International Journal of Research in Engineering, Science and Management
Volume-2, Issue-8, August-2019
www.ijresm.com | ISSN (Online): 2581-5792

Analysis of the Hotel Reviews using Opinion Mining and Machine Learning Concept

Raj Kumar Saini[1], Prakash Dangi[2]

[1]*Student, Department of Computer Science and Engineering, Modern Institute of Technology & Research Centre, Alwar, India*

[2]*Assistant Professor, Department of Computer Science and Engineering, Modern Institute of Technology & Research Centre, Alwar, India*

***Abstract*: Analyzing the importance of internet is really important these days for the businesses. The concept of monitoring opinions of the customers and understanding their needs require a lot of data mining and its overall effects. This paper deals with the inclusion of the concept of the Machine Learning. Because of new computing technologies, machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence wanted to see if computers could learn from data. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results. It's a science that's not new – but one that has gained fresh momentum.**

***Keywords*: Opinion Mining, Sentiment Analysis, Machine Learning, Data Mining, Data Science.**

1. Introduction

In the broadest terms, opinion mining is the science of using text analysis to understand the drivers behind public sentiment. All text is inherently minable. As such, while social media may be an obvious source of current opinion, reviews, call center transcripts, web pages, online forums and survey responses can all prove equally useful.

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. A sentiment analysis system for text analysis combines natural language processing (NLP) and machine learning techniques to assign weighted sentiment scores to the entities, topics, themes and categories within a sentence or phrase.

Whereas sentiment analysis – a predecessor to the field of opinion mining – examines how people feel about a given topic (be it positive or negative), opinion mining goes a level deeper, to understand the drivers behind why people feel the way they do.

Individual opinions are often reflective of a broader reality. A single customer who takes issue with a new product's design on social media likely speaks for many others. The same goes for a member of the public who takes to a political campaigner's web page to praise or criticize the policies proposed.

Gather enough opinions – and analyze them correctly – and you've got an accurate gauge of the feelings of the silent majority. This relates not only to how people feel, but the drivers underlying why they feel the way they do.

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that which makes it more similar to humans: The ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.

2. Literature survey

Sentiment analysis has three levels of analysis; document a level, sentence level, and aspect level [1]. Document level sentiment analysis analyzes a whole document and classifies polarity for the whole document. Sentiment analysis in sentence level providing polarity for each sentence. Aspectbased sentiment analysis identify the aspects of the given target entities and sentiment expressed for each aspect.

Qiu et al., [2] suggested double propagation method for aspect-based sentiment analysis. This method extracts the aspect expression and the sentiment by using syntax relation between opinion words and target to do propagation in order to find the other targets and opinion words. Propagation is used to extract aspect expression or target and also expand opinion words dictionary.

Yauris and Khodra [3] also employed double propagation to develop aspect based summarization system for game review. This system extracted aspect and its sentiment, aggregated it with aspect categorization, and presented it in structured summary. From their experiment, the best performance achieved by using Zagal opinion lexicon, opinion word dictionary for validation, and using clause pruning without global pruning. [3]

Chung and Tseng [4] proposed a framework to design business intelligence (BI) system to extract the relation between customers rating with its review. The system used two data mining methods to extract the decision setting automatically, therefore it can be a manual to understand the relation between rating and the review. In designing BI system, there are some steps needed to be accomplished. The first step is feature extraction, which means extracting every feature consisted by textual terms showed up in the review. These features are filtered because not all terms are necessarily taken, except some important terms. This term selection process is conducted using TF-IDF. The next step is finding the decision rules from review. The method used in this research is association rule mining method and rough set theory (RST).

Das [5] proposed business intelligence system to analyze consumer opinions obtained from Twitter. Opinion analysis is conducted by OLAP and data cubes to form three dimension data (product name, date, polarity). The system counted tweet frequency for each product on particular day to generate report.

Lexicon-based method is to formulate a series of sentiment dictionaries and rules, to split and parse paragraphs, to find out the emotional words, negative words and degree adverbs in the document, and finally calculate emotional values to determine the emotional tendencies of the text [6].

Machine learning-based method is to use statistical learning from marked corpus to extract features which can effectively express the type characteristics of comments and form the classification model by using those features [7]. In order to translate the corpus into computer understandable form, the corpus needs to be transferred into corresponding vector representation [8]. This paper introduced an open source toolkit named Word2vec which was developed by Google in2013. It was used for translating the corpus into word vectors [9]. Word vectors are used as training features for text classification by using machine learning algorithms such as Support Vector Machine (SVlv1), Random Forest, and so on.

