BOOK CHAPTER | Sentiment Analysis

Opinion Mining and Sentiment Analysis: Concepts & Methodologies

[1]Abdulsalami B.A., [2]Baale A. A., [3]Fenwa O.D., & [4]Ismaila W.O.
[1]Department of Mathematical and Computer Sc.., Fountain University, Osogbo, Nigeria.
[2,3,4]Department of Computer Sc.., Ladoke Akintola University of Technology, Ogbomoso, Nigeria
Corresponding Author's Email: basiratabdusalam@gmail.com

Abstract

A great attention has been drawn to the web as a new source of individual opinions due to the ubiquitous nature of the internet as well as ease of accessing documents on the web. The dynamically expanding web and social media and other micro-blogging sites keep generating huge amount of opinion data daily. Individuals, businesses and government organizations can gain more insights about different products or services, and capture the needs of people towards improved policy formulation and plan new strategies for improved service delivery. However, the vast availability of these opinions becomes overwhelming to users especially when there is too much to digest. Analyzing such comments or review manually can be time consuming. As a result, there has been a tremendous need to design methods and implement algorithms which can process a wide range of these text applications. Opinion mining has been found to be one of the best methods or appropriate ways which can automatically generate information by extracting new insights and discover some knowledge from customers' reviews, or peoples' comments on subject matter. This method has been implemented at various levels of abstraction namely document, sentence and aspect level. In recent times, researchers are also investigating concept-level sentiment analysis, a form of aspect-level opinion mining. Researches in opinion mining have included semantic approaches such as ontology and topic modeling techniques. In this book chapter, we cover the brief about opinion mining, also known as sentiment analysis, some of its applications in different domains and the main challenges in opinion mining and sentiment analysis are discussed. Also, existing approaches that have been used to address these challenges.

Keywords: Opinion mining, Sentiment Analysis, Machine learning, Lexicon, Ontology, Cyberspace Abstractions, Probabilistic models, Text Applications,

Introduction

Human life consists of emotions and opinions (Hajmohammadi et al., 2012). Opinions describe peoples' sentiments or feelings towards a particular entity or entities, events or their properties (Liu, 2012; Thakor and Sasi, 2015). In recent years, a great attention has been drawn to the web as a new source of individual opinions.

Citation: Abdulsalami B.A., Baale A. A., Fenwa O.D., & Ismaila W.O. (2022)
Opinion Mining and Sentiment Analysis: Concepts & Methodologies
SMART-IEEE-ACity-ICTU-CRACC-ICTU-Foundations Series

The dynamically expanding web, social media and other micro-blogging sites are generating huge amount of opinion data on daily basis. People post their views on various subject matters via internet forums and social networking sites such as Facebook, Twitter, Instagram, News portal, Blogs, E-commerce sites etc. This presents opportunities and challenges. Individuals, businesses and government organizations can gain more insights about different products and/or services. Besides, it can assist organizations to capture the needs of people, thereby enhancing improved service delivery through improved policy formulation and planning of new strategies to meet customers' requirements and needs. Organizations may no longer need to conduct surveys, opinion polls, and focus groups in order to gather public opinions because there is an abundance of such information publicly available (Liu, 2015; Liu and Liu, 2016).

However, the vast availability of these opinions become overwhelming to users especially when there is too much to digest. Analyzing such opinion comments or review manually can be time consuming. This makes summarization of the web very critical (Kim and Ganesan, 2011). As a result, there has been a tremendous need to design methods and implement algorithms which can process a wide range of these text applications. With this development, there has been an increasing interest in methods for automatically extracting and analyzing people's opinions from web documents (Hajmohammadi et al., 2012). Researchers have begun to explore Opinion mining as one of the best and appropriate method that can automatically generate information by extracting new insights and discover some knowledge from customers' reviews, or people's comments on subject matters.

Opinion Mining (OM), also called Sentiment Analysis (SA), Sentiment Mining, Sentiment classification or Review Mining, refers to the application of Natural Language Processing (NLP), Computational linguistics, and text analytics to identify and extract subjective information in source materials (Tuchowski, 2014). Its basic task is to determine the subjectivity and polarity (positive, negative or neutral) of a piece of text in other words. Research in OM is gaining more popularity among researchers and commercial companies due to its pervasive real-life applications. It is a highly challenging research topic that covers many novel sub-problems (Liu, 2012). It combines many techniques from Text mining (TM), Data Mining (DM) and NLP research areas such as Information Retrieval (IR) and Knowledge Discovery (KD). It emphasizes the analysis of users' opinions, comments, feelings, and characteristics among others based on features from products, services, individuals, organizations, and events (Liu, 2012; Yaakub et al., 2012). The term Sentiment analysis is more commonly used in the industry, while in academia both SA and OM are frequently employed (Liu, 2015). They basically represent the same field of study.

