DOI: http://dx.doi.org/10.26483/ijarcs.v16i6.7370

Volume 16, No. 6, November-December 2025

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

SEMANTICALLY ENRICHED DOCUMENT – LEVEL SENTIMENT ANALYSIS: A COMPREHENSIVE STUDY

C. Indrani, V. Dharani, Dr. P. Kalarani
Assistant Professor, Department of CT and IT
Kongu Arts and Science College (Autonomous)
Erode, Tamilnadu, India

Abstract: In the era of digital communication, vast amounts of textual data are generated daily through social media, reviews, blogs, and forums. Extracting meaningful insights from this unstructured data presents both opportunities and challenges. Sentiment analysis sometimes referred to as opinion mining, it is a branch of Natural Language Processing (NLP) that focuses on recognizing and classifying opinions in written material in order to ascertain if the author has a positive, negative or neutral attitude toward a certain subject. Document-level sentiment analysis is a vital task within Natural Language Processing (NLP) that aims to determine the overall sentiment expressed in an entire document, rather than individual sentences or phrases. Unlike sentence-level or aspect-level sentiment analysis, the document-level approach evaluates cumulative sentiment to understand the general opinion conveyed by the author. This method is crucial in domains such as product reviews, movie critiques, customer feedback and news articles, where the holistic sentiment of the text is more informative than isolated fragments. This study explores various methodologies for document-level sentiment classification, starting from traditional machine learning models such as Naïve Bayes, Support Vector Machines (SVM) and Random Forests, which rely on bag-of-words, TF-IDF and n-gram features. While these models offer baseline performance, they often fail to capture the deeper context and inter-sentence dependencies present in longer texts.

Keywords: Document-Level Sentiment Analysis, Natural Language Processing, Machine Learning, Support vector Machine, Opinion Mining, Random Forest.

1 INTRODUCTION

Sentiment analysis is concerned with the automatic extraction of sentiment information from natural language text. The word sentiment is generally defined as a thought, attitude or judgment promoted by feelings. It is also defined as a specific view or notion: opinion and emotion. An individual usually asked his/her friends or family for opinions before making a decision and an organization normally conducted opinion polls, surveys and focus groups to find out the sentiments of the general public about its product or services. Due to the development of technology, internet is widely used where people can express their opinions and emotions by posting reviews of products or services.

Especially with the explosion of Web 2.0 platforms such as blogs, Twitter, Facebook and various other types of social media, an individual has unprecedented channels and powers by which to share his/her opinions and brand experience regarding any product or service. Moreover, the organizations can modify their marketing strategies through social media monitoring and analysis. However, it can still be difficult task to find out opinion sources and monitoring them on the World Wide Web, because there is large number of diverse sources such as discussion groups, online forums and blogs.

In sentiment analysis, text is classified according to the following criteria:

- ➢ The polarity of the outcome.
- ➢ The polarity of the sentiment expressed as positive, negative or neutral.
- ➢ Pros and cons
- ➢ Agree or disagree with a topic
- ➢ Support or opposition
- ➢ Good or bad news

2. LEVELS OF SENTIMENT ANALYSIS

Sentiment analysis has been mainly investigated at three different levels [PRI16] are,

- ➢ Document level sentiment analysis
- ➢ Sentence level sentiment analysis
- ➢ Aspect level sentiment analysis

The following figure 1.1. shows the level of sentiment analysis.

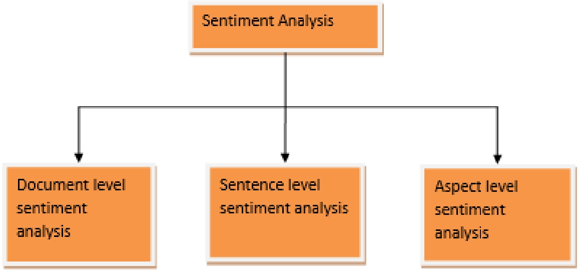
Sentiment Analysis
Document level sentiment analysis
Sentence level sentiment analysis
Aspect level sentiment analysis

