Research article *University of Swabi Journal (USJ); Open access*

Sentiment Analysis of Twitter for Recommender System

[1]Sajida Fayyaz, [2]Amatul Musawir, [3]Hafiz Ali Hamza Gondal*, [4]Syed Muhammad Mahdi

[1-4]Department of Computer Sciences, The University of Lahore, Sargodha

Abstract

The web is flooded with evaluative text that proves to be worthwhile resource of opinions available on different products, markets, occasions or individuals and so on. People are more interested to shop online and prefer to go through user reviews before making a purchase. Opinions within reviews prove to be functional for manufacturers as well that it may assist enhancing the design plus quality. Numerous research works have been carried out, on diverse sort of reviews available online for various products, named as text mining and opinion mining. Few earlier approaches to calculate polarity are keyword spotting (identifying the keywords from certain text), the lexical affinity (probabilistic affinity used to allocate arbitrary words for a specific category to demonstrate either result is negative or else positive), the statistical method (functions on distinct patterns as well as word or event co-occurrences) plus concept level techniques (make use of semantic networks to infer information from concepts of natural language). Recommender systems so far have not mechanized for hybrid cars yet. The system of recommendation is to propose the latest trending hybrid cars. Important parameters are to be discovered from the information, experience and feedback of twitter users about hybrid cars. Investigating every single review even for a single hybrid car is not that much easy. Hence this study concentrates on sentiment classification techniques among which lexicon-based dictionary is implemented to carry out Natural Language processing (NLP) for optimizing the outcomes. System recommends on the basis of polarity mapping and mining fractional parts of data to designate in terms of positive or negative review. The obtained percentage of positivity of each car model after comparison determines which hybrid brand is prominent and trending.

Key words: Opinion Mining, Natural Language Processing NLP, Sentiment analysis, Lexicon based approach

**Corresponding author address:* Hafiz Ali Hamza Gondal

Department of Computer Sciences, The University of Lahore, Sargodha

+92-302-5151829, alihamzagondal12@yahoo.com

1. Introduction

The Information Extraction (IE) is a field of NLP concerned with discovering factual information within the text. Despite of additional tractable necessity, it's alive with challenges and hence a motivating research subject. Discovery of opinions and their extraction from reviews sited on internet is a fraction of latest research area advanced in previous decade. Opinion mining, known as sentiment analysis as well in compositions of science, learns the purpose and categorization of opinions or way of thinking articulated in the form of text, utilizing computing machines.

Opinions to extract positive/negative meanings on different blogs are carried out through two step procedures. A java wrapper named J-Twitter accompanied by twitter APT. Practical example of this setup is about the tweets regarding the battery of iPhone. Keyword ―iPhone battery finds tweets containing both words. It summarizes the statistics contrasting other applications [1]. Efficiently filtering the content that possesses negativity is a great deal as compared to an entire website filtration. Proposed technique performs text sentiment analysis as well as methods of feature engineering for text classification [2].Retailers with small online businesses are targeted to offer a recommendation system implementing ARM Association Rule Mining by C. Junnanet. al [3]. Research deals with using an open source database running for three years. Some of the implemented rules considered by the system are history of user's purchase, location, time and date plus current purchase made by the customer. Because of limited dataset pattern storage is divided for an easy access into tables. Process is carried out in four different stages the order acceptation, tables with queries, results of weight queries and finally the result summarization. Processing time is divided by the system using ARM plus utilizes the weighted criteria for an efficient recommendation proving itself worthy for small datasets. A systematize approach presented by A. Sajid et al [4] to extract topic plus the main title from short text of a single document. The method makes use of text mining online as well as the techniques of Natural Language Processing. Content overview can be easily grasped by its title reducing the headache of reading summary providing heading preview for which three types of approaches have been presented. Classification of document and pathway for new researchers is the twofold of the research.

