International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-12, Special Issue-1, March-2024
https://doi.org/10.55524/CSISTW.2024.12.1.3
Computing for Sustainable Innovation: Shaping Tomorrow's World (CSISTW-2024) on 16th-17th February, 2024
Organized by B.N. College of Engineering and Technology, Lucknow, India

A Review on Challenges in Recent Opinion Extraction Techniques

Devendra Kumar[1], and Faiyaz Ahamad[2]

[1]Research Scholar, Department of CSE, Integral University, Lucknow, UP, India
[2]Associate Professor, Department of CSE, Integral University, Lucknow, UP, India

Correspondence should be addressed to Devendra Kumar; devendrakumar17@gmail.com

ABSTRACT-With the introduction of Web 2.0, people are now encouraged to share information about events in their community as well as to voice their opinions and beliefs. People utilize social media, blogs, review sites, news sites, user feedback portals, and other platforms to express themselves regarding events, locations, decisions made by leaders, policies, and other topics. Opinion extraction is a method that automatically extracts and analyzes people's sentiments from their thoughts. Manufacturers may find this collected data helpful in analyzing their products. Sentiment analysis, also referred to as opinion extraction, has been the subject of intense research over the past 15 years by academics, research communities, and service corporations in an effort to gather and examine public opinions and moods.

This work focuses on the difficulties encountered in opinion extraction at many levels, including word-, sentence-, and language-related difficulties as well as some generic difficulties. The study provided clear examples and a thorough discussion of 30 hurdles, which aid in presenting problems for text or opinion extraction research. Opinion extraction tools are briefly presented in this section to help novice researchers get started. The published material spans the years 2004 to 2022.

KEYWORDS: Opinion extraction, sentiment analysis, opinion extraction tools, problems in opinion extraction.

I. INTRODUCTION

People have the need and desire to communicate who they are and what they think about the world. The World Wide Web (WWW) offers people a variety of platforms on which to voice their thoughts, and these viewpoints are highly beneficial to a particular group for several reasons. The social networking sites (Facebook, Twitter), blogs (generic blogs, such wix.com, web.com, or specific purpose blogs, like Mkyong.com). [1]

When transforming it into a valuable information package, these viewpoints are beneficial for making decisions, enhancing services, improving the quality of a product, etc. Opinion extraction is a field that emerged from the analysis of user opinions and the reduction of product information overhead, among other reasons. [2]

Each of these situations has a number of difficulties; they include not just gathering textual opinions in an effective manner but also analyzing and presenting them in a way that maximizes their potential for the community in question. Text extraction, language processing, and text classification are the basic components of opinion extraction.Mkyong.com), social media platforms (Twitter, Facebook).

Opinion and text extraction now face several shortcomings and difficulties that text extraction has a huge potential to solve. Unfortunately, because human language is so complex, natural language processing, or NLP, faces several challenges. The problems with sentiment analysis and opinion extraction, as well as certain well-known tools that have been mentioned in the literature, are the two main topics of this work. The report addressed challenges in detail in one section, and opinions in another. [3]

II. OPINION EXTRACTIONS

Opinion extraction is the process of extracting biased text from many sources and summarizing it so that the reader can understand it [4]. Opinion extraction is a technique used to extract meaningful information from textual sources.[5] The opinions and views of users are gathered from many sources, including as news websites, blogs, review websites, and social media platforms, and are then categorized based on their degree of polarity. Polarity is categorized as follows:

Positive connotes goodness or a favorable opinion; negative connotes unfavorable or neglectful opinions; and neutral connotes surprise or the absence of particular preferences.

Sentiment analysis is crucial for closely extraction public opinion and figuring out the following three issues [6]:

A. Subjectivity

The text is examined to see if it is subjective, which indicates that any opinionated content in the text is being examined.

B. Polarity

There are two categories of polarity: positive and negative opinions.

C. Strength of Polarity

There are several degrees of polarity, such as extremely high, high, low, and extremely low, among others. Highly positive, mildly positive, weekly positive, weekly negative, mildly negative, and highly negative are examples of polarity strength representations.

