International Journal of Advances in
Soft Computing and Intelligent Systems (IJASCIS)
2024, Vol 03, Issue 02, 242-252
ISSN: 3048-4987

Science Transactions © 2024

Original Paper

Opinion Mining in Resource-Poor Languages: Techniques and Challenges

Nidhi N. Solanki[a], Dipti B. Shah[b]

[a] M.K. Institute of Computer Studies, Bharuch, Gujarat, India, nidhi17.solanki@gmail.com
[b] Post Graduate Department of Computer Science and Technology, Sardar Patel University,
Vallabh Vidyanagar, Anand, Gujarat, India, dm_shah@spuvvn.edu

Abstract

Human emotions automatically arise from the experience of gratification, sorrow, failure, or success. They are a reflection of behavior and cerebral thoughts. Opinion mining (OM) is crucial in all fields of work today, such as cyber security, smart cities, medicine, education, e-commerce, governance, and agriculture. Consumer understanding is required to increase business profits and strategy building as they are valuable assets. OM is an inevitable task to understand users' beliefs. Ample resources are available for the English dialect, but there is not enough for code-mixed and non-English dialects. This survey focuses on the OM of resource-poor languages. It depicts the various techniques, datasets, and polarities of emotion mining used by scholars in previous research. It illuminates the various practical challenges with suitable examples and possible solutions to develop an intelligent system. It will help to make noteworthy strides in discovering and filling the OM research gaps. We have also analyzed the applications of OM. Our study will hopefully guide future researchers, academicians, industrialists, and learners in accomplishing their tasks.

Keywords

Code-mixed language, Techniques, Dataset, Challenges & Applications

1. Introduction

Social media is a rich source of communication. It is an extending home of information around the globe. Many persons desire to give their judgment in their inherent language, in which they are additionally comfortable and relaxed. The content in a regional language is increasing at a high rate. The Unicode (utf-8) results in bulky non-English internet data. Today, many applications facilitate functions in regional languages. Google search engine, many social media websites, and microblogging sites also support many languages. Retrieval and processing of the same data are mandatory for every service provider group. OM is an intelligent computerized application that helps to get consumers' views from their textual comments [1]. Widespread OM applications prove its scope of need. Today, almost every domain of education, hospitality, entertainment, medical science, government, business, production, purchase, and transaction processing applies OM. OM identifies, customer behavior, satisfaction analyses, and document analysis. Brand monitoring tools, social media analytics platforms, market research platforms, and E-commerce analytics tools are a few cases that pull OM in business intelligence. Experimentations have happened on various regional code-mix languages like Margalish (Marathi in English), Bangalish (Bengali in English), Devanagari, Hinglish (Hindi in English), Tanglish (Tamil in English), Kanglish (Kannada in English), Manglish (Malayalam in English). Still, some more languages and domains need to be studied.

Opinion extraction is possible from subjective sentences only [2]. Objective sentences do not guarantee opinion knowledge. For example, “This book is very nice” is a subjective sentence as it provides an opinion on a book. “I am from India” is an objective sentence that gives some generalized information only. We accomplish this literature survey by reviewing various respectable and noble resources like Springer, IEEE, Google Scholar, Research Gate, and Elsevier. This paper presented a survey of varied techniques and datasets used in the past epoch to perform OM in cross-lingual languages. The diligent motivation behind the article is to unite all the reviews and experiments done in the publications and deliver direction for future work in the area of OM.

2. Techniques of Opinion Mining

Based on existing research documents, opinion mining practice is categorized into classes of lexicon, deep learning, machine learning, and hybrid approach [3].

2.1. Lexicon Based Technique

The lexicon approach is synonymous with the rule-based approach. The opinion lexicon is a source of words with their corresponding opinion score. Scores characterize the nature of the word as positive, negative, or neutral [4]. For an input preprocessed text, the scores are taken collectively based on arithmetic operations to calculate the final score [5]. The final score decides the sentiment polarity. The corpus-based technique & dictionary-based techniques are two subcategories of the lexicon approach.

The dictionary-based approach intakes a list with two attributes of a word and their polarity score. This list may expand by adding new records later. The focus is more on formal sentences. The Corpus method trusts relative statistics, syntactic trace [5] of textual corpora, and a collection of reserved negative and positive kernel words called seed words. Semantic and statistical methods are the types of corpus-based methods [6]. The semantic method uses the semantic knowledge of the word. Words with the same meaning will get the same opinion score. The statistical approach considers the context of words. If words fall together in the same context frequently, they will get the same opinion score [7].