Nowadays content generated by users is trending a lot on social media especially that is a means for sharing one's opinion on any subject for which blogs is one of the important aspect. Influential bloggers are being targeted by implementing efficient algorithm for measuring factors based on semantics of posted blogs, the content being analyzed quantitatively as well as associated readerships along with comments [5]. Sentiment classification, at sentence level, of reviews available over the web is carried out using machine learning. Subjective sentences are being extracted and labeled positive or else negative by Naïve Bayesian classifier analyzing its words. The Bags of Sentences (BOS) are been passed through trained Support Vector Machine for polarity calculation of sentences. Semantic orientation is figured out through contextual information and passed through simulation for evaluation [6]. Naïve Bayes (NB) together with Support Vector Machine (SVM) is implemented for text classification of Urdu language by comparing the statistical techniques. An enormous corpus is used for classifier's training. Particular techniques of pre-processing are executed, as classifiers are not able to understand the raw data, to generate a lexicon with reduced-feature. SVM resulted better in classification analysis than NB [7]. Table 1 shows sentiment techniques of classification implemented in various research works.

Table 1: Sentiment classification techniques performed in earlier researches

Technique	Language Dependent	Lexicon Usage	Tagged Review	Dataset Used	References
NB, ME, SVM	Yes	No	Yes	IMDB	[8]
SVM, NB	Yes	No	Yes	Amazon, CNET	[9]
Lexicon, Tagged reviews	Yes	Yes	Yes	Amazon, CNET	[10]
Lexicon	Yes	Yes	No	Luce, Yoka	[11]
Lexicon	Yes	Yes	No	IMDB, Skytrax	[12]
ML, Lexicon	Yes	Yes	Yes	Multi-Domain Sentiment	[13]
Lexicon, NLP	Yes	Yes	Yes	News	[14]
Lexicon	Yes	Yes	No	Enron Email Corpus	[15]
NB, SVM	Yes	No	Yes	Movie reviews	[16]
Corpus based	Yes	Yes	No	Blogs data	[17]
Apriori, NB	Yes	Yes	No	Movie reviews	[18]
Lexicon, ML	Yes	Yes	Yes	Twitter	[19]

The dare of current area under research is to mine information from unstructured records. The reviews hold point of view conveyed in natural language, recognizable to humans but not interpretable by machines [20]. It's great deal for customers nowadays to go through many online user reviews in order to make an informed preference for a product. Information is present in bulk that needs to be transformed into meaningful content. Opinions by the customers on different products are a valuable source of information for progress of ecommerce business that couldn't be ignored by the researchers. Thus, it's the demand of current age to develop efficient procedures which will also save time. It's a big challenge to accumulate the reviews of everyone's interest manually. The study concentrates on sentiment classification techniques among which lexicon-based approach is been implemented. It's a classification technique with its three basic levels, for subjective analysis of text data in bulk. Every chunk of text is designated as positive or negative or neutral. The methods applied for lexicon may vary consistent with the context. It demands that semantic orientation of every individual term must be known for complete semantic orientation of the text [21]. We considered using Lexicon-based approach as this method has high precision as compared to Machine Learning approach where large amount of data is required with a doubt in the training set and thus in the results.

Lexicon-based dictionary approach uses already defined words of dictionary where every word in the list is assigned a particular sentiment polarity and its value. Technique is described by Sharma et. al. used in dictionary-based method [22]. It's considered an efficient approach with advantage of reducing the data learning process as it already contains pre-defined list of words with their orientation.Sentiment analysis [23] is performed on user provided opinions within tweets of Twitter captured through Standard Search API that searches tweets based on keywords for hybrid cars in real-time. Text pre-processing is performed

in steps for cleaning of data. Next phase deals with the concept extraction by comparing polarities loaded in our database with SenticNet 4 of 50,000 concepts resulting the polarity sense. Results of each hybrid model are analyzed statistically to obtain top most trending one. The system of recommendation is for proposing the trending hybrid cars with specified technology. Automobiles with hybrid proficiency are enhancing the fuel efficiency provided by lessening the need of oil from transportation sector. Future trend predicts that almost 60% of the population will shift to cities so it's considered an important aspect to be reviewed for technology improvement [24].

