelligent Systems **25**(3) (8 2010) 76–80
20. Kumar, S., Morstatter, F., Liu, H.: Twitter data analytics. Springer Science & Business Media (2013)
21. Khan, F.H., Bashir, S., Qamar, U.: Tom: Twitter opinion mining framework using hybrid classification scheme. Decision Support Systems **57** (2014) 245–257
22. Kontopoulos, E., Berberidis, C., Dergiades, T., Bassiliades, N.: Ontology-based sentiment analysis of twitter posts. Expert systems with applications **40**(10) (2013) 4065–4074
23. Ortigosa, A., Martín, J.M., Carro, R.M.: Sentiment analysis in facebook and its application to e-learning. Computers in Human Behavior **31** (2014) 527–541
24. Li, W., Xu, H.: Text-based emotion classification using emotion cause extraction. Expert Systems with Applications **41**(4) (2014) 1742–1749
25. Cruz, F.L., Troyano, J.A., Enríquez, F., Ortega, F.J., Vallejo, C.G.: A knowledge-rich approach to feature-based opinion extraction from product reviews. In: Proceedings of the 2nd international workshop on Search and mining user-generated contents, ACM (2010) 13–20
26. Kouloumpis, E., Wilson, T., Moore, J.D.: Twitter sentiment analysis: The good the bad and the omg! Icwsm **11**(538–541) (2011) 164
27. Bravo-Marquez, F., Mendoza, M., Poblete, B.: Meta-level sentiment models for big social data analysis. Knowledge-Based Systems **69** (2014) 86–99
28. Giatsoglou, M., Vozalis, M.G., Diamantaras, K., Vakali, A., Sarigiannidis, G., Chatzisavvas, K.C.: Sentiment analysis leveraging emotions and word embeddings. Expert Systems with Applications **69** (2017) 214–224
29. Poria, S., Chaturvedi, I., Cambria, E., Hussain, A.: Convolutional mkl based multimodal emotion recognition and sentiment analysis. In: Data Mining (ICDM), 2016 IEEE 16th International Conference on, IEEE (2016) 439–448
30. Mukherjee, S., Joshi, S.: Author-specific sentiment aggregation for polarity prediction of reviews. In: LREC. (2014) 3092–3099
31. Cambria, E., Olsher, D., Rajagopal, D.: Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In: Proceedings of the twenty-eighth AAAI conference on artificial intelligence, AAAI Press (2014) 1515–1521
32. Poria, S., Cambria, E., Winterstein, G., Huang, G.B.: Sentic patterns: Dependency-based rules for concept-level sentiment analysis. Knowledge-Based Systems **69** (2014) 45–63
33. Poria, S., Gelbukh, A., Cambria, E., Hussain, A., Huang, G.B.: Emosenticspace: A novel framework for affective common-sense reasoning. Knowledge-Based Systems **69** (2014) 108–123
34. Krishnamoorthy, S.: Linguistic features for review helpfulness prediction. Expert Systems with Applications **42**(7) (2015) 3751–3759
35. Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., Stoyanov, V.: Semeval-2016 task 4: Sentiment analysis in twitter. Proceedings of SemEval (2016) 1–18
36. Saif, H., He, Y., Fernandez, M., Alani, H.: Contextual semantics for sentiment analysis of twitter. Information Processing & Management **52**(1) (2016) 5–19
37. Wang, L., Cardie, C.: A piece of my mind: A sentiment analysis approach for online dispute detection. arXiv:1606.05704 (2016)
38. Palomino, M., Taylor, T., Göker, A., Isaacs, J., Warber, S.: The online dissemination of nature-health concepts: Lessons from sentiment analysis of social media relating to nature-deficit disorder. International journal of environmental research and public health **13**(1) (2016) 142
39. Poria, S., Peng, H., Hussain, A., Howard, N., Cambria, E.: Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. Neurocomputing (2017)

40. Ali, F., Kwak, D., Khan, P., Islam, S.R., Kim, K.H., Kwak, K.: Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling. Transportation Research Part C: Emerging Technologies **77** (2017) 33–48
41. Narayanan, R., Liu, B., Choudhary, A.: Sentiment analysis of conditional sentences. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1–Volume 1. EMNLP '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 180–189