



Comparative Study of Opinion Mining and Sentiment Analysis: Algorithms and Applications

Rana Zuhair Alobaidy¹, Ghaydaa Abdulaziz Altalib² and Zainab S. Attarbashi³

^{1,2} Computer Science Department, College of Computer Sciences and Mathematics, University of Mosul, Iraq.

³ School of Computing, Universiti Utara Malaysia, 06010 Kedah, Malaysia

¹rza2451990@gmail.com, ²ghaydabdulaziz@uomosul.edu.iq, ³zainab.senan@uum.edu.my

Abstract

The massive amount of data available online increases the ability to analyze and understand how people are thinking. The internet revolution has added billions of customer's review data in its depots. This has given an interest in sentiment analysis and opinion mining in the recent years. People have to depend on machines to classify and process the data as there are terabytes of review data in stock of a single product. So that prediction customer sentiments is very important to analyze the reviews as it not only helps in increasing profits but also goes a long way in improving and bringing out better products. In this paper, we present a survey regarding the presently available techniques and applications that appear in the field of opinion mining, such as: economy, security, marketing, spam detection, decision making, and elections expectation. The survey is based on the techniques used with English-written data however it is important for future studies on other languages like Arabic and Malay.

Keywords: Data Mining, Social Media, SVM.

1. Introduction

Humans are biased creatures and the opinions are very important to them; therefore sentiment analysis(emotion) aim to build a system that analyze what an individual want from a product, topic, or event. People express their opinions in review, posting, comment, or tweet texts. Mining in available user opinions is difficult but very useful. Therefore, sentiment analysis has done on three levels:

1. *Document level*: it reduces whole document into one opinion, but on most cases, it doesn't represent one opinion, perhaps, one document has been inconsistent and can contain different opinions for the same topic.
2. *Sentence level*: it classifies emotions for each sentence in a document. First task is to classify each sentence into objective or subjective.
3. *Attribute level*: it aims to classify different attributes for the same topic. The first task is recognition of different attributes and extracting the opinions for them

This research highlights on studying of human opinions to know what people think about any topic or event by looking on their sentiments and emotions in their texts. Therefore, sentiment analysis is field of natural language processing (NLP).

Social media is one of the best ways for getting opinions; because it is open area for expressing opinions freely. Opinion mining is a branch of web content mining, and the latter is branch of data mining, figure (1): shows opinion mining in the structure of data mining. There are another analysis's in this topic, these analysis's may appear in another researches in social media, they can be under another names:

- Subjectivity Analysis.
- Review Mining.
- Appraisal Extraction.

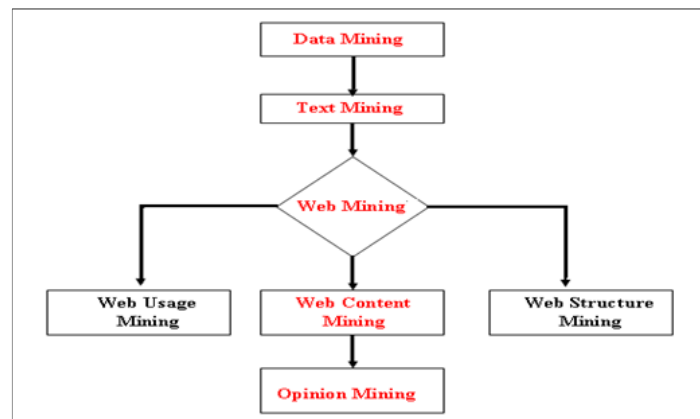


Figure 1. structure of data mining [20]

Components of opinion mining have shown in figure (2) and containing: 1) Object 2) Opinion Holder 3) Opinion.

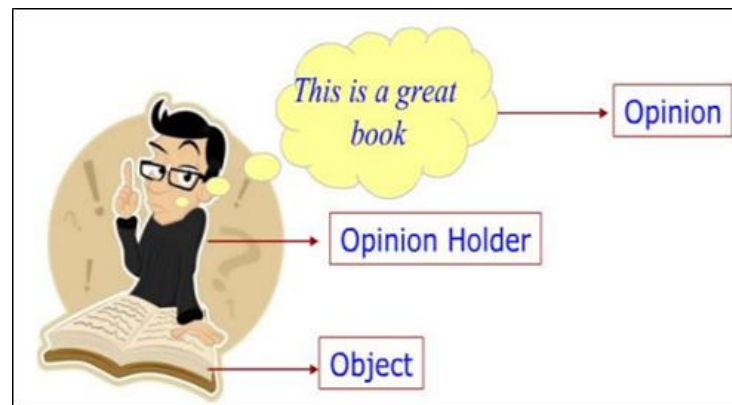


Figure 2. Components of opinion mining [11]

2. Data Resources

Opinions are presented by people. Therefore, the companies evaluate opinions by data studying that are written in blogs, review websites, in addition to, web discourse and news articles.