2. Materials and methods

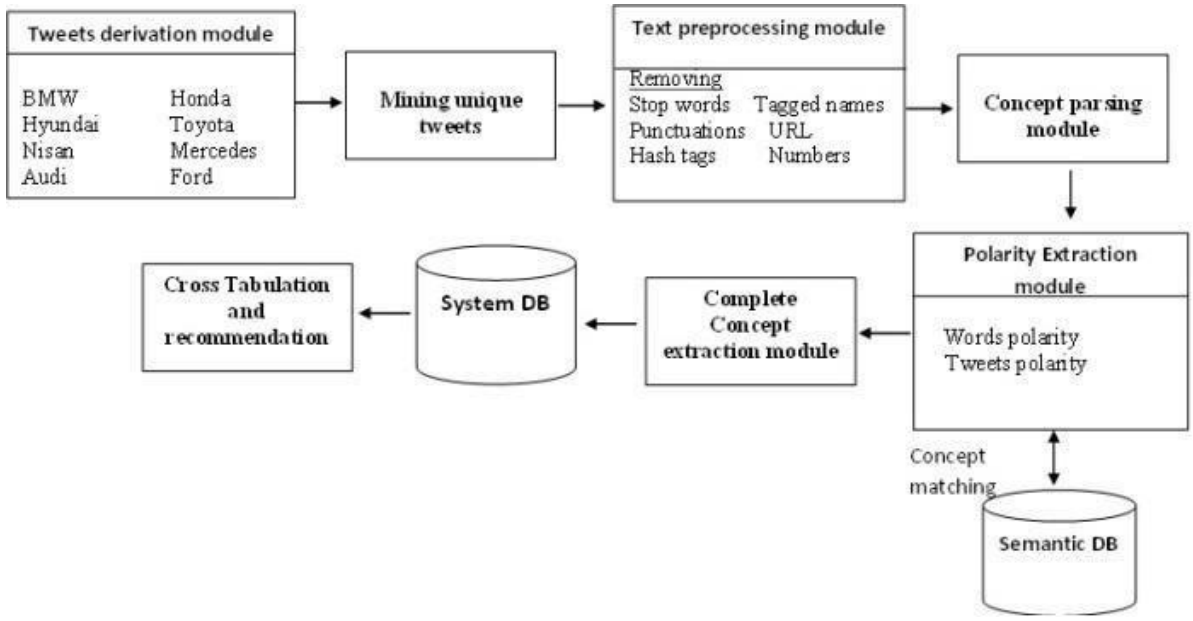
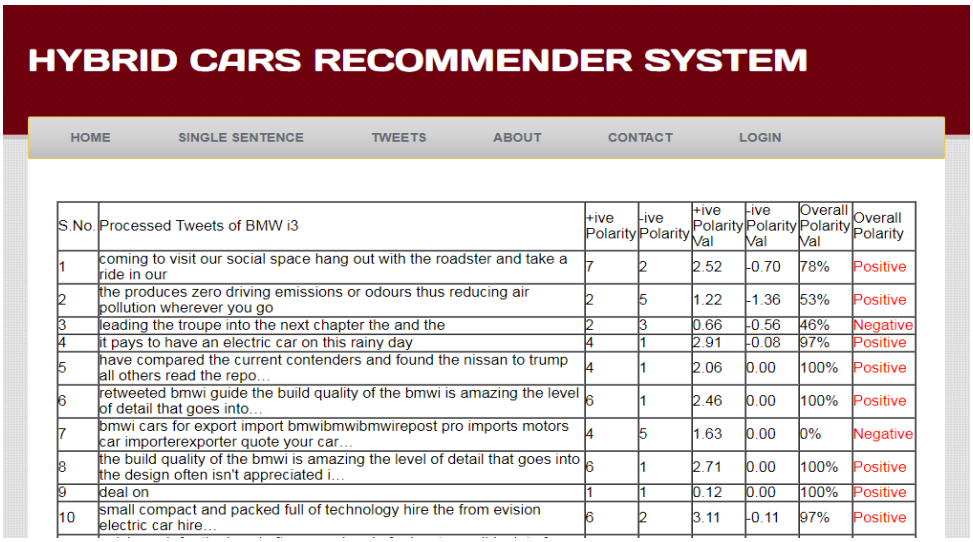