2.2. Machine Learning Technique

As the human brain learns from repeating events, the machine also learns from data and has the power to improve predictions [8]. This procedure is called machine learning, in which the machine learns from historical data. It is an artificial technique. Train and test data are the two portions of a dataset. Training data trains algorithms. Test data [9] tests the output predictions. Feature extraction plays a vital role in this direction. Some commonly used feature sets by published work are unigram, bigram, trigram, statistical features, local context, POS tags, character n-grams, bag-of-words, emoticons, punctuation, intensifiers, hashtags, etc.

Some commonly used feature selection methods by published work are term frequency-inverse document frequency (TF-IDF), term frequency-based bag-of-terms, SentiWordNet, semantic orientation score, chi-square, and token score. OM may be in supervised, unsupervised, or semi-supervised mode [10]. The supervised model uses labeled data. The unsupervised model uses unlabeled data. The Decision Tree (DT), Random Forest (RF), K-Nearest-Neighbors (KNN), Multinomial Naive Bayes (MNB), Gaussian Naive Bayes (GNB), Logistic Regression (LR), and Support Vector Machine (SVM) are the popular and good-performing OM classifiers, proven in past research.

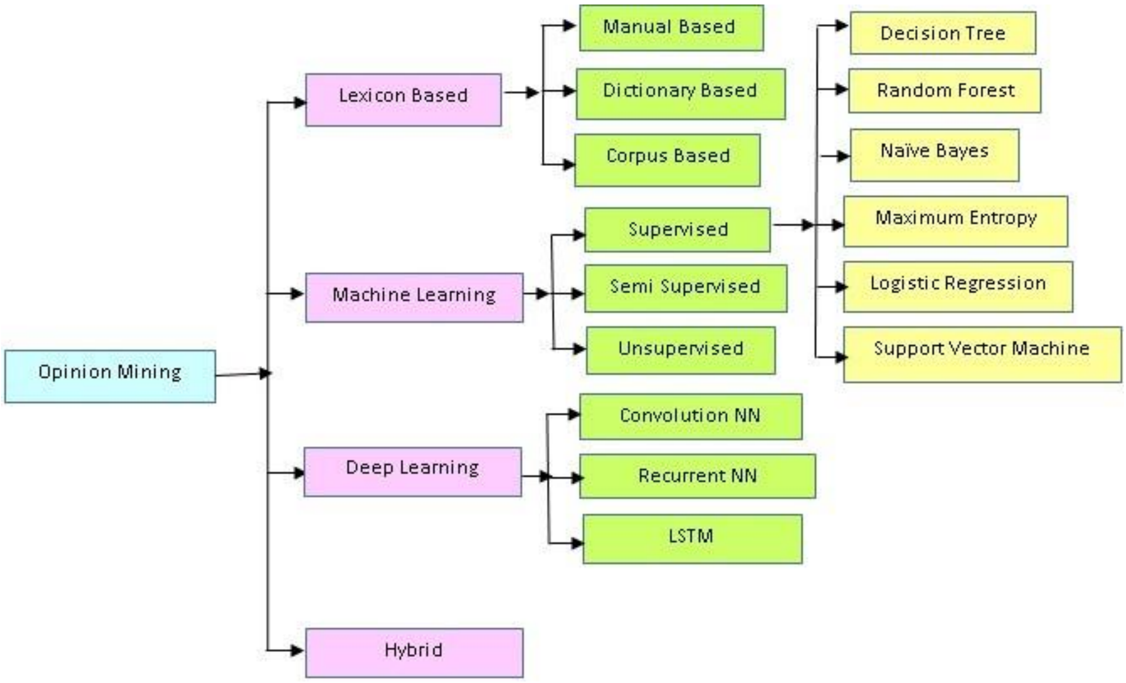

Figure 1. Opinion mining techniques

2.3. Deep Learning Technique

Deep learning (DL) is a specialized form of artificial intelligence given by G. E. Hinton in 2006 [11]. The idea behind it came from the many neurons of the human brain, so deep learning is termed a neural network [12**]**. This network consists of an input layer, many hidden layers, and an output layer [13]. The deep models predict the result based on weight values, data features, functions, and gradients. Forward propagation helps in finding results. Back propagation helps in learning and fine-tuning that result. In many studies, deep learning proved the best strategy for solving real-world applications such as pattern recognition, handwriting recognition, speech recognition, computer vision, image processing, and so on [14].