2.1 Reviews

The users who generated reviews available on the web that helps a customer to buy a product. Sites of E-commerce such as www.amazon.in, www.flipkart.com and www.reviewcentre have millions of customer reviews for products, whereas www.rediff.com/movies/reviews, www.indiaglitz.com and www.rottentomatoes.com have movies reviews and www.yelp.com, www.burrrp.com have reviews for restaurants.

2.2 Web Discourse

A user records its opinions in blog which is a personal website, it links to other sites. Blog is simple and no daft post writing approach that uploading it on the internet made the blogs a rich source of data for sentiment mining. Twitter is micro blogging site and overfilled with opinions that are resolute in determining the results of election. Opinions, feelings and emotions are expressions which people record them daily as events on blog or on twitter.

2.3 News Articles

There are many web sites have news articles like *www.thesun.co.uk*, *www.cnn.com* and *www.thehindu.com* that allow readers to comment. This is good way to recording people opinions in important and relevant issues.

3. Applications of Opinion Mining and Sentiment Analysis

The main applications of Opinion mining and sentiment analysis are as given below [16]:

3.1 Purchasing Product or Service

The extraction, analysis and presentation of opinions and sentiments of the web to people in understandable manner help them to take right purchasing decision of a product or service. By this way, people don't need to external consultant because they can easily take evaluation of other's opinions and experience about service or product and comparing competing brands. By this technique, people save consulting expenses.

3.2 Quality Improvement of Product or service

The manufactures use opinions about product or service as feedback for making decision to improve the quality of them or not. Therefore, opinion mining and sentiment analysis use them to analyze consumers or customer's opinions online from website such as News websites to determine advantages and disadvantages of product or service, this process save amount of money which had spent to take customers or users opinions previously.

3.3 Marketing research

Marketing research is one of areas which uses sentiment analysis techniques to analyze consumers trends about particular products or services, it also use to determine success of ads campaign and make studying about "what does individual need from products or services that is not available in market, therefore, it is very important point, if it use in right way economically, it will be a big benefit for businesses.

3.4 Detection of "flame"

Opinion mining can be used to determine opinions and thoughts those use argumentativeness, annoying, bullying, over heating words or hatred language among individuals from same background or with different culture, race or country. Opinion mining can help to study different ideologies in one society and how these ideologies are been spread throughout the Internet, especially the social media platforms.

3.5 Opinion spam detection

Spam content is one of problems that need to solve, because anyone can put any content on web. People may put spam content to mislead people because internet is available to all. Therefore, opinion mining and sentiment analysis techniques can classify web content into "spam" and "not spam" contents.

3.6 Decision Making

Making decision is one of difficult processes in companies, originations and governments that need requirements. In opinion mining and sentiment analysis techniques use people's opinions and experiences about products, services or events which analyze to give good decision making. In this way, the decision makers partake people who are beneficiary side, such as "Government ask people's opinion about refugees receiving in their country", therefore, making decision may be effect on external policy.

3.7 Elections Expectation

Mining in opinions help in expectation "who is the winner in the election" by comparing people opinions and know answer the question "what are trending names on web pages especially social media", positive and negative opinions about candidates' achievement or through elective program.

4. Classification Methods

From table (1), illustrate three methods which used in sentiment analysis, they are:

- *Lexicon-based method*: used lexicon for explanation of text.
- *Machine learning*: used classifiers to classify opinions such as: SVM, Naïve Bayes, etc.
- *Hybrid Approach*: depends on two approaches (lexicon-based method + machine learning method).

Algorithms are using frequently in this area:

4.1 Support Vector Machine (SVM)

SVM is non-probabilistic linear classifier that use for classifying, regression and pattern recognition, it aims to recognize between two classes of training data by decision boundary. The idea of SVM is finding optimal hyper plane that uses to classify data, as in figure (3).