Figure 1.1 Level of Sentiment Analysis

2.1 Document level sentiment analysis

The process of document level sentiment analysis [1] is to find out whether an opinionated document which express public opinions as overall positive or negative opinion. For an instance, a system for sentiment analysis classifies the overall polarity of a customer review about a specific product. Such type of sentiment classification assumes that one document expresses opinion on a single object, such as customer reviews of services or products, because usually the result of sentiment analysis only has two outputs such as positive or negative or three outputs such as positive, negative or neutral. However, it is common that there might be a few different opinions in one document, hence it is not applicable to documents in which opinions are expressed on multiple products. There are various researches have been carried out on document level sentiment analysis. The main intend of document level sentiment analysis is how to separate the positive texts form negative texts automatically. Due to the simple output of sentiment classification, the major disadvantage of document-level sentiment analysis is the lack of in-depth analysis.

2.2 Sentence level Sentiment analysis

Sentence-level sentiment analysis focuses on identifying the sentiment expressed in individual sentences. It aims to determine whether a sentence conveys a positive, negative, or neutral emotion. This level of analysis is particularly useful when a text contains multiple opinions or viewpoints about different aspects, as it allows for a more fine-grained understanding of sentiment. The sentence level sentiment analysis process consists of two sub tasks where one of the tasks determine whether the sentence is a subjective sentence or objective sentence and another task determines whether sentences express positive or negative opinion. This level is related to subjectivity classification which is to differentiate the subjective sentences that express sentiments or views from objective sentences that expresses factual information. The subjectivity classification filtered out those sentences which contain no opinions. But the sentence level classification process considers that one sentence expresses a single opinion from a single opinion holder.

2.3 Aspect level Sentiment analysis

Classifying opinions at document level or sentence level is useful in many cases, but they are insufficient to provide necessary details needed for applications, because they do not identify sentiment targets or assign opinions to these targets. The aspect level sentiment analysis [3] focused on opinions itself instead of looking at the constructs of documents, such as paragraphs, sentences and phrases. Identifying the opinion targets is also essential along with the determination of polarity of the opinions. The task of aspect extraction can also be seen as an information extraction task, which aims to extract the aspects that opinions are on. The basic approach of extracting aspects is finding frequent nouns or noun phrases, which are defined as aspects. Then the text containing aspects are classified as positive, negative or neutral.

3. WORKFLOW OF DOCUMENT - LEVEL SENTIMENT ANALYSIS

In general, sentiment analysis tries to figure out the sentiment of a writer about the specific aspect and also the overall contextual polarity of a document. A core issue in this field is an opinion classification where a review is classified as a positive, negative or neutral evaluation of a subjected object. The assessment of sentiment can be done in two ways are direct opinions and comparisons. The direct opinions give positive, negative or neutral about the product directly. The comparison means to compare the subject with any other similar objects. The workflow of the sentiment analysis is depicted in figure 1.2.

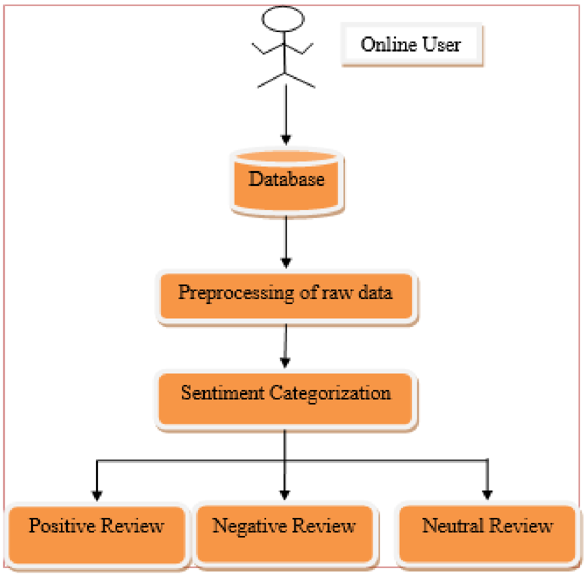

Figure 1.2 Workflow of Sentiment Analysis

The views are being extracted from writer reviews over their comment. Preprocessing phase has been divided into number of sub phases are tokenizing, normalization, feature selection and feature reduction mechanism. Tokenization is the process to split the reviews into tokens by removing whitespaces, commas and other symbols etc. Feature extraction phase identifies the types of features for sentiment analysis. Feature selection and feature weighting mechanism are select good features for sentiment classification and assigns weights for each feature for good recommendation. The features are optimized in reduction mechanism for classification process.