In deep learning, a large amount of data is a prerequisite for a robust system [15]. It may acquire knowledge from the hidden semantics of a large corpus [16]. The model’s efficacy depends on the datasets [17] and the selection of deep learning algorithms [18]. In deep learning algorithms, feature extraction is in combination with classification. There is no need for explicit feature engineering [19]. The various DL algorithms for opinion mining tasks used by existing research work and are still commonly in use are Long Short-Term Memory (LSTM) , Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Multi-layer Perceptron (MLP), Deep Belief Network (DBN), and Deep Neural Networks (DNN) [20].

2.4. Hybrid Technique

The hybridization of two or more algorithms gives birth to the hybrid model [21]. It is synonymous with ensembling. It results in fast computation, searching, and extended performance [22]. It may bring pinpoints of algorithms. Comparatively, there are fewer research studies on the hybrid approach.

3. Related Work

OM can be a worthy research area due to the large number of languages with rare resources. The study is focused more on regional and code-mixed language. A summarized tabular structure of some research on OM is present in Table 1. Jabreel et al. [23] proposed OM with binary polarity on the

Spanish dataset of the OpeNER and MultiBooked Catalan. OM performed with transfer learning concept. Knowledge gained from the pre-train English model and then smeared to both languages. Spanish embedding has used English embedding. The author has used one-hot encoding, KNN ranking, Adam, nearest neighbour ranking, and softmax activation function. An encoder consists of BiRNN with two LSTMs for forward and backward. The result is 83.7 precision for Spanish and 84.3 precision for Catalan.

Bhadka and team [24] have presented a supervised term rank-based algorithm. Text or voice from a mobile device may be input into the model. The team implements Android technology, PHP web panel, Google Translate API, MySQL, and JSON. Database of 6000 words with respective opinion ranks developed to implement a rank-based method. Ternary classification of opinions carried in multiple languages like English, Gujarati, Hindi, and Marathi. Opinion rank value found for each input word by comparing the database. They found an overall rank value by conjoining all input words. The overall opinion score value outputs a class. They achieved an average accuracy of 89% on the various Gujarat Government schemes dataset. Kaur et al. [25] have generated two new datasets from YouTube API. Datasets belong to the top two cookery channels, Nisha Madhulika and Kabita's Kitchen. Datasets contained 4900 Hinglish comments. The classification was on seven labels: praise, suggestion, gratitude, recipe, video, hybrid, and undefined. The data results in 73.74% accuracy and 75.15% precision in a semi-supervised approach. They created a list of Hinglish stop words and took an English language-based Porter stemmer algorithm. The DBSCAN (density-based spatial clustering of applications with noise) did the clustering. They created a total of 80 clusters with seven labels. The team practiced both parametric and nonparametric machine learning algorithms with experiments using DT, LR, NB, RF, and SVM. SVM linear kernel with TF-IDF vectorizer is best.

The Ternary classification of Hinglish OM was well-thought-out by Singh et al. [26]. They created a model on a dataset of 14000 tweets from the Semeval-2020 competition website. The training set consists of 4102 negative, 4634 positive, and 5264 neutral polarity tweets. Predefined functions have converted the emoticons into text. They created a stop words list and a list of the different spellings of the word. They have considered Doc2Vec, Word2Vec, and FastText embeddings. SVM, KNN, DT, GNB, MNB, LR, and RF were trained and tested under machine learning. The trained models were MLP and CNN under deep learning. An ensemble of SVM, LR, and RF was applied. The study results best with a 69.07 F1-score using an ensemble voting classifier. Out of different vectorizers, the TF-IDF vectorizer gave the best result.

Shah et al. formed two datasets from YouTube cookery channels in Margalish and Devanagari languages [27]. The Margalish dataset consists of 14453 comments, and another dataset consists of 4145 comments. Emoticons were expelled. Multi-layer perceptron and Bernoulli Naïve Bayes implemented on the count vectorizer. The five classification labels were gratitude, recipe, extraneous, hybrid, and suggestion. They achieved a 62.68% accuracy on the Margalish dataset and a 60.60% accuracy on the Devanagari dataset. Kuriyozov et al. [28] prepared a binary class prediction model on the Uzbek with two datasets. An add-up of 100 Uzbekistan applications from the Google Play Store was applied to create a manually annotated dataset of 4300 comments. The MTRANSLATE Google Translate API created the translated dataset with 18485 comments. This study has applied CNN, RNN, LR, SVM, and N-gram. The author guaranteed that annotated Uzbek corpora are the first of this type in the world. The result proved that deep learning does not perform better than machine learning.