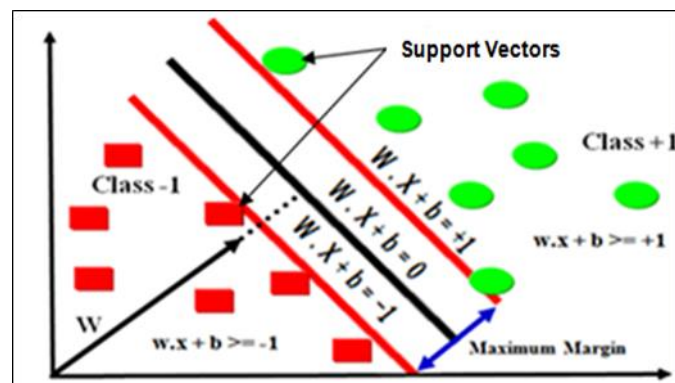


Figure 3. Optimal hyper plane for binary classification

One of the SVM algorithm properties is getting high accuracy in classifying data and it applies in text categorization, image recognition and medical application.

4.2 Naïve Bayes

One of supervised machine learning methods also called simple Bayes, idiot's Bayes and independence Bayes. It is very important for these reasons: naïve bayes is easy to construct, may be readily applied to large data sets and easy to interpret, so users unskilled in classifier technology can understand why it is making the classification it makes.

4.3 K-Nearest Neighbor (KNN)

It uses for classification and regression and also it is a non-parametric machine learning method. It works to searches for the pattern space for the k training samples that are closer to the unknown samples. The effectiveness of KKN depends upon applied value.

5. Evaluation and Description

Performance of algorithm evaluates depending on one or more of the following measurements, accuracy, precision, F1-measure and recall. This research depends on many accuracy metrics for getting the evaluation. Table (1) illustrate equation of each metric.

Table 1. illustrate measurements for classification results

Measurements	Equations
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$
F1- measure	$\frac{(2 * Precision * Recall)}{(Precision + Recall)}$
Recall	$\frac{TP}{TP + FN}$

Table (2) illustrate studies in opinion mining and sentiment analysis field using different techniques, it shows data resources, technique used, accuracy of results. The efficiency of algorithms is very difficult to estimate according to researchers' opinions. All the input data are written by Non-Arabic language.

Table 2. Evaluation Metrics of different algorithms

Author Name	Year	Technique used	Data source	Accuracy
Tan [25]	2008	SVM, Centroid classifier, KNN, Winnow	ChinSentiCorp	SVM : 90%
Prabowo [14]	2009	SVM and Hybrid	Movie review and MySpace comment	89%
Melville [12]	2009	Bayesian Classification	Blogs	91.21%
Rushdi Saleh M. [19]	2011	SVM	Blogs and Product reviews	91.51%
Khan [10]	2011	Naïve Bayes	Movie review, Hotel review, Airline and Airport review	86.6%
Xia [27]	2011	SVM, Naïve Bayes, Maximum Entropy	Amazon reviews	SVM: 86.4% NB : 85.8%
Zhang [28]	2011	Naïve Bayes	Cantonese reviews	93%
B. Chen [4]	2011	NB, KNN, Modified K Means + NB and K Means + NB + KNN	Mobile reviews	NB : 79.66 %
				KNN :83.59%
				MKM+NB 89%
				MKM + NB + KNN : 91%
Basari & Hussin Ananta [2]	2012	SVM and SVM-PSO	Movie reviews	SVM 71.87% SVM-PSO 77%
A. Shrivatava and B. Pant [23]	2012	SVM	facebook Comments	74.8268%
K. Saras Wathi , A.	2014	SVM with Polykernel,	Movie reviews	Poly Kernel 87%

Tamil Arasi [7]		SVM with RBF Kernel and Bagging kernel		RBF Kernel 73.33%
				Bagging Kernel 88%
Rastogi & Singhal Kumar [18]	2014	SVM, Lexicons + MPQA+SVM	Camera reviews	SVM 48.7% SVM 67.3%
Sharma & Nigam Jain [22]	2014	Document based opinion mining system	Movie reviews	63%
Preety & S. Dahiya [15]	2015	SVM, NB, NB+ MKM	Mobile reviews	NB 79.66%
				SVM 83.59%
				NB+MKM 98%
S. Kethavath [9]	2015	NB, Multinomial, NB DecisionTree, SVM	Tweets	NB 81.21% Multinomial NB 89.75% Decision tree 80.08% SVM 94.45%
Raj. S. Suresh [24]	2015	NB, SVM, Max Entropy, J48	Twitter	NB 76% SVM 88% Max E. 89.2% J48 92%
Bholane & Gore [3]	2016	SVM	Data Twitter	97.54%
Rocha, Pacheco & Mendonza [17]	2016	SVM	Weka Software and Movie DataBase	93.7%
Tanesab, Sembiring & Purnomo [26]	2017	SVM	Youtube Comments	84%
Aggarwal L. & Gupta [1]	2017	NB,KNN,MKM + NB and MKM+ NB+ KNN	Mobile reviews	NB 79.66% KNN 83.59% MKM + NB 89% MKM + NB + KNN 91%

