International Journal of Engineering & Technology, *6 (4) (2017) 108-115*

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v6i4.8295
Review Paper

Perspectives of the performance metrics in lexicon and hybrid based approaches: a review

Meesala Shobha Rani [1], Sumathy.S [2] *

[1] *Research Scholar, School of Computer Science and Engineering VIT University, Vellore-632014, India*
[2] *Associate Professor, School of Information Technology and Engineering, VIT University, Vellore-632014, India*
**Corresponding author E-mail: ssumathy@vit.ac.in*

Abstract

Online social media and social networking services experience a drastic development in the present scenario. Contents generated by hundreds of millions of users are used for communication in general. Users mark their opinion and review in various applications such as Twitter, Facebook, YouTube, Weibo, Flicker, LinkedIn, Online-e commerce sites, Microblogging sites, etc. User generated text is spread rapidly on the web, and it has become tedious to analyze the opinionated text in order to arrive at a decision. Sentiment analysis, a sub-category of text mining is the major active research domain in current era due to greater quantity of opinionated text present in the Internet. Semantic detection is the sub-class in the sentiment analysis which is used for measuring the sentiment orientation in any text. Opinionated text is used for analyzing and making the decision simple. This interdisciplinary field draws various techniques from data mining, machine learning, natural language processing, lexicon based and hybrid based approaches. This paper provides a broad perspective with the highlight of the current state-of art techniques emphasizing the various research challenges and gaps present. The performance metrics in terms of detection rate, precision, recall, f-measure/score, average mean, auto-Pearson correlation, cosine similarity and ratio of time on various algorithms is discussed in detail. An analysis of the text mining approaches in different domains is presented.

Keywords: *Corpus Based Approach; Dictionary Based Approach; Hybrid Based Approach; Lexicon Based Approach; Sentiment Analysis; Text Mining.*

1. Introduction

With the tremendous reliance of social media on web, users show keen interest to state their opinions and reviews on various applications in terms of product pricing, forecasting elections, competitive intelligence, national relationship analysis, market stock prediction and risk detection in banks, healthcare and industrial domains [1]. The data present in the Internet is in a textual and unstructured form. Text mining is a field of study that converts unstructured data to structural data using knowledge discovery, information extraction and retrieval [2]. Pre-processing is a stage to convert unstructured textual data into raw data after removing noisy data. For organizing the raw data, tokenization, stemming, stops word removal, lemmatization and parts of speech (POS) are used.

Sentiment analysis, also called as opinion mining is a field used in web mining, data mining, text mining, machine learning and natural language processing. Opinion mining is the field to study people's opinion, sentiments, attitudes, emotions, and evaluations towards certain entities such as products, services, organizations, individuals, issues, topics, events and their attributes [3]. Sentiment analysis is widely used for classification (Positive, Negative and Neutral), clustering and categorization of text.

Sentiment analysis is defined as an opinion in a set (e_i a_{ij}, S_{ijkl}, h_k, t_l), where e_i is the term of the i^{th} entity; a_{ij} is the j^{th} attribute of entity e_i ; h_k is the K^{th} opinion holder ; S_{ijkl} is the opinion on i^{th} the entity on j^{th} attribute by k holder at t_l time; t_l is the time at which opinion is given by the k^{th} holder. For example in "The screen of the laptop was good", screen is the aspect of entity laptop and overall positive sentiment opinion is expressed [1, 3].

Figure 1 presents various text mining approaches in sentiment analysis. Sentiment analysis approaches are categorized as machine learning based approaches, lexicon based approaches and hybrid based approaches. Machine learning (ML) algorithms are widely used in sentiment analysis. Machine learning algorithms give better classification performance, and it takes more time to compute. They are broadly classified into three categories such as Supervised based learning approach, Unsupervised based learning approach and Semi-supervised based learning approach. To overcome the limitations of machine learning approaches, lexicon based approaches are used to speed up the process. It does not require any annotated corpus and training data.

This paper mainly focuses on a lexicon based approaches and hybrid based approaches on sentiment analysis in different domains. Lexicon based approaches depend on domain knowledge and co-occurrence of words. A Lexicon is a group of vocabulary of sentiment words used to determine sentiment polarity and strength of the sentiment word [4]. Lexicon based approach is categorized into dictionary and corpora based approaches. Dictionary based approach uses the domain knowledge such as Ontology. Corpus based approach determines the similarities of words based on the co-occurrence of words in the document [5]. Hybrid based approach is an integration of machine learning based approaches and lexicon based approaches.