The CNN achieved an 88.89% accuracy on a manually annotated set. The LR achieved an 89.56% accuracy on the translated set.

Table 1. Models of opinion mining.

Ref.	Polarity	Language	Dataset	Technique	Result
[23]	Binary	Spanish, Catalan	Spanish OpeNER and MultiBooked Catalan	KNN Ranking Adam, BiRNN with two LSTMs	Spanish :83.7 precision, Catalan: 84.3 precision
[24]	Ternary	English, Gujarati, Hindi, Marathi	24 Agriculture schemes of Gujarat Government	Android technology. PHP web panel , MySQL, JSON	Average Accuracy 89%.
[25]	7 labels	Hinglish	4900 comments from two YouTube cookery channels.	DT,LR,NB-B,NB-G,NB-M, RF, SVM, DBSCAN	Accuracy 73.74%, Precision of 75.15%.
[26]	Ternary	Hinglish	Semeval-2020 competition website, 14000 tweets	Ensemble Voting Classifier, LR	F1-score 69.07
[27]	5 labels	Marglish, Devanagari	YouTube Cookery Chanels,14,453 Marglish and 4145 Devanagari comments	MLP, Bernoulli Naïve Bayes	62.68% Accuracy on Marglish, 60.60% Accuracy on Devanagari
[28]	Binary	Uzbek	100 Google Play Store applications, 4300 size manually annotated dataset and 18485 size translated dataset	CNN,RNN,LR,SVM,	Manually annotated set: CNN 88.89% accuracy, Translated set: LR 89.56% accuracy.
[29]	7 labels	Malayalam -English mix-code.	4291 comments from 2 YouTube cooking channels	BERT, RF, Cross-lingual Language Model (XLM)	XLM accuracy 67.31%. RF accuracy 63.59%.
[30]	Binary	Banglish	2 datasets. One is on customer product review and other is device list data from Wikipedia.5300 labeled data	NER, LSTM	87.99% accuracy in Spacy Custom Named Entity recognition, 95.51% in Amazon Comprehend Custom NER
[31]	Binary	Gujarati	500 movie reviews using Python crawling	MNB, KNN	MNB 87.14% accuracy, KNN 81.43% accuracy
[32]	Binary	Gujarati	Gujarati SentWordNet developed by translating the English synsets with Google translation tool. 3100 opinions	Hybrid with CNN and Rule-based method	75% accuracy

Kazhuparambil et al. [29] identified multiple classes of text data with seven labels: praise, suggestion, gratitude, recipe, video, hybrid, and undefined. All records are labeled manually. The team has generated two new datasets using YouTube API on Malayalam-English mix-code. One dataset of comments belongs to the cookery channel Veena's Curry World, and another belongs to the cookery channel Lekshmi Nair. Each dataset consists of 4291 comments. The study has applied BERT, DISTILBERT, XLM, MNB, KNN, SVM, RF, and DT. The XLM was the top-performing model, with an accuracy of 67.31 %. The result of the random forest classifier was 63.59% accuracy.

Hossain et al. [30] classified Banglish data with binary polarities on Bangladeshi smartphone market demand analysis using Natural language processing (NLP). The classification was on two datasets. One dataset is on customer product review and other is device list data from Wikipedia. The dataset contained 5300 manually labeled data. The edit distance algorithm and levenshtein ratio performed stemming. They have translated Banglish data to Bangla using the Google Cloud Translation API. Then, feed Bangla text into the pre-trained gender prediction model. They implemented both named entity recognition (NER) and LSTM. The spacy custom NER achieved an 87.99% accuracy. The Amazon comprehend custom NER achieved a 95.51% accuracy.

Shah et al. [31] programmed the binary classification of 500 Gujarati movie reviews obtained using Python crawling. The study has implemented MNB and KNN. The MNB model achieved an 87.14% accuracy. The KNN model achieved an 81.43% accuracy. Patel et al**. [**32] made a binary classification model in the Gujarati language in an educational domain. They have collected 3100 opinions from more than 440 users. These opinions renewed into 9371 lines consisting of 4799 positive and 4572 negative lines. The team has developed a Gujarati SentWordNet by translating the English synsets with the Google translation tool. A hybrid model of CNN and the rule-based method was applied. The team has framed the rules for the conjunction and negation words. The result is down with the conjunction rules because of the ineffective scores of some sentences. There is a scope for some more improvements in conjunction rules for future researchers. The study results with 75% accuracy.