Fig. 1 Hybrid cars recommender system (HCRS)Module

To execute the script twitter authentication tokens (OAUTH) are required which are captured through twitter app. Our system will always grab fresh data from the social site (Twitter) in real time. API is provided with keywords of hybrid model names to capture tweets that only grab last seven day's tweets in every run which ensures no data redundancy. Script is directed with required tweets attributes i.e. ID and tweet text only. Every time the system will run, data will be unique and fresh. This feature will allow comparing the latest trends with previous ones obtained so far.Total 8 popular hybrid brands have been selected in this research with almost 4 models of each brand. This all together makes huge data set of 32 cars approximately. Data needs to be pre-processed efficiently when applying Natural Language Processing [25] [26] [27]. Our system is efficiently pre-processing the data going through elimination process of stop words, punctuations, hash tags, URLs, numbers, and names of users. Data is then divided into chunks of words for matching with SenticNet database concepts. Entire data is matched with the available list of 50,000 concepts [28] of SenticNet and polarity of each word is figured out.

Proper fraction is applied to determine what major part of sentiment lies within a tweet. Since the sum of positive concepts polarity from SenticNet database will always be greater than negative due to polarity value range (-1 to +1), thus system don't consider just the obtained highest polarity as in previous works

[29]. Larger the dataset, more the probability of accuracy within results, so maximum number of obtained sentic words of each tweet is targeted; either positive or negative, and so their respective polarity is picked. This polarity is divided by the total polarity of the tweet to obtain the fractional part of sentiment that resides within a tweet.Below figure 2 is a screenshot of our system. Tweets of a hybrid car after being processed and with respective polarities are illustrated.



S.No.	Processed Tweets of BMW i3	+ive Polarity	-ive Polarity	+ive Polarity Val	-ive Polarity Val	Overall Polarity Val	Overall Polarity
1	coming to visit our social space hang out with the roadster and take a ride in our	7	2	2.52	-0.70	78%	Positive
2	the produces zero driving emissions or odours thus reducing air pollution wherever you go	2	5	1.22	-1.36	53%	Positive
3	leading the troupe into the next chapter the and the	2	3	0.66	-0.56	46%	Negative
4	it pays to have an electric car on this rainy day	4	1	2.91	-0.08	97%	Positive
5	have compared the current contenders and found the nissan to trump all others read the repo...	4	1	2.06	0.00	100%	Positive
6	retweeted bmwi guide the build quality of the bmwi is amazing the level of detail that goes into...	6	1	2.46	0.00	100%	Positive
7	bmwi cars for export import bmwibmwibmwirepost pro imports motors car importerexporter quote your car...	4	5	1.63	0.00	0%	Negative
8	the build quality of the bmwi is amazing the level of detail that goes into the design often isn't appreciated i...	6	1	2.71	0.00	100%	Positive
9	deal on	1	1	0.12	0.00	100%	Positive
10	small compact and packed full of technology hire the from evision electric car hire...	6	2	3.11	-0.11	97%	Positive

Fig. 2 Fetched processed tweets with sentiment ratio

Mathematically it is expressed as:

$$D = P + abs(-N) \qquad (1)$$

$$Z = (X / D) * 100 \qquad (2)$$

Where Z is the resulting value after proper fraction, X is the tweet polarity against maximum word count; D is the sum of overall polarity, P and N denotes positive and negative polarities respectively. Absolute value of N is considered. Resulting ratio of the proper fraction will be categorized according to the range defined. If results against each tweet obtained are above 50%, it is declared positive. Below 50% are declared negative while equal is neutral. Final phase deals with the overall polarity calculation of all tweets against a single hybrid car model generating its sentiment ratio. Same process will be applied against each brand's tweets. Ratio of every hybrid model will be compared for final recommendation. In figure 2, 2nd and 3rd tweets resulted with more negative words with probability of negative outcomes, but results are opposite due to proper fraction mining sentiments precisely. We consider this method efficient as the results proved 80% accuracy.