Copyright © 2017 Meesala Shobha Rani, Sumathy.S. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

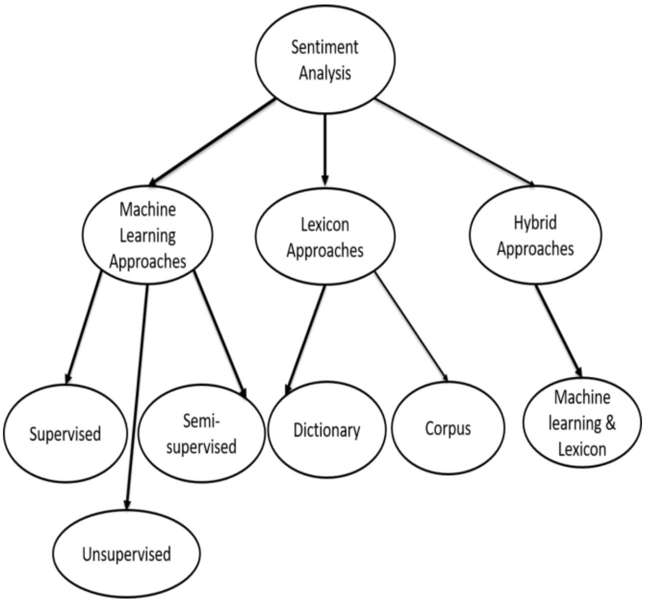

Fig. 1: Sentiment Analysis Approaches.

The paper is organized as follows: first section presents the introduction, next section presents the lexicon based approaches which contain dictionary based and corpus based approaches. The hybrid based approaches are described in the third section followed by the comparative analysis and discussion in the fourth section. Open issues and challenges identified from the existing literature are presented in section five followed by conclusion in sixth section.

2. Lexicon based approaches

Lexicon based approach is categorized into two types such as dictionary based approach and corpus based approaches.

2.1. Dictionary based approach

Dictionary based approach uses the prior domain knowledge. Unsupervised parsing based polarity determination which combines natural language processing and sentiment features from sentiment lexicons are presented [6]. This approach is compared with baseline algorithms that determine positive, negative and neutral opinion. It is a real and complex problem in order to extract noise and maintain the semantic and sentics from media, leveraging multi-model framework [7]. The proposed approach focuses on event summarization and concept-level sentimental analysis.

A sentiment embedding logic based on context and similarity of the sentiment text is presented [8]. Sentiment embedding is applied to word level, sentence level and in constructing sentiment lexicons. The hybrid approach is compared with BL-Lexicons, MPQA, NRC-Lexicon, Hashtag Lex, Sentiment140Lex, SE-Pred, SE-Rank, SE-HyPred, and SE-HyRank. Their proposed method focuses on sentiment embedding, word embedding and building lexicon feature embedding.

An ensemble classifier technique for semantic short text opinion mining is proposed [9]. This method is used for short text or tweets and opinion detection. The technique is performed and compared with Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Gaussian Naïve Bayes, Linear SVM, Radial SVM, Polynomial SVM, Decision trees, K-NN and Logistic regressions. An ensemble method is proposed and is used for model selection and classification. The final weights w_j (degree of confidence) are calculated for the model selection.

$$w_j = \frac{1}{|\log_2\left(\frac{Acc_j}{\max\left(Acc_j\right)+T}\right)|}, 1 \le j \le n \tag{1}$$

Where 0<T<1 is used to control the higher and lower weights. If the value of T is small, then the change between weights of the classifier with low and high accuracy is more. As T reaches to zero, the variance between the voting weights w_j of the classifier with the greater accuracy (Acc_j = max (Acc_j)) and all other possible classifiers with (Acc_j < max (Acc_j)) is higher. If T is one, the difference in the voting of all classifiers is small.