6. Conclusion

Opinion mining and sentiment analysis is an domain of data mining which is applied to summarize the knowledge from huge amount of data such as people's tweets, comments and reviews on any product, topic or event etc. However, the rise of social media services such as: Twitter, Facebook and YouTube allowed people to express and share their reviews, comments, tweets, "what they like" and "what they dislike" about something they concern it in freely and openly way. Therefore, sentiment analysis emerged to benefit from written texts people's sentiment. This field is very important for all people not only commercial companies but any person can use these techniques if he thinks in business. Web has a large amount of thoughts and opinions, so, this research focus on this field. Mining in opinions can play role in extracting hatred opinions which are written on web for bad goals. On other hand, there are many challenges to design sentiment analysis system such as: there are large numbers of dialects and slangs for the same language and each one of them has different meaning and grammar to formulate sentence. This work suggests that future work in this area can be done for Arabic tweets.




References

- Aggarwal R. and Gupta L. (2017) A Hybrid Approach for Sentiment Analysis using Classification Algorithm, International Journal of Computer Science and Mobile Computing, 149-157.
- Basari A.S.H., Hussin B., and Ananta I.G.P. (2012) Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization, Malaysian

- Technical Universities Conference on Engineering and Technology (MUCET), 545-552.
- Bholane Savita D. and Prof. Deipali Gore (2016) Sentiment Analysis on Twitter Data Using Support Vector Machine, *International Journal of Computer Science Trends and Technology (IJCTST)*, 365-370.
- Chen B. (2011) Topic Oriented Evolution and Sentiment Analysis, The Pennsylvania State University, Information Sciences and Technology, PhD Thesis, 1-126.
- Gaur M. and Pruthi J. (2017) A Survey on Sentiment Analysis and Opinion Mining, *International Journal of Current Engineering and Technology*, p. 445.
- Govindarajan M., Romina M. (2013) A Survey of Classification Methods and Applications for Sentiment Analysis, *The International Journal of Engineering and Science (IJES)*, page: 12.
- Saraswathi K. and Tamilarasi A. (2014) Investigation Of Support Vector Machine Classifier For Opinion Mining, *Journal of Theoretical and Applied Information Technology*, 291-296.
- Kang H., Yoo S.J. , Han D. (2012) Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, pp. 39.
- Kethavath S. (2015) Classification of Sentiment Analysis on Tweets using Machine Learning Techniques, PhD thesis, National Institute of Technology Rourkela, India. 1-28.
- Khan A., Baharudin B. and Khan K. (2011) Sentiment Classification Using Sentence-level Lexical Based Semantic Orientation of Online Reviews, *Trends in Applied Sciences Research*, 1141-1157.
- Kulkarni A.A., Hundekar V.A., Sannakki S.S. and Rajpurohit V.S. (2017) Survey on Opinion Mining Algorithms and Applications, *International Journal of Computer Techniques*, p. 9.
- Melville and Gryc W. (2009) Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification, *KDD'09*, June 28–July 1 2009, France.
- Sahare S.A. (2017) A Survey Paper on Opinion Mining and Sentiment Analysis, *IJARIE*, pp. 5075.
- Prabowo R. and Thelwall M. (2009) Sentiment analysis: A combined approach, *Journal of Informetrics*, pp. 143–157.
- Preety and Dahiya S. (2015) Sentiment Analysis Using Svm And Naïve Bayes Algorithm, *International Journal of Computer Science and Mobile Computing*, pp. 212-219.
- Rababah O.M., Hwaitat A.K. and Al Qudah D.A. (2016) Sentiment analysis as a way of web optimization, *Scientific Research and Essays*, pp. 90-91.
- Rocha D., Pacheco M.A. and Mendonza L.F. (2016) Sentiment Analysis on Web-based Reviews using Data Mining and Support Vector Machine, *Int'l Conf. Information and Knowledge Engineering*, 113-117.
- S.S.K. Rastogi , Singhal R. and Kumar A. (2014) An Improved Sentiment Classification using Lexicon into SVM, *International Journal of Computer Applications*, 37-42.
- Saleh M.R., Martín-Valdivia M.T., Montejo-Ráez A. and Ureña-López L.A. (2011) Experiments with SVM to classify opinions in different domains, *Expert Systems with Applications*, 14799–14804.
- Seerat B. and Azam F. (2012) Opinion Mining: Issues and Challenges: A survey, *International Journal of Computer Applications*, p. 42.
- Sharma R., Nigam S. and Jain R. (2014) Opinion Mining of Movie Reviews at Document Level, *International Journal on Information Theory (IJIT)*, 13-21.
- Sharma S.P., Tiwari R. and Prasad R. (2017) Opinion Mining and Sentiment Analysis on Customer Review Documents: A Survey, *International Journal of Advanced Research in Computer and Communication Engineering*, p. 156.
- Shrivatava A. and Pant B. (2012) Opinion Extraction and Classification of Real Time