Let T representing Twitter, hybrid cars as keyword q, number of tweets t, unwanted character's list as c, list of stop words s, hashtags h, tagged names n, URL's as u, and digits d, SenticNet database dictionary SEN_D, polarity sense POL_S, polarity POL, positive polarity POS_P, negative polarity NEG_P, positive polarity value POS_V, negative polarity value NEG_V, total polarity TOTAL_P, total positive and negative words of all tweets of a car as POS_W and NEG_W, total positive and negative polarities of all tweets of a car POS_POL and NEG_POL.

1. Start
2. Search data array against q from T
3. Repeat step 4 while t = 300
4. Filter the array for unique text
    a. Remove c
    b. Remove s
    c. Remove h
    d. Remove n
    e. Remove u
    f. Remove d
5. Explode t into chunks k
6. Repeat step 7 to 10 while t = 300
7. Match words where k = SEN_D
    a. If POL_S = positive
    b. POS_P += 1;
    c. else if POL_S = negative
    d. NEG_P +=1;
    e. If POL > 0
    f. POS_V ++;
    g. Else NEG_V ++;
8. POS_V + abs (-NEG_V) = D
    a. If POS_P == 0 && abs(-NEG_P) == 0
    b. Result R = undefined
    c. Else if POS_P > abs(-NEG_P)
    d. R = POS_V
    e. Else R = NEG_V
9. If D > 0
10. round(R/D)*100 = TOTAL_P
    a. If TOTAL_P > 50
    b. POL_S = positive
    c. Else if TOTAL_P < 50
    d. POL_S = negative
    e. Else POL_S = neutral
11. POS_W = POS_W + POS_P

12. NEG_W = NEG_W + NEG_P
13. POS_POL = POS_POL + POS_V
14. NEG_POL = NEG_POL + NEG_V
15. POS_POL + abs (-NEG_POL) = D
    a. If POS_W == 0 && NEG_W == 0
    b. R = undefined
    c. Else if POS_W > NEG_W
    d. R = POS_POL
    e. Else NEG_POL
16. If D > 0
17. round(R/D)*100 = TOTAL_P
    a. If TOTAL_P > 50
    b. POL_S = positive
    c. Else if TOTAL_P < 50
    d. POL_S = negative
    e. Else POL_S = neutral
18. Stop

Consider first tweet in Fig 2 that shows total positive words count is greater than negative words count hence tweet's positive polarity value will be picked for proper fraction that is 2.52. From equation (1)

$$D = P + abs(-N)$$

$$2.52 + 0.70 = 3.22$$

Putting the values in equation (2)

$$Z = (2.52 / 3.22) * 100$$

$$Z = 78\% > \text{Range } 50 \text{ hence positive}$$

3. Results and discussion

Almost 300 tweets are scrapped from twitter against each hybrid model. Data is processed in Statistical Package for the Social Science (SPSS) tool for statistical analysis. Tweets polarity value against each model is accumulated and summarized by cross tabulation method. Larger the dataset, more the probability of accuracy within results [30] hence models with larger no. of tweets are extracted from the list. Table 2 shows Brand models with higher number of tweets extracted for classification, ranging from 80 to 100 tweets against each brand.

Table 2: Models with maximum number of tweets extracted from the list

Brand Model	Max. No. of Tweets	Percentage of Positivity
Audi A6 hybrid	100	73.00%
Audi A4 hybrid	99	72.00%
Audi Q5 hybrid	99	72.00%
Ford escape	99	51.00%
Toyota Prius V	97	75.00%
BMWi8	97	76.00%

Highest percentage resulted models are then selected and classified by cross tabulation where results in Figure 3 shows that BMWi8 have the higher value up till now among positive polarities. As the procedure is implemented on real time data, rationalized results are obtained each time data is fetched for recommendations which is the main characteristic of a recommender system.