Web mining, semantics and entity relations are discussed [10]. Sentiment detection in entity and tweet level are explained [11]. Their proposed methodology uses Senticircle and lexicon based approaches for sentiment analysis on twitter data. Senticircle is used to update and assign the polarity and strength of the values to the words. Experimental analysis presents that their approach outperforms several baseline algorithms. This approach is used for runtime analysis to validate scalability, build and release STS gold dataset used for calculating both entity and tweet sentiment analysis. The contextual sentiment value is computed using the Senti-median metric. It is used to calculate new sentiment score in the Senticircle as follows.

$$g = arg \min_{g\epsilon\mathbb{R}^2} \sum_{i=1}^{n} \left|\left|p_i - g\right|\right|_2 \tag{2}$$

Where p is the set of n points (p_1, p_2….p_n) in the Senticircle in the geometric median g(x_k,y). If the Euclidean distance of all points p_i is minimum, then it is termed as Senti-median. The Senti-median captures the sentiments in y coordinates and strength in x coordinates of Senticircle in terms of m. The entity-level sentiment detection is calculated by the following equation.

$$\mathcal{L}(g_e) = \begin{cases} -ve & if\ y_g < -\lambda \\ +ve & if\ y_g > +\lambda \\ neutral & if\ \left|y_g\right| \le \lambda\ \&\ X_g \ge 0 \end{cases} \tag{3}$$

Where, Sentiment median g_e of an entity is e_i , $\mathcal{L}$ is the entity sentiment function. λ is the threshold which defines the y-axis neutral border region. If g lies in the neutral region then the entity has neutral sentiment. If g lies in the +ve quadrants, then the entity has +ve sentiment and if g lies in the -ve region then the entity has -ve sentiment. The tweet-level sentiment detection is calculated using the following equation such as

$$\hat{s} = arg \max_{s\epsilon S} \mathcal{H}_s(p) = arg \max_{s\in S} \sum_{i}^{N_p} \sum_{j}^{N_w} \mathcal{H}_s(p_i,\ w_j) \tag{4}$$

Where pivot method is used to identify the sentiment $\hat{s}$ which obtains the maximum sentiment influence in the tweets. $s\epsilon S$ is the positive, negative and neutral sentiment label, p is the vector of all pivot terms in a tweet. N_p and N_w are the set of pivot terms in the tweets. $\mathcal{H}_s(p_i,\ w_j)$ is the impact of sentiment function, which yields the impact $\hat{s}$ of term w_j in the pivot term $p_{i.}$

A semantic framework for textual data is used as sustained tool for recommender systems. This approach combines and uses different NLP tasks to measure the similarity, semantic analysis and rate of the sentiments are discussed [12]. A methodology to identify emotions in online customer reviews in order to determine the various emotion dimensions in various different products are analysed. This approach utilizes the mining of emotion lexicon from the emotion terms. It is used to build the model based on review quality. Review quality is measured based on the online customer's helpfulness rating. The methodology compares two emotional dictionaries, one with crowd funded lexicon and other created by experts. The crowd funded lexicon outperforms and gives the classification precision rate. The quality of the review is measured by helpfulness score with the following equation[13].

$$h_r = \log_{10} \frac{x_r+1}{y_r-x_r+1}\ for\ x_r \leqq y_r \tag{5}$$

Where the helpfulness score h_r of each review is defined as the logarithmic ratio between the number of times the positive reviews are voted (x_r) and the number of times the negative review are voted (y_r-x_r). If the h_r score is higher, the review is more helpful. +1 is added in the denominator to avoid division by 0 and extended to the numerator. If no votes are received the helpfulness score is assigned as (x_r=y_r=0) and h_r=0.

A new methodology for developing the links between source and target domain using user emotion profiles and items is presented [14]. The cross domain recommendation uses emotion lexicons to determine the emotions in the target domain. The proposed approach is evaluated and performed on movie lens and book datasets which gives an improvement of 28% and F1-measure of 71.1% as compared with the existing semantic clustering approaches. Table 1 presents an analysis on various methods in dictionary based approaches and their performance with the current state-of-art techniques.