We have provided the overall tactic of OM reviewed work with techniques, datasets, and results. Its reasonable learning supports in delineating new research.

4. Challenges and Solutions

In designing an OM algorithm, a researcher meets with several challenges [33]. In the digital era of social media, people actively share their views. It gives birth to a bulk of unstructured data in the network. An unformatted data set results in numerous anomalies [34]. Some challenges may be language-dependent. Some challenges detected during the literature survey are brief below.

- Emoji: In today's growing cohort, people are keenly involved in social media. The public posts their messages hastily using emojis. Emojis are specific graphical symbols that display pertinent ideas or reactions [10]. Scholars have to take care of emoji. Emoji play an active role in conveying sentiments from informal text [35]. NLP provides a way of expanding emoji symbol meaning into words. Several techniques to deal with emojis are using predefined or tailored emoji corpus consisting of emoji and relevant opinion class, emoji embeddings, multimodal approaches with textual and visual emoji features, and deep learning techniques.
- Abbreviations: Social media netizens recurrently use small forms of well-defined phrases or word collections. People use self-made short forms as well. These short forms are entitled abbreviations. For example, before as b4, communicate as communic8. They are some user-made abbreviations. 'ASAP' and 'E.g.' are examples of some well-defined abbreviations. An ASAP stands for as soon as possible. An E.g. stands for, for example. A developer should smartly feed knowledge of these abbreviations into the model during training. In preprocessing, abbreviations expanded into their complete forms. Hence, the model can recognize it well and output the desired result. A programmer may craft his lexicon of abbreviations or may use any pre-defined lexicons. Some abbreviations pointedly contribute to OM [36]. Like LOL, that stands for lots of love and reflects a positive emotion. Contextual analysis harmonizing with NLP techniques, rule-based systems, and machine learning models benefit in attaining the envisioned expansion of the abbreviation.
- Idioms: The media populace interestingly uses idioms in their communication. Idioms are linguistic expressions that specify a meaning. The meaning of an idiom may oppose the meaning of its words. For example, consider the idiom 'break a leg' means good luckiness. It illustrates a negative implication of violence, but its actual meaning of 'good luck' portrays a positive wishing sense. If input text contains idioms, the model should be intelligent enough to understand its meaning. It is challenging for the model to differentiate whether a particular phrase is an idiom or a simple word. To cope with this exception, researchers may use a pre-defined idiom's corpus or create their own corpus of idioms.

- Multiple Words with One Meaning: A lone word can impart many forms to convey a similar meaning. A single core word may generate numerous words in any language. Hinglish may have many similar words, for example: “aaya, aaja, aayi, aayega, aa rahi hai”. All these words have the core word ‘aa’; stemming [37] and lemmatization techniques handle this variety of words. These techniques convert a word into its root word. It helps a model to capture the chief word in data for efficient analysis [38].
- Sense Ambiguity: Sometimes, informal text may give two senses [39]. Let us consider two Hinglish sentences. The first one is “Tum sahi nahi ho, tumhe operation karwa lena chaiye”. The second sentence is: “Tum sahi nahi ho, tumhe aapne aap ko badalna chaiye”. Both sentences have ‘Tum sahi nahi ho’ as a common phrase. The first sentence is in a caring form. The second sentence is in a negative criticism form. Both sentences have a common phrase with them but provide two unlike senses. So, a researcher should understand the authentic context of the sentence prudently [40] and handle it plausibly. The factual meaning of the sentence should not compromised. Though it is an intricate task, state-of-the-art packages, and programming features may resolve it with an amalgamation of linguistic humanoid logic.
- Negation Words: Reversing words like not, neither, nor may obscure a model by opposing the meaning of an input text [39]. For example: “I am not happy”. A model may give a positive label in this sentence by netting a positive word, happy. It is a text with a negative opinion. Therefore, scholars should work on these kinds of situations cautiously. The maximum manual labeling of negating sentences may solve such kind of defy. An alternative way to handle the abnormality is the convention of operative error-free labeling function for programmed labeling.
- Insufficient Resources: A righteous preprocessing will result in an eminence model [20]. During preprocessing, several resource tools and collections are prerequisites. In an OM of a regional language, sometimes a researcher has to suffer from a deficit of desirable sources [41] like scarce sentiment dictionaries, inadequate or no list of stop words and communal words, deficient part of speech tagger, nonappearance of language-specific lemmatizer, and stemming algorithm. Explicit development of resources like user-defined dictionaries, POS taggers, stemming, and lemmatizing algorithms may solve this problem. Some researchers use English language resources for non-English models that may anguish a model's exactitude.
- Wrong Spelling: In the real world, people write informal text with different wrong spellings of a single word. Implementations should handle wrong spellings carefully so that the basic meaning of input text will not change. It requires good conceptual skills. Wrong spellings create noise in data. It tends to perform dilapidation [37]. Let us consider some English words: “pish, piece, peis, peeish”. All these above words represent a single-word piece. Existing spell checkers help in handling such issues. NLP facilitates several good spell checkers. A researcher may develop a spell checker. In some cases, the spell checker also flops. Consider an informal comment: “This is very cut”. A person wants to say, this is very cute. In place of correct spelling cute, cut is typed in. It will engender a neutral label instead of a positive label.
- Word Sequence: The English language follows a structured sentence that consists of a Subject, Verb, and Object called an SVO sequencing structure [42]. In code-mixed language, this sequence structure may change. For example, if we consider an English sentence like “Dhruvin is watching a movie”. The sentence is in the form of SVO. Dhruvin is a subject, watching is a verb, and the movie is an object. But the sentence in Hinglish is: “Dhruvin movie dekh raha hai”. Here, the format is SOV. Dhruvin is a subject, a movie is an object, and ‘dekh’ is a verb. This change in the sequence of the words in a sentence may affect the polarity of the sentence. Therefore, a good knowledge of a relevant input language is needed.