D. Object

An object is an entity about which the user has expressed opinions or feelings in light of their experience or knowledge. A person, issue, product, organization, thing, place, etc. can all be considered entities.

E. Opinion Holder

An individual using the internet to voice opinions or feelings about a product, or an individual acting on behalf of a company or client in the case of a product vendor.

F. Opinion

An opinion is a written statement that expresses the writer's feelings or point of view toward a subject.

G. Peer

A user, organization, or customer may profit from the ratings that peer—opinion holders—determine based on their opinions.

III. CHALLENGES IN OPINION EXTRACTION AND SENTIMENT ANALYSIS

The field of opinion extraction is new and constantly changing, which presents a lot of difficulties. For instance, which textual passage contains the opinion, how polarized the opinion is, how strong the polarity is, whether the opinion is genuine or credible, whether spam is detected, and the sentiment.

The opinion, identifying biased or irrelevant opinions, identifying phony opinions, identifying unclear facts, etc. are open topics for additional study and improvement.

Opinions are usually expressed informally and in a wide variety of ways, which presents a number of difficulties.

This investigation clearly identified the difficulties covered in the next section. Preserving the authenticity of the guidelines maintaining the integrity of the specifications [7-9].

3.1. Authority: It is imperative that the opinion be accepted. The information should come from a reliable source that is respected by the community; in other words, the authority of the opinion holder can be examined, for example, to determine whether the view was formed by a persuasive force or a domain expert (an expert on the issue) [10–12].

3.2. Non-Expert Opinion: Non-expert opinions negatively impact blogs and online forums. Opinion extraction is impacted by specialized blogs, such as programming blogs, where expert opinion counts but isn't properly reviewed or expressed [11- 13].

3.3. Spam Opinion Detection: Bing Liu and Nitin Jindal [14] talked about spam opinions and the veracity of internet opinions. A detailed analysis of the spam opinion has been conducted on 5.8 million reviews.

3.4. Biased or Spam Opinion: Venal reviews or opinions are added on purpose to have an impact the aggregate opinion research regarding a good or service [15]. Ascertaining the

According to the opinion extraction model, which is depicted? There are four stockholders in the opinion extraction ecosystem.

presence of these viewpoints is crucial for sentiment analysis and a challenging task.

3.5. Opinion Credibility: One of the major issues is deter extraction how to evaluate an opinion's credibility or if it is believable and trustworthy [16].

3.6. Sentiment Analysis's Domain Dependency: The sentiment analysis's resultant generalization poses a significant issue. Due to the domain dependence of sentiment terms, a sentiment analysis technique may work well in one domain while performing poorly in another [17].

3.7. Divergent from the Topic (Opinion Relevancy): Relevant comments may or may not be found on the websites that the crawler restricted to particular topics found. Users' comments may touch on topics unrelated to the topic at hand or even unrelated to other topics. On Twitter, for instance, users possess a variety of interests, making it challenging to apply a consistent lexical model to every user. In a similar vein, we discovered critiques of other chat shows that were broadcast on various TV networks or shows that air both before and after the "Off the Record," which is unpredictably irrelevant [18–20].

3.8. Distinction between Search Keyword and Opinion Topic: This is a difficulty mostly related to online sentiment analyzers; the topic of the webpage that is returned is not always the subject of the sentiment expressed on it. For instance, the target of the viewpoint may be entirely different if the search term has nothing to do with the tweet's purpose [22].

3.9. Volatility of Views over Time: As information develops and become more clear, opinions on a subject shift over time, drastically shifting from negative to positive and vice versa. Views on social networking sites, particularly Twitter, show a very significant temporal dynamism. For instance, monitoring public perceptions of politicians' changes over time and party support in political activities.

3.10. Object Features Extraction: A supervised pattern extraction strategy is suggested in this noun-based approach, which was established by [21]. Using commonly used words, the noun-based approach determines the true characteristic of the object. For instance, the review "The voice on my phone was not clear" associates the negative attitude "not clear" with the feature "voice" of the object "phone." With the existing capabilities, it is challenging to extract all of this information from opinion [22].