Many research works are done on Sentiment Analysis of English. Cui et al. [10] worked on online product reviews. They classified the reviews to two major classes: positive and negative. They considered around 100k product reviews from different websites. Jagtap et al. [11] applied Support Vector Machine (SVM) and Hidden Markov Model (HMM). Their hybrid classification model to extract the sentiment of teacher feedback assessment performed well. Alm et al. [12] Separated seven emotional words to three polar classes of positive emotional, negative emotional and neutral. They used Winnow parameter tuning approach and got 63% accuracy. Agarwal et al. [13] applied unigram, tree model and feature based model to extract twitter sentiment. Unigram model is outperformed by tree model and feature based model. The accuracy they got is around 61%. Zou et al. [14] introduced a model of learning bilingual word embeddings from a large and unlabeled dataset. They showed that their model outperforms baselines in semantic similarity of words. Turian et al. [15] worked on Brown clusters, embeddings of Collobert and Weston (2008) and hierarchical log-bilinear embeddings. Chen et al. [16] proposed some approaches that can differentiate the released word embeddings models. They showed that embeddings can detect surprising semantics of the sentences even without having the structure. Tang et al. [17] introduced a technique of gathering both contextual and sentiment information of the words by learning Sentiment-Specific Word Embedding. They applied their model to extract twitter sentiment. The accuracy they got is around 83%. Levy et al. [18] Worked on skip-gram model with negative sampling of Mikolov et. al. (2013) and generalized it. They extracted dependency based contexts and showed that they produce different types of similarities. Andreas et al. [19] showed three possible benefits of word embeddings: Vocabulary expansion, Statistic sharing and embedding structure. Lebret et al. [20] worked on Word Embeddings with Hellinger PCA. They constructed word co-occurrence matrix to find the contextual representation of words. They obtained around 89% accuracy. Levy et al. [21] applied word embedding models with neural network to determine word similarity and detect analogy. They achieved better results than the traditional count based distributional models.

3. Proposed work

The algorithm deals with the opinion analysis posted by the customer over the online platform of about any product or the services. The simplified model for such review by three random users is been shown here.

Table 1
Review Keyword Analysis

Users	Positive Keyword	Negative Keywords	Neutral Keywords	Total Words in Review
User 1	27	7	3	100
User 2	7	35	12	98
User 3	42	13	3	140
Total	76	55	18	338

For the User 1 the entered review is positive or negative that entirely depends over the Naïve Bayes classifier tool and also the formation of the table process. In the proposed thesis work the above table is constructed by the process explained below.

Step 1: Reviews are been saved in the database format, in the current algorithm this has been done using the string comparison and categorization. The datasets of the positive keywords, negative keyword and neutral keywords.

Step 2: Match the keywords with the datasets values and keep updating the count flag for each of the keyword or the word until all the words are looped through and are been converted into the numerical metrics.

To solve the above generated table by the above explained two steps; algorithm will move forward towards the Naïve Bayes Solver developed for the current algorithm.

Step 3: Calculate the conditional probability for user 1, user 2 and user 3.

Step 4: Calculate dependent probability for Positive Review of User 1.

$$p(Positive\ Review|User1) = \frac{p(User\ 1|Positive\ Keywords).p(User\ 1_{positive})}{p(positve\ keywords)}$$

Step 5: Calculate dependent probability for Negative Review of User 1.

$$p(Negative\ Review|User1) = \frac{p(User\ 1|Negative\ Keywords).p(User\ 1_{negative})}{p(negative\ keywords)}$$

Step 6: Calculate dependent probability for Neutral Review of User 1.

$$p(Neutral\ Review|User1) = \frac{p(User\ 1|Neutral\ Keywords).p(User\ 1_{neutral})}{p(neutral\ keywords)}$$

Step 7: After the calculation find out the highest probability ratio which will indicate the Review Polarity, as in the User 1's case the review is positive one.

Step 8: Calculate dependent probability for Positive Review of User 2.

Step 9: Calculate dependent probability for Negative Review of User 2.

Step 10: Calculate dependent probability for Neutral Review of User 2.

Step 11: After the calculation find out the highest probability ratio which will indicate the Review Polarity, as in the User 2's case the review probability score is pretty much on the confusing so as the distance is much lesser between the negative and the neutral review than the positive and the neutral review, review will be considered as the negative review.

Step 12: Calculate dependent probability for Positive Review of User 3.

Step 13: Calculate dependent probability for Negative Review of User 3.

Step 14: Calculate dependent probability for Neutral Review of User 3.

Step 15: After the calculation find out the highest probability ratio which will indicate the Review Polarity, as in the User 3's case the review probability score is pretty much clear and the highest score is with positive verdict hence the User 3's review will come under the positive review category.

4. Results

The above presented algorithm generates verdict after considering the complete test environment and hence is connected to the Machine Learning Approach as well. The below presented graphs shows the accuracy rate generated by the previous algorithms and the above presented algorithm.