- Facebook Status, Global Journal of Computer Science and Technology, 35-39.
- Suresh H. and Raj S. (2015) Analysis of Machine Learning Techniques for Opinion Mining, International Journal of Advanced Research, 375 – 381.
- Tan S. and Zhang J. (2008) An empirical study of sentiment analysis for Chinese Documents, Expert Systems with Applications, 2622–2629.
- Tanesab F.I., Sembiring I. and Purnomo H.D. (2017) Sentiment Analysis Model Based on Youtube Comment Using Support Vector Machine, International Journal of Computer Science and Software Engineering (IJCSSE), 180-185.
- Xia R., Zong C. ,and Li S. (2011) Ensemble of feature sets and classification algorithms for sentiment classification, Information Sciences, 1138–1152.
- Zhang Z., Ye Q., Zhang Z., and Li Y. (2011) Sentiment classification of Internet restaurant reviews written in Cantonese, Expert Systems with Applications, 7674-7682.

Biodata

	<p>Ms. Rana Zuhair Alobaidy received her B.Sc. degree in Computer Science from College of Education, and the M.Sc. degree from College of Computer Science and Mathematics, University of Mosul. Her current research interests include natural language processing, and data mining.</p>
	<p>Dr. Ghayda Abdulaziz Altalib is an assistant professor at Computer Science Department, University of Mosul, Iraq. She received her B.Sc. degree in Physics College of Science, and the M.Sc. and Ph.D. degrees in artificial intelligence from College of Computer Science and Mathematics, University of Mosul. Her current research interests include natural language processing, data mining, information retrieval, and compiler design.</p>
	<p>Dr. Zainab Senan Attar Bashi is a visiting senior lecturer at the School of Computing, Universiti Utara Malaysia. She received her B.Sc. degree in electronic and computer engineering and the M.Sc. and Ph.D. degrees in information and computer engineering from International Islamic University Malaysia. Her current research interests include cyber security, info-centric networks, network mobility, and the Internet of Things (IoT) connectivity.</p>

Arabic Abstract

رنا زهير العبيدي¹، غيداء عبد العزيز الطالب²، زينب عطار باشي³

¹، ²قسم علوم الحاسبات، كلية علوم الحاسوب والرياضيات، جامعة الموصل، العراق

³كلية الحوسبة، جامعة أوتارا الماليزية، ولاية قدح، ماليزيا

zainab.senan@uum.edu.my ، ghaydabdulaziz@uomosul.edu.iq ، rza2451990@gmail.com

الخلاصة. الكم الهائل من البيانات المتاحة على الإنترنت كانت سبباً في زيادة القدرة على تحليل وفهم طريقة تفكير الناس. حيث أضافت ثورة الإنترنت بيانات المليارات من المستخدمين في مستودعاتها. وقد أعطى هذا اهتماماً بالتقريب وتحليل الآراء والمشاعر في السنوات الأخيرة. حيث أصبح ذلك ممكناً بالاعتماد على الأجهزة لتصنيف البيانات ومعالجتها حيث يمكن لمنهج واحد أن يحوي بيانات ضخمة للمراجعة في مخزونه. لهذا فإن التنبؤ بمشاعر وآراء المستخدمين والعملاء مهم للغاية لتحليل المراجعات فهي لا تساعد فقط في زيادة الأرباح ولكنها تقطع شوطاً طويلاً في تحسين وإخراج منتجات أفضل. في هذا البحث، نقدم مسحةً للتقنيات والتطبيقات المتاحة حالياً والتي تبحث في مجال التقريب عن الرأي في نواحٍ مختلفة، مثل: الاقتصاد، والأمن، والتسويق، واكتشاف الرسائل الإقحامية، واتخاذ القرار، وتوقع الانتخابات. يعتمد الاستطلاع على التقنيات المستخدمة عموماً مع البيانات المكتوبة باللغة الإنجليزية ، إلا أنه مهم للدراسات المستقبلية حول لغات أخرى مثل العربية والماليزية وغيرها

الكلمات الجوهرية. التقريب في البيانات، وسائل التواصل الاجتماعي، شبكات دعم التمييز.