Table 1: Various Methods on Dictionary Based Approaches

Approaches	Domain	Advantages	Limitations	Results
[6] Sentiment propagation includes intensification, modifier, negation, adversative/ Concessive	Cornell Movie Review, Obama-McCain Debate, SemEval-2015	Combination of ML and lexicon gives better sentiment classification	Developing different linguistic aspects in same domain gives poor performance	Accuracy 66.59%
[7] Event builder and Event sensor, Sentic-Net3	YFCC100M	Used for event and knowledge summarization	It is monotonous and time consuming. Recall is reduced compared to the baseline techniques	Precision 68.2%, Recall 70.2%, F-Measure 69.4% Cosine Similarity 83.2%
[8] Word level and, Sentence level sentiment Analysis, Building Lexicon level sentiment classification	SemEval from twitter data, Rotten Tomatoes	Sentiment embedding provides outstanding results compared to other baseline sentiment models. Hybrid models outperform the baseline algorithms	Word embedding is very hard to cover all words. Re-embedding of sentiments require more time and slightly degrades the classification performances. Needs to improve ternary classification	Word and Sentiment embedding (SE-HyPred and SE-HyRank) gives better results. Lexicon features append SE-HyPred Lexicon- 81.7% on 2013Test , SE-HyPrank Lexicon - 81.9% on 2014Test
[9] Ensemble method integrates text pre-processing , text normalization, semantic indexing techniques	Social media, Real Tweeter datasets	Increases the performance and reduces the redundancy and irrelevant features. Mainly used for short text classification	Very expensive and requires more computational power. Doesn't perform well on offline analysis. Not suitable for multi-class classification problems.	F-Measure on UMICH dataset is 96.9%
[10] Temporal semantic relations integrate connection entity, lexical syntactic patterns, context sentences, context graph, context communities	LinkedIn dataset, Movie star dataset	Does not contain any domain knowledge for mining	Choosing proper knowledge base and computational method is difficult. Difficult to add time intervals between semantics	Precision - 98% on yahoo movie dataset
[11] Senticircle, Lexicon based approach	OMD, HCR , STS-Gold	Provides better performance compared to other lexicon methods.	Integration of ML method with lexicon gives better performance. All baseline methods used in this approach are syntactical and not semantic.	Subjective detection Accuracy - 81%, F Measure - 80% Binary Polarity detection Accuracy -85%, F Measure- 84% Runtime analysis on STS-Gold dataset - 10ms
[12] ISR-WN Integrated resources including WN,WND,WNA, SUMO, SC, XWN SWN , WN version 1.6 and 2.0	Movie and TV reviews, from IMDB	Integration of Multi-dimensional knowledge network used to detect the alignment problems	Requires ontology based concepts. It does not suit for other domain classifications	House reviews achieve higher accuracy.
[13] Random forest, Finding Helpfulness scoring, extracting emotions	Product categories, online customer reviews	Crowdfunded (NRC) and expert creation (GALC) are used to perform the classification. crowdfunding outperforms and gives the better precision rate	Expert creation of small lexicon reduces the precision. Requires different dictionaries and ensemble methods to improve the precision	NRC data set achieves better performance 40.3% than GALC dataset
[14] Emotion based Cross Domain Recommendation	Movie, Book	Used to determine 6 basic emotions (love, joy, anger, surprise, sadness and fear)	Requires higher computational cost and emotions with other features to strengthen the CDR model	F1-Measure 71.1%

2.2. Corpus based approaches

Corpus based approach determines the resemblances of words based on the co-occurrence of words in the corpus. Corpus based treasures, and word net based approach to increase the text classification performance is presented [5]. The combination of CBT and WN gives better classification results. A hidden de-noising classification model which determines sentiment and emotion classification on different scales of noisy labels is proposed [15]. The proposed method is compared with tuning the variable of parameters and conventional techniques and is used for sentiment and emotion classification.

A new technique termed COS-HMM to handle the class imbalance problems in text classification is presented [16]. Over sampling with HMM technique is to increase the classification performance of SVM. Word generation and word weight generation approach are used to eliminate the class imbalance problems. The proposed methodology is compared with RTOS and SMOTE state-of-art techniques and is used for class imbalance problems in text classification.

A new cross lingual topic model which combines the state of art aspects and sentiment model to improve the sentiment classification in target languages are discussed [17]. This methodology is compared with universal sentiment lexicons collected from HowNet and SVM. The proposed method is used for Cross-Lingual Sentiment classification. [18] Presented a Term Weighted Learning with genetic programming for better text classification. This methodology is used in thematic and non-thematic text classification and image classification, which outperforms conventional schemes and other Term Weighted Learning approaches in Text classification such as Text categorization, Image classification and Authorship attribution. Social-emotional classifications of short tweets based on sentimental analysis are discussed in [19]. Table 2 presents the detailed study of various methods on corpus based approaches and its performance. A novel method termed polaritysim which is used for estimating the word-level contextual polarity that uses online customer rated reviews as a reference corpus with a positive polarity and negative polarity are presented [20]. The overall polarity numerical rating and score of positive and negative corpus are computed in word-level. Mexico-syntactic features are used for word-level polarity determination, which is better than lexical and syntactic features. The proposed approach gives 80% performance in out of domain and achieves 83-91% accuracy within the domain. Evaluating similarity between two vectors is obtained as follows.