Applications and Challenges of Opinion Mining

OM has a wide range of applications from different domains such as commercial, broadcasting, marketing, education, research and development, governmental policies, health care, and others (Pang and Lee, 2008; Alkadri and Elkorany, 2016; Liu, 2012; Yaakub et al., 2012). It has been commonly applied to several areas such as tracking sentiment towards products, movies, politicians, and companies (Pang and Lee, 2008; O'Connor et al., 2010; Yakub, 2012; Mohammad, 2015; Zunic et al., 2020), improving customer relation models (Bougie, Pieters and Zeelenberg, 2003), detecting happiness and well-being (Schwartz, et al., 2013), tracking the stock market, confirming theories in literary analysis (Hassan, Abu –Jbara, and Radev, 2012) and automatically detecting cyber-bullying (Nahar et al., 2012). However, the need for sophisticated methods and approaches for efficient and timely extraction, acquisition, and formalization of knowledge from unstructured text data still poses several challenges.

The continuous overwhelming volume of unstructured text data makes it difficult to manually extract the critical concepts embedded in the data, coupled with the lean use of language and vocabulary which results into inconsistent vocabulary and the different types of noises that are observed in unstructured text data, such as misspellings, additional white spaces, and abbreviations. (Xu et al., 2019). As a result, OM performance is still a subject of concern in research studies till date.

Levels of Opinion Mining

According to (Liu, 2012, 2015; Hajmohammadi et al., 2012), OM research has been investigated at three different levels of granularity: document, sentence and aspect level. These are described in this section.

Document level: This is commonly known as document-level sentiment classification. The OM task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment (Pang et al., 2002; Turney, 2002; Liu, 2012). For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product. This level of analysis assumes that each document expresses opinions on a single entity. Thus, it is not applicable to documents which evaluate or compare multiple entities (Liu, 2012). However, vast majority of OM research works have been implemented at this level. Among the existing works are the works done by Zhang et al., 2011; Pang et al., 2002; Turney, 2002; Prabowo and Thelwal, 2009.

Sentence level: The task of this level is to determine whether each sentence expressed positive, negative or neutral opinion (Liu, 2012). It is similar to document-level classification because sentences can be regarded as short documents. However, it is often harder because the information contained in a typical sentence is much less than that contained in a typical document because of their length difference (Liu, 2015). Most research work in document-level sentiment classification ignore the neutral class because it is more difficult to perform three-class classification (positive, neutral, and negative) accurately. However, for this level, the neutral class cannot be ignored because an opinion document can contain many sentences that express no opinion or sentiment (Liu, 2015). Also, this level of analysis cannot handle sentences with opposite opinions. For example, "*Apple is doing well in this bad economy*". This sentence is often regarded as containing a mixed opinion. Thus, like document sentiment classification, the problem of sentence level sentiment classification is also somewhat restrictive because it is not applicable to many types of sentences owing to its ignorance of opinion targets.

Aspect level: This level was earlier referred to as feature level (feature-based OM and summarization) (Hu and Liu, 2004). It performs finer-grained analysis than document level and sentence level analysis, because the later does not discover what exactly people liked and their dislikes (Liu, 2012, 2015). Classifying opinion text at the document level or sentence level as positive or negative is insufficient for most applications, because these classifications do not identify sentiment or opinion targets or assign sentiments to the targets. If each document evaluates to a single entity, a positive or negative opinion document about an entity does not mean that the author is positive or negative about every aspect of the entity. For a more complete analysis, there is need to discover aspects and determine whether the sentiment is positive, negative, or neutral about each aspect. To achieve this, researchers proposed an aspect-based OM. To support this claim, Hajmohammadi et al., (2012), argued that without knowing the target of an opinion sentence, the polarity detected for such sentence cannot be useful.

Realizing the importance of opinion targets also helps researchers to understand the OM problem better (Liu, 2012). Thus, the goal of this level of analysis is to discover sentiments on entities and their aspects. For example, the sentence "The iPhone's call quality is good, but its battery life is short" evaluates two aspects, call quality and battery life, of iPhone. The sentiment on iPhone's call quality is positive, but the sentiment on its battery life is negative. The call quality and battery life of iPhone are the opinion targets. Based on this level of analysis, a structured summary of opinions about entities and their aspects can be produced, which turns unstructured text to structured data and can be used for all kinds of qualitative and quantitative analyses.