3.1 Types of features used for document level sentiment analysis

There are different types of features [2] are used for sentiment analysis. Some of them are given below:

- Term Frequency: It represents the presence of term in a document carries a weight age.
- Part of speech (POS) information: POS tagger is used to separate POS tokens.
- Opinion words: It expressed the positive, negative and neutral emotions.
- Negations: It represents the shift sentiment orientation in a sentence.
- Term co-occurrence: This feature occurs together like unigram, bigram or n-gram. N-gram is the continuous items of n items from a given review. An n-gram of size 1 is called as unigram and n-gram of size is 2 is called as bigram.
- Syntactic dependency: It is represented as a parse tree and it contains word dependency-based features.

4. FEATURE SELECTION METHODS

Based on the different types of features various feature selection methods [ALI14] such as Document Frequency (DF), Term Frequency-Inverse Document Frequency (TF-

IDF), Information Gain (IG), Mutual Information, chi – square and Principal Component Analysis (PCA) are available which are described as follows.

- Document Frequency

In the document frequency-based feature selection method, features are ordered by document frequency for each feature in the whole document. This method is a simple measure for feature reduction and has a linear time complexity to scale a large dataset.

- Term Frequency-Inverse Document Frequency

TF-IDF is defined by multiplying value of number of times a word occurred in review (TF) and the number of times a word occurred in whole corpus (IDF).

$$TF-IDF_i = t_{i,j} \times \log\left(\frac{N}{df_i}\right) \qquad (1.1)$$

where, $TF-IDF_i$ is the weight of a term i, $t_{i,j}$ represents the frequency of term i in review j, N denotes the total number of samples in the corpus and df_i is the number of samples containing term i.

- Information Gain

One of the most widely used feature selection measure is the Information Gain (IG) in the area of sentiment analysis. It determines the most discriminative features to predict review by analyzing the absence or presence of feature in document.

$$IG(F) = -\sum_{x=1}^{M} P_r(c_x) \times \log P_r(c_x) + P_r(f) \times \sum_{x=1}^{M} P_r(c_x|f) \times \log P_r(c_x|f) + P_r(\bar{f})$$
$$\times \sum_{x=1}^{M} P_r(c_x|\bar{f}) \times \log P_r(c_x|\bar{f}) \qquad (1.2)$$

where, f denotes a feature, $P_r(f)$ indicates the probability situates feature, $P_r(\bar{f})$ means the probability does not situate a feature, $P_r(c_x)$ denotes the situate of class c_x and M denotes the number of categories.

- Mutual Information

It is the process of determine the features which are not uniformly distributed across sentiment classes and selected those features. The non-uniformly distributed features are informative of their classes and MI gives more importance to only few terms.

$$MI(F,C) = \sum_{c\in C}\sum_{f} P(f,c) \log \frac{P(f,c)}{P(f)P(c)} \qquad (1.3)$$

where, P(f,c) represents joint probability distribution function, P(f) and P(c) represent marginal probability distribution of f and c and c is positive and negative classes.

- Chi-Square

Chi-Square measures observed count and expected count and analyzed how much deviation occurs between them.

$$\lambda^2(f,c) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \qquad (1.4)$$

where, A, B, C, D represents the frequencies which denotes the presence of absence of feature in the sample, A is the count of samples in which feature f and c occurred together, $N = A + B + C + D$, f denotes the feature and c represent the class.

5. SENTIMENT CLASSIFICATION APPROACHES

The sentiment classification [VIM16; JAY13] is the process of classifying the sentiments as positive, negative and neutral. Mainly sentiment classification approaches are classified into two categories are supervised learning approach and unsupervised learning approach. Figure 1.4 depicts the sentiment classification approaches.