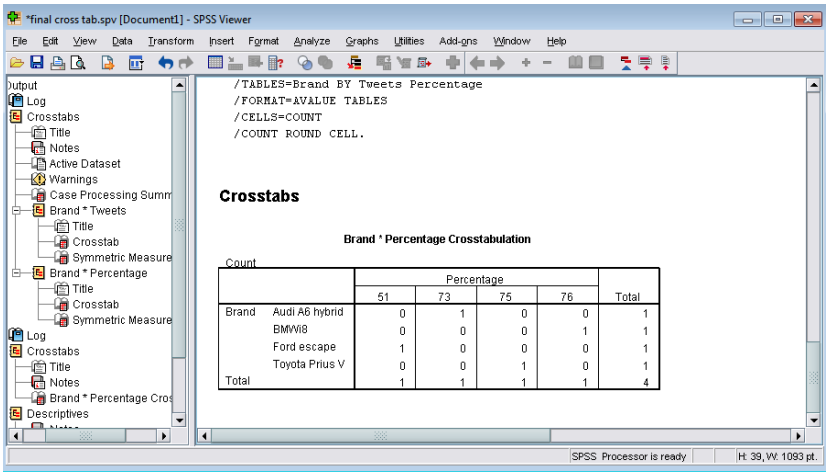


Fig. 3 Highest positive polarities

4. Conclusion

An automated system has been proposed for decision of trending hybrid car brand based on user's views and opinions on Twitter. Since efficient systems yet don't reflect future trends in hybrid technology, this system could be a great support in this sector to be offered on the network. Our system scraps for about 300 tweets for each model that are compared for sentics and semantics. Scrapped data is analyzed and compared with 50,000 natural language concepts embedded within our system. Positive and negative polarities are figured out against tweets of each model calculated by proper fraction. Models with higher number of tweets are been selected for analysis as accuracy demands greater data size. Recommendation of best hybrid brand is provided based on highest positive polarity percentage after analyzing statistically.As the procedure is implemented on real time data, rationalized results are obtained each time data is fetched for recommendations. Descriptive statistics up till now showed that BMWi8 got more number of tweets and polarity percentage as compared to rest of the brands.Text of the tweets may repeat or there might be symbols in the text that could slow down the processing. Function is applied for array

filtration but there is possibility of text repetition. Negations are ignored for now but it will be considered in the future.

As accuracy demands more and more data, in future different social sites will be targeted to include more user opinions. As web is not free of spammers hence data needs to be purified for better results. For high precision and accuracy, data will be passed through spam filtration.

5. Acknowledgements

The authors are grateful to The University of Lahore, Sargodha campus for funding to present this article on 1st international conference on Software Engineering and Computing Discipline hosted by University of Sawabi, KPK Pakistan held on 12-14 Nov 2019

6. References

1. Biancucci, Michele, Jordan Janeiro, Massimo Mecella, and Stephan Lukosch. "Supporting industrial processes by monitoring and visualizing collaborations." In Proceedings of the 2014 InternationalConference on Innovative Design and Manufacturing (ICIDM), pp. 330-335. IEEE, 2014.
2. Khurshid, Sadaf, Sharifullah Khan, and Shariq Bashir. "Text-based Intelligent Content Filtering on Social Platforms." In 2014 12th International Conference on Frontiers of Information Technology, pp. 232-237. IEEE, 2014.
3. Chen, Junnan, Courtney Miller, and Gaby G. Dagher. "Product recommendation system for small online retailers using association rules mining." In Proceedings of the 2014 International Conference on Innovative Design and Manufacturing (ICIDM), pp. 71-77. IEEE, 2014.
4. Sajid, Anamta, Sadaqat Jan, and Ibrar A. Shah. "Automatic Topic Modeling for Single Document Short Texts." In 2017 International Conference on Frontiers of Information Technology (FIT), pp. 70-75. IEEE, 2017.
5. Aziz, Mehwish, and Muhammad Rafi. "Identifying influential bloggers using blogs semantics." In Proceedings of the 8th International Conference on Frontiers of Information Technology, p. 7. ACM, 2010.
6. Baharudin, Bharum. "Sentence based sentiment classification from online customer reviews." In Proceedings of the 8th International Conference on Frontiers of Information Technology, p. 25. ACM, 2010.
7. Ali, Abbas Raza, and Maliha Ijaz. "Urdu text classification." In Proceedings of the 7th international conference on frontiers of information technology, p. 21. ACM, 2009.
8. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." In Proceedings of the ACL-02 conference on Empiricalmethods in natural language processing-Volume 10, pp. 79-86. Association for Computational Linguistics, 2002.