Opinion Mining Methods

Different methods have been proposed and implemented in OM research. The most common methods are explained here.

1. *Association Rules:* One of the popular techniques used for OM is the use of association mining rule to find the product feature from frequent noun, as infrequent noun is hardly referred to as product feature. This technique was proposed by Hu and Liu (2004), and later improved on by Popescu and Etzioni (2005) with the introduction of part-of-relations that removes the frequent noun, which is not a feature. However, this technique is time consuming as it needs to use query on web for finding the product features (Yaakub et al., 2012).
2. *Manual:* According to Liu, (2012), the human approach is the best technique to identify feature and its sentiment as humans know exactly the meaning of every sentence. However, this technique is most expensive and time consuming, as it will take a long time to analyze the enormous amount of opinion sentence. Currently, this technique is used to evaluate automatic technique as it was believed that manual technique is the most accurate technique.
3. *Frequency-Based Aspect Extraction:* This method finds explicit aspect expressions that are nouns and noun phrases from a large number of reviews in a given domain using a Part of Speech (POS) tagger, and then counts their occurrence frequencies using a data mining algorithm, keeping only the frequent nouns and noun phrases using an experimentally determined frequency threshold. The approach works because aspects are usually expressed as nouns and noun phrases, and when people comment on different aspects of an entity, the vocabulary that they use usually converges.
4. *Lexicon Approach:* Opinion lexicon is a set of opinion words compiled by a system such as adjectives, adverbs, verbs and nouns (Yaakub et al. 2012). NLP's technique was used to extract a review based on Part-of-Speech (POS)'s tag. Afterwards, the identification of frequent features was implemented using association rules technique and supervised mining method. Frequent nouns were used to find the nearest opinion words based on adjectives. There are two forms of lexicon-based approach: Dictionary-based and Corpus-based. The former does not need any training to be developed. Thus, it is considered as an unsupervised method. The two most popular applications in OM research area that were developed based on this technique are WordNet and SentiWordNet. The Corpus-based approach was developed based on the weakness of dictionary-based approach and it is domain independent. The advantage of using lexicon is that the model can easily identify opinion polarity in sentences or a product review.
5. *Machine Learning (ML):* Generally, there are two approaches in this method: Supervised and unsupervised. Most of the researches in OM used the supervised technique. The tasks of supervised ML are to train a function, to make the system capable of identifying sentiment orientation and also understand how to use a list of sentiment of sentiment's corpus in document. The current state-of-the-art machine learning methods include *Hidden Markov Models* (HMM) and *Conditional Random Fields* (*CRF*). Other popular

techniques in supervised ML are *Support Vector Machines* (SVM), *Naïve Bayes* (NB), *Maximum Entropy Classification* (MES) and *Neural Networks* (NN).

6. *Topic Modeling: Topic modeling is an unsupervised principled method for discovering topics from a large corpus of text documents. The most common outputs of a topic model are a set of word clusters and a topic distribution for each document. Each word cluster is called a topic and is a probability distribution over words in the corpus. The two majorly used topic models are probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA). LDA is an unsupervised learning model that assumes that each document consists of a mixture of topics and each topic is a probability distribution over words. It is a document generative model that specifies a probabilistic procedure by which documents are generated.*
7. *Ontology-based Approach:* Considering the various hurdles in conducting sentiment analysis of unstructured texts, researchers have moved on to explore new approaches based on both semantic web technologies and domain-dependent corpora for feature-based OM (Alkadri and ElKorany, 2016). This was mostly achieved by utilizing an ontology (Zehra et al., 2017a). In OM, ontology is the ability to share knowledge, exchange information and minimize ambiguity (Xue et al., 2009). According to Kontopoulos et al., (2013), the motivating factor for preferring the use of ontologies in an application includes analyzing domain knowledge and separating it from operational knowledge; enabling the reuse of domain knowledge; making domain assumptions explicit; sharing a common understanding of the information structure among people and software agents. Major works that have deployed ontologies in the micro-blogging domain includes Kontopoulos et al. (2013); Yaakub et al., (2012) ; Zehra et al., (2017a); Thakor and Sasi, (2015); Alkadri and Elkorany, (2016) among others.

Conclusion

We have introduced in this book the field of OM or sentiment analysis. The basic knowledge about OM such as its different levels of opinion analysis, various applications and state -of-the-art techniques. OM has been a very active research area in different fields of computer science fields, such as data mining, web mining, information retrieval and NLP. It has also spread to management science and other social science fields such as communications and political science because of its importance to business and society as a whole (Liu, 2012). With the continuous growth of social media on the web, the importance of sentiment analysis cannot be overemphasized.