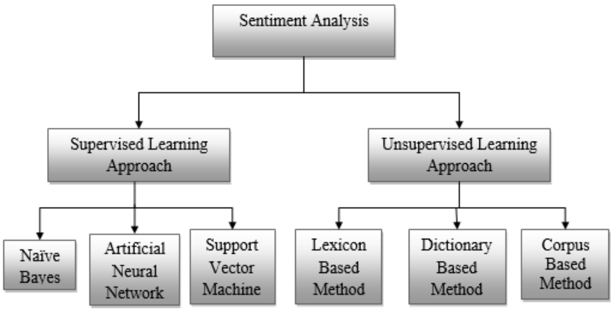

Figure 1.3 Sentiment Classification Approaches

5.1 Supervised Learning Approach

Supervised learning [5] is the machine learning task of inferring a purpose from labeled training data. The training data consists of a set of training examples. The unsupervised learning consists of each example which is a pair consisting of an input object and a desired output value. A supervised learning algorithm studies the training data and produces an inferred function which can be used for mapping new examples. In other words, supervised classification algorithms are rule based classifier, decision tree, probabilistic classifier and linear classifier. These algorithms are based on labeled dataset which is given as input to train the model. It generates output by applying the trained model to test data. Machine learning based sentiment classification consists of two steps. The features are extracted and stored in feature vector which is done in the first step. Then in the second step, machine learning train feature vectors by using classification algorithms. Some of the commonly used supervised learning algorithms for sentiment classification are given as follows:

- Support Vector Machine (SVM)
- Artificial Neural Network (ANN)
- Naïve Bayes (NB)

a) Support Vector Machine

Support Vector Machine (SVM) [1] examines the data, identifies hyperplane that classify the data into two classes with maximum margin. SVM also supports classification and regression in statistical learning. A separate hyperplane is written as:

$$W \times X + b = 0 \qquad (1.5)$$

where, $W = \{w_1, w_2, w_3, \dots w_n\}$, w_n is defined as weight vector of n attributes, b denotes the bias. If the value of hyperplane is greater than 0, then the points are classified as positive category. If the value of hyperplane is lesser than 0, then the points are classified as negative category. If the hyper plane value is equal to 0, then all points are perpendicular to W. A large penalty is assigned to errors or margin errors when the margin value is large. If margin value is small then some points become margin error and orientation of hyperplane is changed.

$$W = \sum_j \alpha_j c_j d_j , \alpha_j \geq 0 \qquad (1.6)$$

where, α_j is the weight of the training samples, c (-1,1) is class (positive, negative) for document d.

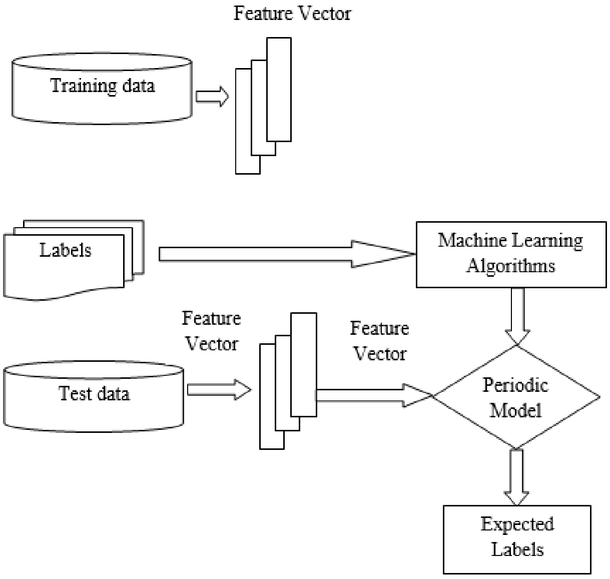

Figure 1.4. Overview of Supervised Learning Algorithm

b) Artificial Neural Network

Artificial Neural Network (ANN) [9] is typically represented by a network diagram which is composed of nodes connected direct links. Nodes are arranged in layers and the structure of the most used neural network consists of three layers are an input, a hidden and an output layer of nodes. It is also classified as a feed forward network, since the nodes are connected only in one direction. Each connection has an associated weight, whose value is estimated by minimizing a global error function in gradient descent training process. Generally, a neuron is a simple mathematical model that produces an output value in two steps. First, the neuron output computes a weighted sum of its input and then applies an activation function to this sum to derive its output. The activation function is typically a non-linear function and it ensures that the entire network can estimate a non-linear function which is learned from the input data.