3.11. Strength of Opinions: Deter extraction the strength of an opinion or the collective view of several users is a significant obstacle for opinion extraction. As the conversation continues, opinions may become stronger or even shift from being positively to negatively polarized, or vice versa. This problem is somewhat addressed by the Senti-WordNet application [23].

3.12. Typographical Errors: People may comprehend terms such as "schol" for "school," "knowledge" for "knowledge," and other typos. However, activities like typo mistake extraction and analysis are difficult for an opinion miner [24].

3.13. Natural Language Processing Overhead: The sentiment analysis tool was hampered by natural language overhead such as ambiguity, co-reference, implicitness, inference, etc.

IV. CHALLENGES AT WORD LEVEL

4.1 Using Polysemy (Context Dependent Words): It might be challenging to analyze in an effective manner when employing words or phrases that have multiple meanings depending on the context in which they are used. For instance, pupil for eye and pupil for a student in "He left the bank an hour ago." situation-dependent terms, which have dual meanings depending on the situation, are outside the scope of lexicon approaches. For instance, the term "small" in "The size of Q Mobile is very small" does not indicate whether it is being used negatively or positively [25].

4.2 Contextual Information (Use of Simile or Metaphor): In order for the reader to comprehend and interpret the user's opinion accurately, contextual information is necessary. For instance, in [26], a politician was associated with the well-known (and wicked) Voldemort from the Harry Potter series in a tweet that cannot be examined with the tools available today. We are able to summarize this since metaphor and simile use is not automatically understood. For instance, the simile "You are beautiful like a flower" compares a person's beauty to a flower, whereas the metaphor "You are a lion" refers to someone who is brave like a lion.
This type of review is difficult to polarize and understand automatically.

4.3 Grouping Synonyms: When two customers commend a mobile device by saying, "The phone's sound quality is very good" and "The voice quality is excellent," they are both praising the device.

4.4 Word Orientation Issue: It's challenging to determine how the words in a view point and may vary depending on the circumstances. For instance, in the assessment "The battery life of this mobile device is impressive," the adjective "small" is used positively."Time is also small," [27] says, using the negative connotation of the word small. In a different context, it can be challenging to determine the polarity of the adjective words [28].

4.5 Sarcastic and Ironic Statement: Owing to incorrect orientation, sarcastic and ironic (humorous sarcastic) statements exist and mislead judgments. Positive words can have a negative meaning in sarcastic statements, which is hard to spot because sarcastic words give misleading directions and incorrect orientations.

4.6 Adjective, Verb, and Adverb as Opinion Words: Liu [17] asserts that only adjectives are capable to be employed as a word of opinion, yet it has been noted that verbs and adverbs can also express opinions. For instance, in the sentence "I like this book," the word like is employed as an word of opinion. The Liu technique finds and removes sentences that include one or more feature words, disregarding the other words [29].

4.7 Employing Short Forms or Abbreviations: In an effort to save time and energy, the trend of using social media (Facebook, Twitter), messaging services (WhatsApp, Tango, Skype), and SMS encourages the use of acronyms, abbreviated versions, or shortcuts for numerous phrases. As an illustration, "I'm f9, dear hru?"

V. DIFFICULTIES WITH SENTENCE OR LANGUAGE LEVEL

5.1 Opinion in Different Languages: Depending on the context, different users express their thoughts in different languages. They also post reviews in their own tongues on English for ums .For instance, someone may say in Roman Urdu, "Murad sahib bohat hi as an tareke se samjate hay" and another person may express the same sentiment in Pashto, and so forth. It is challenging to use the opinion extraction method for effective sentiment analysis in such a situation [30].

5.2 Product Opinion: Depending on the buyer's language preference, product reviews, comments, and feedback may be provided in "English, Arabic, Urdu, and Pashto, etc."Therefore, holding the orientation of each and every language and deter extraction the polarity of the product is a difficult task [31].

5.3 Different Style Writing (Different methods of Expressing Opinions): There are various methods to write a phrase depending on the writer. An opinion might have both advantages and disadvantages. As a result, it is challenging to parse and evaluate an opinion that incorporates both positive and negative sentiment at the phrase level.

5.4 Incorporated Implicit Behavior Data in Opinion: Words that implicitly convey behavior data are incorporated into the views to ascertain the sentiment words' true behavior [32].