References

Alkadri, A. M., and Elkorany, A. M. (2016). Semantic Feature Based Arabic Opinion Mining Using Ontology. 7(5), 1–7.

Bougie, J. R. G., Pieters, R., and Zeelenberg, M. (2003). Angry customers don't come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services. Open access publications from tilburg university, Tilburg University.

Hajmohammadi, M. S., Ibrahim, R., and Ali Othman, Z. (2012). Opinion Mining and Sentiment Analysis: A Survey. International Journal of Computers & Technology, 2(3), 171–178. https://doi.org/10.24297/ijct.v2i3c.2717.

Hassan, A., Abu-Jbara, A., Jha, R., and Radev, D. (2012). Identifying the semantic orientation of foreign words. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: short papers (ACL-2011).

Hatzivassiloglou, V. and Wiebe, J. (2000). Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In Proceedings of International Conference on Computational Linguistics (COLING-2000).

Hu, M., and Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, pp. 168-177, New York, NY, USA.al of Computers & Technology, 2(3), 171–178. https://doi.org/10.24297/ijct.v2i3c.2717.

Kim, H., and Ganesan, K. (2011). Comprehensive review of opinion summarization. Illinois Environment for ..., 1–30. http://www.ideals.illinois.edu/handle/2142/18702.

Kontopoulos, E., Berberidis, C., Dergiades, T., and Bassiliades, N. (2013). Ontology-based Sentiment Analysis of Twitter Posts. https://doi.org/10.1016/j.eswa.2013.01.001.

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–184. https://doi.org/10.2200/S00416ED1V01Y201204HLT016.

Liu, B. (2015). Sentiment analysis: Mining opinions, sentiments, and emotions. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, 1–367. https://doi.org/10.1017/CBO9781139084789.

Liu, B., and Liu, B. (2016). Sentiment Analysis : A Multi-Faceted Problem Sentiment Analysis : A Multi L. -Faceted Problem. March, 2–7.

Mohammad, S. M. (2015). Challenges in Sentiment Analysis. Section 7.

Nahar, V., Unankard, S., Li, X., and Pang, C. (2012). Sentiment analysis for effective detection of cyber bullying. In Web Technologies and Applications, pp. 767-774.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86.

Pang, B., and Lee, L. (2008). Presentation: Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 1(2), 1–135. https://www.nowpublishers.com/product.aspx?product=INR&doi=1500000 001.

Popescu, A. M., and Etzioni, O. (2005). Extracting product features and opinions from reviews. In Proc. Conf. Human Language Technology and Emprical Methods in Natural Language Processing, pp 339–346.

Prabowo, R. and Thelwall, M. (2009). "Sentiment Analysis: A Combined Approach", Journal of Informetrics, 3 (2), 143-157.

Schwartz, H., Eichstaedt, J., Kern, M., Dziurzynski, L., Lucas, R., Agrawal, M., and Park, G., (2013). Characterizing geographic variation in well-being using tweets. In Proceedings of the International AAAI Conference on Weblogs and Social Media.

Thakor, P., and Sasi, S. (2015). Ontology-based Sentiment Analysis Process for Social Media Content. Procedia - Procedia Computer Science, 53, 199–207. https://doi.org/10.1016/j.procs.2015.07.295

Tuchowski, J. (2014). MSc Katarzyna Wójcik 1 , MSc Janusz Tuchowski 2 Computational Systems Department Cracow University of Economics.

Turney, P. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In Association for Computational Linguistics (ACL2002), pages 417–424.

Xu, Y., Rajpathak, D., Gibbs, I., and Klabjan, D. (2019). Automatic ontology learning from domainspecific short unstructured text data. arXiv preprint arXiv:1903.04360.

Xue, Y., Wang, C., Ghenniwa, H.H., and Shen, W. (2009). A tree similarity measuring method and its application to ontology comparison. Journal of Universal Computer Science, 15(9):1766–1781.

Yaakub, M. R., Li, Y., Algarni, A., and Peng, B. (2012). Integration of opinion into customer analysis model. Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, WI-IAT 2012, June, 164–168.

Zehra, S., Wasi, S., Jami, I., Nazir, A., Khan, A., and Waheed, N. (2017a). Ontology-based Sentiment Analysis Model for Recommendation Systems. Keod, 155–160. https://doi.org/10.5220/0006491101550160.

Zunic, A., Corcoran, P., and Spasic, I. (2020). Sentiment Analysis in Health and Well-\being: \systematic \review JMIR Med Inform; 8 (1): e16023 doi: 10.2196/16023.