c) Naïve Bayes

Naïve Bayes [8] is utilized for prediction of a given tuple to belong to a particular class. It is used because of its easiness in both during training and classifying steps. Pre-processed data is given as input to train input set by classifier using naïve bayes and that trained model is applied on test data to generate either positive or negative sentiment. The bayes theorem is given as follows:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \qquad \textbf{(1.7)}$$

where, H is the hypothesis, X denotes the tuples, P(H|X) denotes posterior probability of H conditioned on X i.e., the Probability that a hypothesis holds true given the value of X, P(H) represents Prior probability of H i.e., the probability that H holds true irrespective of the tuple values, P(X|H) represents posterior probability of X conditioned on H i.e., the probability that X will have certain values for a given hypothesis, P(X) represents prior probability of X i.e., the probability that X will have certain values.

5.2 Unsupervised Learning Approach

Unsupervised learning approaches [7] for sentiment classification can solve the problem of domain dependency and reduce the need for training data. There are no labels are provided in the unsupervised learning approaches. These approaches are departure it on its individual structure in its input. It explained the private structure from unlabeled data. Because the instance is provided to the learner are unlabelled, there is no faults or reward to estimate a possible result. The unsupervised learning is nearly related to the troubles of density evaluation in statistics. On the other hand, it also relates with the greater number of techniques that seeks to recapitulate and describe the key features of the data. The structural design [OBU17] of the unsupervised learning is demonstrated in Figure 1.6.

Some of the unsupervised learning algorithms for sentiment classification are listed as follows:

- ➢ Lexicon based approach
- ➢ Dictionary based approach
- ➢ Corpus based approach

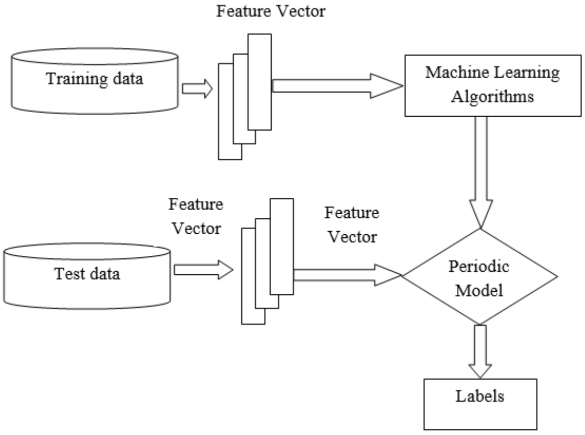

Figure 1.5. Overview of Unsupervised Learning Algorithm

a) Lexicon based approach

The lexicon-based sentiment classification approach [7] is further splitted into two categories are corpus-based sentiment classification approach and dictionary-based sentiment classification approach. Corpus based sentiment classification approach is classified as semantic and statistical approach. In semantic approach, terms are represented in semantic space to discover the relation between terms. The statistical approach identifies the sentiment based on the calculation of co-occurrence of words. The sentiments are identified in dictionary-based approach by using synonym and antonym of words from lexical dictionary like WordNet. The sentiment classification accuracy is improved by including WordNet, word sense disambiguation, and semantics which help to find Synset of words.

WordNet

The similarity between words is checked and the sentiment score are calculated by using WordNet [7]. It links to sets of syntactic categories which are noun, adjective, verb and adverb and these are linked with semantic relations those are called as entailment, hyponymy, troponymy, synonym, antonym, meronymy, etc. From the feature vector list the WordNet find out synonym of features and those are compared with each feature of feature list. With the help of WordNet adjectives are scored, it helps for classification of polarity. WordNet has 166000 above sense pairs and word form. A sense is represented by set of synonyms and forms are represented by string of ASCII characters.

b) Dictionary based approach

The sentiment words are compiled by using dictionary-based approach [6] is an obvious approach because most dictionaries like WordNet have list of synonyms and antonyms for each word. In this approach, a simple technique is used for sentiment classification. A few seed sentiment words are used in this technique to bootstrap based on the synonym and antonym structure of a dictionary. Initially in this approach, a small set of sentiment words called as seeds with known positive or negative orientations are collected manually which is very easy. Then this algorithm grows the collected set by searching in the WordNet or another online dictionary for their synonyms and antonyms. The latterly found words are included to the seed list. Then the next iteration begins. The iteration process is continued until no more new words can be found. After the process completes, a manual inspection step is included to clean up the list.

c) Corpus based Approach

The corpus-based approach [9] matches lexicon opinion with the list of opinion words that identifies correct domain or context of specific opinion. The opinionated lexicons are obtained in two step process. In the first step, the word which contain opinion in the corpus are identified and then in the second step, polarity is assigned to opinionated word.