5.5 Opinion Aggregation from Multiple Sources: This process identifies opinions that have been extracted from various sources and can be combined easily if, for example, there are no discrepancies in the attributes of an entity. However, views act differently in this context: we recommend populating a knowledge base since numerous opinions may be associated to an entity and require independent modeling.

5.6 Incorporated Implicit Behavior Data in Opinion: Words that implicitly convey behavior data are incorporated into the views to ascertain the sentiment words' true behavior.

5.7 Opinion Aggregation from Multiple Sources: This process identifies opinions that have been extracted from several sources and can be combined easily if, for example, there are no discrepancies in the attributes of an entity. Here, however, views act differently: we recommend populating a knowledge base because several opinions may be associated to an entity and require independent modeling.

5.8 Crossed Lingual Opinion: Users in a certain community may utilize numerous languages while expressing difficult-to-analyze opinions about a good or service. For instance, the Pashto phrase "Da Mobile der khasta da, it’s slim" words) as well as English.

5.9 Colloquial Language: Also known as colloquial dialect, colloquial language is any informal, casual, or spoken language that an employee uses in discussion. It can take many different forms. To put it simply, the dialect is used for everyday conversation or casual communication. It might be challenging to evaluate words or phrases like "what's up?" for "What is going on?" and "yeap" for "yes," among others, in their actual sense.

VI. RESOURCES TO OPINION EXTRACTION

The tools utilized in the opinion extraction process are presented in this section. Sentiment analysis is a technique that opinion extraction systems employ to identify the polarity of user reviews. Opinion extraction's primary goal is to extract sentiment or feelings from the brief textual evaluation. This process presents a variety of difficulties, all of which are covered in detail in section 3. In this section, we will quickly go over a few well-known, readily available tools for opinion extraction. These tools serve a variety of functions. For example, pre-processing comprises sentiment extraction, opinion analysis, text categorization, clustering, cleaning, and cleansing.
A range of tools are available for opinion extraction; some are free and open source, while others must be purchased.

- **Opinion Finder:** Researchers from Cornell University and the University of Pittsburgh created Opinion Finder. It is a system that reads papers and recognizes subjective statements and different types of subjectivity automatically. Phrases It encompasses direct subjective expressions and speech events, emotion expressions, and agents who are opinion sources (opinion holders).
- **Opinion Observer:** is a tool for analyzing and contrasting user-generated content from various websites. Additionally, it displays the outcomes in a graphical manner in relation to the ideas produced for each feature of the product. It assigns prior polarity using WorldNet exploring techniques [8-12].
  A toolset for machine learning-based natural language text processing is the Apache OpenNLP library. Along with doing typical NLP tasks like chunking, tokenization, extracting named entities, parsing, and a part-of-speech tagger, it also includes machine learning approaches based on perception and entropy. Advanced text processing services also need these tasks [1-5].
- **Natural Language Tool Kit (NLTK**) One of the best platforms for developing Python applications to work with statistical and symbolic natural language data is called Natural Language Tool Kit (NLTK). Lexical resources like WorldNet and a collection of text processing libraries are available for parsing, tokenization, classification, stemming, tagging, and semantic reasoning. NLTK utility for classification, clustering and parsing research and education offers user-friendly interfaces for more than fifty corpora. Linux, Mac OS X, and Windows can all use NLTK [10-15].
- **Stanford Parser**: A breakthrough in 1990, Stanford Parser is a natural language parser. It is mostly utilized for the sentence's grammatical structure. For instance, which words belong in phrases and which words stand for verb, subject, or object [16-20]?
- **Red Opal** is an opinion extraction tool that lets people find out what products' opinions are oriented toward. Based on the extraction of product features through an analysis of user reviews, scores are allocated to the products. The outcomes of the red opal tool are displayed via the web-based interface. When compared to other tools, it performs admirably.
- **Ling Pipe** is a toolbox for opinion extraction that processes text using computational linguistics. It is used to specify named entities, such as the name of a person, place, or organization. It is also used to automatically classify tweets and recommend relevant queries. A Java API called Ling Pipe features multilingual, multi-domain, and multi-genre models along with source code and unit tests.
- **Examine the Seer Tool** The work performed by aggregation sites is automated with the help of this instrument. In order to give the extracted feature words a score, positive and negative feedback is gathered using the Naive Bayes classifier approach. The findings are displayed in a straightforward opinion statement.
- **SVM with several classes** An SVM technique implementation can be found in the SVM multiclass. Among its many functions is the ability to solve regression and classification issues. [15].Developed a model that can mine comparative opinions using this method.
- **CRF toolkit** The CRF approach is implemented by the CRF toolbox, a Matlab toolbox. Using this technique, [16] constructed a two-level CRF model for the purpose of obtaining comparative relations.
- **General Architecture for Text Engineering (GATE)**: In 1995, the University of Sheffield developed General Architecture for Text Engineering (GATE), a development environment for creating software that can process text and be used for text extraction, opinion extraction, and computational language processing. The primary functions of the GATE are tokenization, named entity extraction, parsing, a part-of-speech tagger, Gazetteer, and coreference tagger. Today, scientists, engineers, businesses, educators, and students use it all across the world [19].
- **Stanford CoreNLP framework**: A suite of natural language analysis capabilities, including tokenization, named entity recognition, parsing, and coreference, are offered by the Stanford CoreNLP framework. Kessler and Kuhn (2013) employed Stanford CoreNLP for tokenization and text segmentation [32].