9. Dave, Kushal, Steve Lawrence, and David M. Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews." In Proceedings of the 12th international conference on WorldWide Web, pp. 519-528. ACM, 2003.
10. Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177. ACM, 2004.
11. Zhang, Wenhao, Hua Xu, and Wei Wan. "Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis." Expert Systems with Applications 39, no. 11 (2012): 10283-10291.
12. Khan, Aurangzeb, Baharum Baharudin, and Khairullah Khan. "Sentiment classification from online customer reviews using lexical contextual sentence structure." In International Conference on Software Engineering and Computer Systems, pp. 317-331. Springer, Berlin, Heidelberg, 2011.
13. Fang, Ji, and Bi Chen. "Incorporating lexicon knowledge into SVM learning to improve sentiment classification." U.S. Patent 8,352,405, issued January 8, 2013.
14. Moreo, Alejandro, M. Romero, J. L. Castro, and Jose Manuel Zurita. "Lexicon-based comments-oriented news sentiment analyzer system." Expert Systems with Applications 39, no. 10 (2012): 9166-9180.
15. Mohammad, Saif M. "From once upon a time to happily ever after: Tracking emotions in mail and books." Decision Support Systems 53, no. 4 (2012): 730-741.
16. Rui, Huaxia, Yizao Liu, and Andrew Whinston. "Whose and what chatter matters? The effect of tweets on movie sales." Decision Support Systems 55, no. 4 (2013): 863-870.
17. Keshtkar, Fazel, and Diana Inkpen. "A bootstrapping method for extracting paraphrases of emotion expressions from texts." Computational Intelligence 29, no. 3 (2013): 417-435.
18. Heerschop, Bas, Frank Goossen, Alexander Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. "Polarity analysis of texts using discourse structure." In Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 1061-1070. ACM, 2011.
19. Zhang, Lei, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. "Combining lexicon-based and learning-based methods for Twitter sentiment analysis." HP Laboratories, Technical Report HPL-2011 89 (2011).
20. Bucur, Cristian. "Using opinion mining techniques in tourism." Procedia economics and finance 23 (2015): 1666-1673.
21. Kaur, Amandeep, and Vishal Gupta. "A survey on sentiment analysis and opinion mining techniques." Journal of Emerging Technologies in Web Intelligence 5, no. 4 (2013): 367-371.
22. Sharma, Richa, Shweta Nigam, and Rekha Jain. "Supervised opinion mining techniques: a survey." The Proceedings of the International Journal in Foundations of Computer Science & Technology (IJFCST) 4, no. 3 (2014).
23. Jebaseeli, A. Nisha, and E. Kirubakaran. "A survey on sentiment analysis of (product) reviews." International Journal of Computer Applications 47, no. 11 (2012).
24. Phillips, Judith, and Shauna McGee. "Future ageing populations and policy." In Geographies of Transport and Ageing, pp. 227-250. Palgrave Macmillan, Cham, 2018.

25. Cambria, Erik, Daniel Olsher, and DheerajRajagopal. "SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis." In Twenty-eighth AAAI conference on artificial intelligence. 2014.
26. Cambria, Erik, BjörnSchuller, Bing Liu, Haixun Wang, and Catherine Havasi. "Knowledge-based approaches to concept-level sentiment analysis." IEEE intelligent systems 28, no. 2 (2013): 12-14.
27. Cambria, Erik. "Affective computing and sentiment analysis." IEEE Intelligent Systems 31, no. 2 (2016):102-107.
28. Cambria, Erik, Robert Speer, Catherine Havasi, and Amir Hussain. "Senticnet: A publicly available semantic resource for opinion mining." In 2010 AAAI Fall Symposium Series. 2010.
29. Baradwaj, Brijesh Kumar, and Saurabh Pal. "Mining educational data to analyze students' performance." arXiv preprint arXiv:1201.3417 (2012).
30. Cambria, Erik, Catherine Havasi, and Amir Hussain. "SenticNet 2: A semantic and affective resource for opinion mining and sentiment analysis." In Twenty-Fifth International FLAIRS Conference. 2012.