6. WORD SENSE DISAMBIGUTION

Word Sense Disambiguation (WSD) [8] is the task of identifying the meaning of words in context in a computational manner. It is often used in sentiment analysis to disambiguate the sense of word. Lesk algorithm is one of the word sense disambiguation algorithms which compares the dictionary definitions of two terms and the appropriate sense for each word is determine by the frequency of common words between the definitions. The extended gloss overlap has its foundations on the original lesk algorithm. The extended gloss overlap algorithm gives a higher score to phrasal matches instead of treating glosses as a bag of terms regardless of word order. Glosses of related words obtained through WordNet also play a role in determining the sense of a term. Using machine learning features for word sense disambiguation, terms are labeled as subjective or objective depending on how it is used in a sentence and whether it has an impact on overall polarity of a sentence.

CONCLUSION

Document-level sentiment analysis plays a pivotal role in understanding the overall emotional tone of lengthy text content such as articles, reviews, or social media posts. By evaluating sentiment at the document level, organizations and researchers can capture the dominant sentiment expressed in entire documents, which is especially useful for summarizing opinions, detecting trends, and informing decision-making processes. It is a valuable tool for large-scale sentiment monitoring, particularly when high-level emotional insight is more important than granular interpretation. Future improvements may come from integrating it with sentence and aspect-level analyses to provide a more comprehensive understanding of sentiment across diverse textual sources.

REFERENCES

[1] Maryem Rhanoui, Mounia Mikram, Siham Yousi, and Soukaina Barzali. 2019. A CNN-BiLSTM model for documentlevel sentiment analysis. Machine Learning and Knowledge Extraction 1, 3 (2019), 832ś847.

[2] Ramadhani, A.M.; Goo, H.S. Twitter sentiment analysis using deep learning methods. In Proceedings of the 2017 IEEE 7th International Annual Engineering Seminar (InAES), Yogyakarta, Indonesia, 1–2 August 2017; pp. 1–4.

[3] Vimali, J.; Murugan, S. A Text Based Sentiment Analysis Model using Bi-directional LSTM Networks. In Proceedings of the 2021 IEEE 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 8–10 July 2021; pp. 1652–1658.

[4] Hannah Kim and Young-Seob Jeong. 2019. Sentiment classiication using convolutional neural networks. Applied Sciences 9, 11 (2019), 2347.

[5] L. C. Chen, C. M. Lee, and M. Y. Chen, "Exploration of social media for sentiment analysis using deep learning," Soft Computing, vol. 24, no. 11, pp. 8187–8197, Jun. 2020.

[6] M. Arbane, R. Benlamri, Y. Brik, and A. D. Alahmar, "Social media-based COVID-19 sentiment classification model using Bi-LSTM," Expert Systems with Applications, vol. 212, Feb. 2023, Art. no. 118710.

[7] R. Sanjana, C. Tandon, P. J. Bongale, T. M. Arpita, H. Palivela, and C. R. Nirmala, "Comparative Analysis of Various Language Models on Sentiment Analysis for Retail," in Soft Computing for Problem Solving, Singapore, 2021, pp. 725–739.

[8] S. Halder, "Tokenization, Stemming and Lemmatization | TechGenizer," Mar. 16, 2021. https://techgenizer.netlify.app/blog/2021/03/16/tokenization-stemming-lemmatization/.

[9] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis," Informatics, vol. 8, no. 4, Dec. 2021, Art. no. 79.

[10] B. Ahmed, G. Ali, A. Hussain, A. Baseer, and J. Ahmed, "Analysis of Text Feature Extractors using Deep Learning on Fake News," Engineering, Technology & Applied Science Research, vol. 11, no. 2, pp. 7001–7005, Apr. 2021.