VII. CONCLUSION

There are many applications for opinion extraction and sentiment analysis, but there are also many research obstacles to overcome. This paper provides a thorough, up-to-date analysis of the issues surrounding opinion extraction based on research conducted in a variety of areas between 2004 and 2022.The study's comprehensive examination of problems addresses every facet of opinion extraction, with a focus on generic, word or phrase-level, sentence or opinion-level challenges.

With every task, we provide a brief example to help you grasp the material better. Opinion extraction and sentiment analysis become increasingly intriguing and provide opportunities for deeper study to better understand the issue and offer a solution. Opinion extraction tools were covered in brief in one section; for more information, see the relevant references. The paper's thorough list of opinion extraction challenges is one of its primary contributions. The report will assist the inexperienced researcher in comprehending the issue, effectively addressing the difficulties, and anticipating a resolution.

REFERENCES

1. Alslaity, A., & Orji, R. (2024). Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions. Behaviour& Information Technology, 43(1), 139-164.
2. Seerat, B., & Azam, F. (2012). Opinion Mining: Issues and Challenges (A survey). International Journal of Computer Applications, 49(9), 42-51.
3. Shamshiri, A., Ryu, K. R., & Park, J. Y. (2024). Text mining and natural language processing in construction. Automation in Construction, 158, 105200.
4. Kaur, A., & Gupta, V. (2013). A survey on sentiment analysis and opinion mining techniques. Journal of Emerging Technologies in Web Intelligence, 5(4), 367-371.
5. Alqaryouti, O., Siyam, N., Abdel Monem, A., & Shaalan, K. (2024). Aspect-based sentiment analysis using smart government review data. Applied Computing and Informatics, 20(1/2), 142-161.
6. Chen, Z. S., Liang, C. Z., Xu, Y. Q., Pedrycz, W., &Skibniewski, M. J. (2024). Dynamic collective opinion generation framework for digital transformation barrier analysis in the construction industry. Information Fusion, 103, 102096.
7. Just, J. (2024). Natural language processing for innovation search–Reviewing an emerging non-human innovation intermediary. Technovation, 129, 102883.
8. Zulqarnain, M., Ghazali, R., Aamir, M., & Hassim, Y. M. M. (2024). An efficient two-state GRU based on feature attention mechanism for sentiment analysis. Multimedia Tools and Applications, 83(1), 3085-3110.
9. Das, R., & Singh, T. D. (2023). Multimodal sentiment analysis: A survey of methods, trends and challenges. ACM Computing Surveys.
10. Das, R., & Singh, T. D. (2023). Multimodal sentiment analysis: A survey of methods, trends and challenges. ACM Computing Surveys.
11. Chen, M., Jiang, L., Ma, C., & Sun, H. (2019, February). Bimodal Emotion Recognition Based on Convolutional Neural Network. In Proceedings of the 2019 11th International Conference on Machine Learning and Computing (pp. 178-181).
12. Mishra, N., & Jha, C. K. (2012). Classification of opinion mining techniques. International Journal of Computer Applications, 56(13).
13. Liu, B. (2022). Sentiment analysis and opinion mining. Springer Nature.
14. Shi, X., Hu, M., Deng, J., Ren, F., Shi, P., & Yang, J. (2023). Integration of Multi-Branch GCNs Enhancing Aspect Sentiment Triplet Extraction. Applied Sciences, 13(7), 4345.