Several challenges during the enactment of OM cannot be restricted to the above only. A scholar may face more challenges. It may include issues of community-created words, different writing styles, pairs of opposite words, and conditional sentences. We have listed only some rampant challenges. It is a complex task to reduce or eliminate all possible anomalies in an opinion mining model.

5. Applications

Opinion mining technique is being used in every work field analysis today. Some prevalent applications of opinion mining are given below.

Table 2. Applications of opinion mining.

Domain	Applications
Medical Science	Find patient mental health and check is a medical emergency required?
	Public reaction analysis to any new medical invention or treatment.
Business Intelligence	Explore customer behavior to improve products and services.
	Structuring marketing maneuvers and customer relationship management.
	E-commerce analytics by acquiring competitor information and market trends.
Recommender Systems	Choosing any entertainment show, service, product, or decision.
Government	Make predictions on the next minister and election.
	Assessing party's work, social issues, applied schemes, public vision, and terrorism.
	Planning of the foreign strategies.
Travel	Planning of the food and hospitality industry.
	Supports visitors in selecting destinations, accommodations, and transportation.
Financial Market	Price prediction by stock exchange, market exploration, and trading.
	Risk management and anticipation of cryptocurrencies.

6. Conclusion

The OM application aids service providers in comprehending their strengths and weaknesses, which helps in deciding the marketing strategies and goals of the concerns. Opinions can be mine on various levels like multi-level, ternary (positive, negative, or neutral), or binary (positive or negative). Researchers may find research gaps with the help of presented challenges and detailed reviewed research work. By fostering notion-building and problem-solving, this study aids developers in opinion-mining for resource-poor languages. Successful pertinent research is proportionate to the maximum number of challenges untangled during OM. Challenges will depend on the dataset, domain, language, tools, and techniques. But, a good researcher should identify possible challenges and try to fix them. This paper presented several relevant challenges with elucidating examples and suggestions. We have investigated numerous OM techniques. NB, SVM, LR, and RF are popular machine-learning techniques. Machine learning has SVM as its most popular technique. Deep learning is popular due to its advantages over machine learning. CNN, RNN, and LSTM are popular techniques. A review also focuses on the varied applications to provide a good understanding of the scope of OM. Fitting a single OM model on different domains and languages is challenging due to the domain and language-specificity. Innovative and new researchers will surely benefit from our study as they build their knowledge and understand the OM models.

References

[1] Torregrosa, J., D'Antonio-Maceiras, S., Villar-Rodríguez, G., Hussain, A., Cambria, E., & Camacho, D. (2022). A Mixed Approach for Aggressive Political Discourse Analysis on Twitter. Cognitive Computation, 1-26.

[2] Alessia, D., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. International Journal of Computer Applications 125(3).