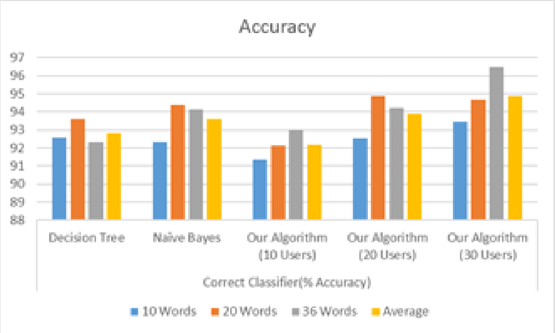

Fig. 1. Accuracy graphs

5. Conclusion

Inspired by the increasing hotel information on the Internet, we investigated customer satisfaction and the impact of different hotel aspects on customer satisfaction. This is done through innovative text mining techniques for topic modeling, which allows us to form an overall picture of these dependencies without any preconceptions about the meaning of these important aspects.

Forming such a holistic picture allows us to create information systems and management programs that systematically collect, monitor and analyze key online customer reviews and social media posts, and estimate how these key changes will lead to overall customer satisfaction or other aspects. Controlling these aspects will help the hotel build its brand and customers at a low cost. Hotel managers can understand how customers respond to specific aspects of the hotel and identify areas where they are underperforming. The system from here can develop itself for dynamic database technique for self-evolving and learning, so that this system will never get outdated and will keep on analyzing the reviews and will predict its verdict accurately.

References

[1] B. Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, 2012.

[2] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion Word Expansion and Target Extraction through Double Propagation," Comput. Linguist., vol. 37, no. September 2009, pp. 9–27, 2011.

[3] K. Yauris and M. L. Kodra, "Aspect-based Summarization for Game Review Using Double Propagation", Advanced Informatics, Concepts, Theory, and Applications (ICAICTA), 2017 International Conference on. IEEE, 2017.

[4] W. Chun and T.L. Tseng, "Discovering business intelligence from online product reviews: A rule-induction framework", Expert Syst. Appl., 2012.

[5] T. K. Das, "Business Intelligence through Opinion Mining. Handbook of Research on Advanced Data Mining Techniques and Applications for Business Intelligence", IGI Global, 2017

[6] W. Cheng, Y. Song, Y. Zhu and P. Jian, "Dimensional Sentiment Analysis for Chinese words Based on synonym lexicon and Word Embedding," 2016 International Conference on Asian Language Processing (IALP), Tainan, 2016, pp. 312-316.

[7] Z. Fan, L. Su, X. Liu and S. Wang, "Multi -label Chinese question classification based on word2vec," 2017 4th International Conference on Systems and Informatics (ICSAI), Hangzhou, 2017, pp. 546-550.

[8] N. Chirawichitchai, "Emotion classification of Thai text based using term weighting and machine learning techniques," 2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE), Chon Buri, 2014, pp. 91-96.

[9] A. H. Ombabi, o. Lazzez, W. Ouarda and A. M. Alimi, "Deep learning framework based on Word2Vec and CNN for users' interests classification," 2017 Sudan Conference on Computer Science and Information Technology (SCCSIT), Elnihood, 2017, pp- 1-7.

[10] Hang Cui, Vibhu Mittal and Mayur Datar, "Comparative Experiments on Sentiment Classification for Online Product Reviews," Proceedings of the 21st National Conference on Artificial Intelligence, AAAI, Boston, MA, 2006.

[11] Balaji Jagtap and Virendrakumar Dhotre, "SVM and HMM Based Hybrid Approach of Sentiment Analysis for Teacher Feedback Assessment," International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 3, May-June 2014.

[12] C. Alm and D. Roth and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," EMNLP, 2005.

[13] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau, “Sentiment Analysis of Twitter Data,” LSM '11 Proceedings of the Workshop on Languages in Social Media, Pages 30-38, 2011.

[14] Will Y. Zou, Richard Socher, Daniel Cer and Christopher D. Manning, “Bilingual Word Embeddings for Phrase-Based Machine Translation,” Sem Eval, 2012.

[15] Joseph Turian, Lev Ratinov and Yoshua Bengio, “Word representations: A simple and general method for semi-supervised learning,” Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 384–394, Uppsala, Sweden, 11-16 July 2010.

[16] Yanqing Chen, Bryan Perozzi, Rami Al-Rfou, and Steven Skiena, “The Expressive Power of Word Embeddings,” ICML 2013 Workshop on Deep Learning for Audio, Speech, and Language Processing, Atlanta, USA, June 2013.

[17] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu and Bing Qin, “Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification,” Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1555–1565, Baltimore, Maryland, USA, June 23-25, 2014.

[18] Omer Levy and Yoav Goldberg, “Dependency-Based Word Embeddings,” Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 302–308, Baltimore, Maryland, USA, June 23-25, 2014.

[19] Jacob Andreas and Dan Klein, “How much do word embeddings encode about syntax?,” Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), June 2014.

[20] Remi Lebret and Ronan Collobert, “Word Embeddings through Hellinger PCA,” Idiap Research Institute, 2013.

[21] Omer Levy, Yoav Goldberg and Ido Dagan, “Improving Distributional Similarity with Lessons Learned from Word Embeddings,” Transactions of the Association for Computational Linguistics, vol. 3, pp. 211–225, 2015.