15. Seerat, B., & Azam, F. (2012). Opinion mining: Issues and challenges (a survey).International Journal of Computer Applications, 49(9).
16. Dhokrat, A., Khillare, S., & Mahender, C. N. (2015). Review on techniques and tools used for opinion mining. International Journal of Computer Applications Technology and Research, 4(6), 419-424.
17. Doma, V., & Pirouz, M. (2020). A comparative analysis of machine learning methods for emotion recognition using EEG and peripheral physiological signals. Journal of Big Data, 7(1), 1-21.
18. Zhang, L., & Liu, B. (2011, June). Identifying noun product features that imply opinions. In Proceedings of *the* 49th annual meeting of the Association for Computational Linguistics: human language technologies (pp. 575-580).
19. Roy, S., Dhar, S., Paul, A., Bhattacharjee, S., Das, A., & Choudhury, D. (2013). Current Trends of Opinion Mining and Sentiment Analysis in Social Networks. International Journal of Research in Engineering and Technology, 2(2).
20. Maynard, D., Bontcheva, K., & Rout, D. (2012). Challenges in developing opinion mining tools for social media. Proceedings of@ NLP can u tag# usergeneratedcontent, 15-22.
21. Fujiyoshi, S., Yarimizu, K., Perera, I., Abanto, M., Jorquera, M., & Maruyama, F. (2023). Learning from mistakes: challenges in finding holobiont factors from environmental samples and the importance of methodological consistency. Current Opinion in Biotechnology, 80, 102897.
22. Devendra, Faiyaz. (2024). Application of Machine Learning Algorithm for Optimal Model Design for Opinion Extraction. TELEMATIQUE, ISSN: 1856-4194(pp215– 227).
23. Elbagir, S., & Yang, J. (2018, December). Sentiment analysis of twitter data using machine learning techniques and scikit-learn. In Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence (pp. 1-5).
24. Sharma, A. K. A Comparative Study between Web Mining Tools over some WUM Algorithms to Analyze Web Access Logs. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN, 2278-3075.
25. Kumar, D., & Ahamad, F. (2023, December). Opinion extraction from big social data using machine learning techniques: A survey. In AIP Conference Proceedings (Vol. 2916, No. 1). AIP Publishing.
26. Balakrishnan, V., Shakouri, M. R., &Hoodeh, H. (2012). Predictions using data mining and case-based reasoning: a case study for retinopathy. International Journal of Biomedical and Biological Engineering, 6(3), 55-58.
27. Liu, D., Wu, Q., Han, W., & Zhou, B. (2016). Sockpuppet gang detection on social media sites. Frontiers of Computer Science, 10, 124-135.
28. Baroni, M., &Vegnaduzzo, S. (2004, September). Identifying subjective adjectives through web-based mutual information. In Proceedings of KONVENS (Vol. 4, pp. 17-24).
29. Medhat, W., Hassan, A., &Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, 5(4), 1093-1113.
30. Potdar, G. S., &Phursule, R. N. (2015). Twitter Blogs Mining using Supervised Algorithm. International Journal of Computer Applications, 126(5).
31. Wang, X., & Fu, G. H. (2010, July). Chinese subjectivity detection using a sentiment density-based naive Bayesian classifier. In 2010 International Conference on Machine Learning and Cybernetics (Vol. 6, pp. 3299-3304). IEEE.
32. Kolovou, A. (2019). Machine Learning Methods for Opinion Mining in Text: The Past and the Future. Machine Learning Paradigms: Applications of Learning and Analytics in Intelligent Systems, 429-457.