[3] Ahmad, G. I., Singla, J., & Nikita, N. (2019). Review on sentiment analysis of Indian languages with a special focus on code mixed Indian languages. In International Conference on Automation, Computational and Technology Management (ICACTM), pp. 352-356. IEEE.

[4] Mitra, A. (2020). Sentiment analysis using machine learning approaches (Lexicon based on movie review dataset). Journal of Ubiquitous Computing and Communication Technologies (UCCT) 2(03), 145-152.

[5] Taj, S., Shaikh, B. B., & Meghji, A. F. (2019). Sentiment analysis of news articles: A lexicon based approach. In 2019 2nd international conference on computing, mathematics and engineering technologies (iCoMET), pp. 1-5. IEEE.

[6] Al-Tameemi, I. K. S., Feizi-Derakhshi, M. R., Pashazadeh, S., & Asadpour, M. (2022). A Comprehensive Review of Visual-Textual Sentiment Analysis from Social Media Networks. arXiv preprint arXiv:2207.02160.

[7] Rajput, R., & Solanki, A. K. (2016). Review of sentimental analysis methods using lexicon based approach. International Journal of Computer Science and Mobile Computing 5(2), 159-166.

[8] Meng, G., & Saddeh, H. (2020). Applications of machine learning and soft computing techniques in real world. International Journal of Computer Applications & Information Technology 12(1), 298-302.

[9] Yuan, J., Chen, C., Yang, W., Liu, M., Xia, J., & Liu, S. (2020). A Survey of Visual Analytics Techniques for Machine Learning. arXiv e-prints, arXiv-2008.

[10] Anees, A. F., Shaikh, A., Shaikh, A., & Shaikh, S. (2020). Survey paper on sentiment analysis: Techniques and challenges. EasyChair2516-2314.

[11] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[12] Otter, D. W., Medina, J. R., & Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. IEEE transactions on neural networks and learning systems 32(2), 604-624.

[13] Colón-Ruiz, C., & Segura-Bedmar, I. (2020). Comparing deep learning architectures for sentiment analysis on drug reviews. Journal of Biomedical Informatics, 110, 103539.

[14] Dargan, S., Kumar, M., Ayyagari, M. R., & Kumar, G. (2020). A survey of deep learning and its applications: a new paradigm to machine learning. Archives of Computational Methods in Engineering 27, 1071-1092.

[15] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems 32.

[16] Bian, Y., Ye, R., Zhang, J., & Yan, X. (2022). Customer preference identification from hotel online reviews: A neural network based fine-grained sentiment analysis. Computers & Industrial Engineering, 172, 108648. https://doi.org/10.1016/j.cie.2022.108648

[17] Tessore, J. P., Esnaola, L., Lanzarini, L. C., & Baldassarri, S. (2021). Distant Supervised Construction and Evaluation of a Novel Dataset of Emotion-Tagged Social Media Comments in Spanish. Cognitive Computation 14.

[18] Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. Electronics 9(3), 483.

[19] Jin, Z., Tao, M., Zhao, X., & Hu, Y. (2022). Social media sentiment analysis based on dependency graph and co-occurrence graph. Cognitive Computation 14(3), 1039-1054.https://doi.org/10.1007/s12 559-022-10004-8

[20] Solanki, N.N., & Shah, D.B. (2022). A Comparative Assessment of Deep Learning Approaches for Opinion Mining. In Rajagopal, S., Faruki, P., Popat, K. (eds) Advancements in Smart Computing and Information Security. ASCIS 2022. Communications in Computer and Information Science, vol. 1759. Springer, Cham. https://doi.org/10.1007/978-3-031-23092-9_5

[21] Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). Affective computing and sentiment analysis. A practical guide to sentiment analysis, 1-10. https://doi. org/10.1007/978-3-319-55394-8_1

[22] Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., Acharya, & U. R. (2021). ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. Future Generation Computer Systems 115, 279-294.

[23] Jabreel, M. H. A. (2020). Sentiment analysis of textual content in social networks, from hand-crafted to deep learning-based models (Doctoral dissertation, Universitat Rovira i Virgili).

[24] Bhadka, D. B., Shah, D. B., & Patel, N. S. (2019). Mobile Computing Opinion Mining on Agriculture Schemes of Gujarat Government. Wadhwancity: C. U. Shah University.

[25] Kaur, G., Kaushik, A., & Sharma, S. (2019). Cooking is creating emotion: A study on hinglish sentiments of youtube cookery channels using semi-supervised approach. Big Data and Cognitive Computing 3(3), 37. https://doi.org/10.3390/bdcc3030037

[26] Singh, G. (2021). Sentiment Analysis of Code-Mixed Social Media Text (Hinglish). arXiv. https://doi.org/10.48550/arXiv.2102.12149

[27] Shah, S. R., Kaushik, A., Sharma, S., & Shah, J. (2020). Opinion-mining on marglish and devanagari comments of youtube cookery channels using parametric and non-parametric learning models. Big Data and Cognitive Computing 4(1), 3.

[28] Kuriyozov, E., Matlatipov, S., Alonso, M. A., & Gómez-Rodrıguez, C. (2019). Deep learning vs. classic models on a new Uzbek sentiment analysis dataset. Human Language Technologies as a Challenge for Computer Science and Linguistics, 258-262.

[29] Kazhuparambil, S., & Kaushik, A. (2020). Cooking is all about people: Comment classification on cookery channels using bert and classification models (malayalam-english mix-code). arXiv preprint arXiv:2007.04249

[30] Hossain, M. S., Nayla, N., & Rassel, A. A. (2022). Product Market Demand Analysis Using NLP in Banglish Text with Sentiment Analysis and Named Entity Recognition. In 2022 56th Annual Conference on Information Sciences and Systems (CISS), pp. 166-171. IEEE.

[31] Shah, P., Swaminarayan, P., & Patel, M. (2022). Sentiment analysis on film review in Gujarati language using machine learning. International Journal of Electrical and Computer Engineering 12(1), 1030.

[32] Patel, H. H., Patel, B. C., & Lad, K. B. (2022). Opinion Mining of Gujarati Language Text Using Hybrid Approach. United International Journal for Research & Technology 3(4), 105-110.

[33] Jain, A., Jain, G., & Tewari, D. (2024). KNetwork: advancing cross-lingual sentiment analysis for enhanced decision-making in linguistically diverse environments. Knowledge and Information Systems, 1-19.

[34] Farooq, M. U., & Khattak, A. J. (2015) Investigating Highway–Rail Grade Crossing Inventory Data Quality’s Role in Crash Model Estimation and Crash Prediction. Applied Sciences 13(20), 11537 (2023).

[35] Schouten, K., & Frasincar, F.: Survey on aspect-level sentiment analysis. IEEE transactions on knowledge and data engineering 28(3), 813-830.

[36] Kanojia, D., & Joshi, A. (2023). Applications and challenges of SA in real-life scenarios. Computational Intelligence Applications for Text and Sentiment Data Analysis, 49-80.

[37] Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. Information Systems 121, 102342.

[38] Sharma, H. D., & Sharma, S. (2024). Enhancement of the Lexical Approach by N-Grams Technique via Improving Negation-Based Traditional Sentiment Analysis. International Journal of Intelligent Systems and Applications in Engineering, 12(15s), 63-69.

[39] Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems, 226, 107134.

[40] Mulatkar, S., & Bhojane, V. (2014). Sentiment classification in hindi. International Journal of Scientific and Technology Research 3(5).

[41] Ren, Y., Kaji, N., Yoshinaga, N., & Kitsuregawa, M. (2014). Sentiment classification in under-resourced languages using graph-based semi-supervised learning methods. IEICE TRANSACTIONS on Information and Systems 97(4), 790-797.

[42] Kulkarni, D. S., & Rodd, S. S. (2021). Sentiment Analysis in Hindi—A Survey on the State-of-the-art Techniques. Transactions on Asian and Low-Resource Language Information Processing 21(1), 1-46.

Authors

Nidhi N. Solanki

Nidhi N. Solanki is a research student in the field of Computer Science and Technolog at Sardar Patel University, Vallabh Vidyanagar, Anand, Gujarat, India. She is currentl a Veer Narmad South Gujarat university approved Assistant professor in the M. K Institute of Computer Studies, Bharuch, Gujarat, India. With over 13 years c experience, she possesses skills in the Data Science, Deep Learning, Object Oriente Concepts, Database Management System and Artificial Intelligence. She also possesse knowledge of various programming languages.

Dipti B. Shah

Dr. D. B. Shah is a Professor at Post Graduate Department of Computer Science & Technology, Sardar Patel University, Vallabh Vidyanagar, Anand, Gujarat, India. Sh is working as a faculty in the same department since 1989. She has guided; 16 researc scholars for Ph. D in Computer Science and 3 for M.Phil. Her publication includes books and more than 150 research papers in international and national level journals Her area of interest is Image processing, Multimedia and Medical Informatics.