$$probability(p) = \frac{freq(p)}{\sum_{i=1}^{p} freq(i)} \qquad (6)$$

Where p is the patterns, $freq(p)$ is the frequency of patterns occurring in the corpus C and the total number of patterns p mined from the C corpus. If the probability of pattern p is higher than the positive corpus then the negative corpus is assigned as positive. Similarly, negative corpus is applicable as vice versa. The feature expansion of word2Vec is calculated as

$$PolSim_{w2v} = PolSim_{TF.IDF} + \sum_{f=1}^{F} \sum_{r=1}^{R} TF_r.IDF_r.W2V_r \qquad (7)$$

Where r is the word connected to the word in feature f over Word2vec. R is the higher rank related words in the Word2Vec. The cosine similarity score $W2V_r$ is assigned to r in the Word2Vec.

The creation of health associated sentiment lexicon with hybrid approach are explained [21]. This methodology uses the bootstrapping, health opinions dataset and corpus based sentiment lexicon for preprocessing, creation and lexicon expansion, removing irrelevant words, polarity determination and polarity score modification. The polarity score is allocated to words for presenting the weighing approach. The proposed method achieves better performance in terms of accuracy, recall, precision and f1-score. The final maximum polarity of sentiment score is measured as follows,

$$pol^{swn}(w_i) = \begin{cases} pol^{+} & if\ \max(pol^{+}, pol^{-}, pol^{0}) = pol^{+} \\ pol^{-} & if\ \max(pol^{+}, pol^{-}, pol^{0}) = pol^{-} \\ pol^{0} & else \end{cases} \qquad (8)$$

Where polarity sentiment scores of the word $pol^{swn}(w_i)$ is positive, if the average positive score is higher than negative and objective scores. It is same as negative sentiment score. If the average sentiment score of positive and negative is same, then it is objective. The polarity class detection is measured in specific domain

$$pol\ class(w) =$$

$$\begin{cases} +ive,\ if \left(\frac{freq(w\in T_{+})}{|T_{+}|}\right) > \left(\frac{freq\ (w\in T_{-})}{|T_{-}|}\right) \\ -ive,\ \ else \end{cases} \qquad (9)$$

Where the polarity class detection is defined to verify the frequency of terms in a specified labelled class. $\left(\frac{freq(w\in T_{+})}{|T_{+}|}\right), \left(\frac{freq\ (w\in T_{-})}{|T_{-}|}\right)$ are the probabilities of word w occurring as positive and negative reviews in the training corpus. T_+ and T_- indicate the positive and negative opinions. . Here the words are not found in eq (7) and (8), polarity modification score is proposed. The polarity modification score is calculated by the given equation.

$$pol^{modified} =$$

$$\begin{cases} tfxidfx \left(\frac{freq(w\in T_{+}))}{|T_{+}|}\right), if \left(\frac{freq(w\in T_{+})}{|T_{-}|}\right) > \left(\frac{freq(w\in T_{-})}{|T_{-}|}\right) \\ tfxidfx \left(\frac{freq(w\in T_{-})}{|T_{+}|}\right),\ \ if \left(\frac{freq(w\in T_{+})}{|T_{+}|}\right) < \left(\frac{freq(w\in T_{-})}{|T_{-}|}\right) \end{cases} \qquad (10)$$

Where if the polarity score is not found in the SWN, the predicted class polarity modification score is used. The polarity modification score integrates the tf (Term Frequency), idf (Inverse Document Frequency) and count based probability.

Table 2: Various Methods on Corpus Based Approaches

Method	Domain	Advantages	Limitations	Results
[5] KNN, BPNN, MRBP, LPEBP	Reuters 21578 data and 20 Newsgroups corpus	Achieves high categorization performance equally measured by precision, recall, F-measure	Not suitable for high dimensional data	Combination of CBT+WN achieves better performance compared to dataset2 LPEBP which is 93%
[16] COS-HMM, Support Vector Machine	Medical document corpora OHSUMED, TREC	Supports class imbalance problems	Not suitable for multi labelled data. Lack of advanced machine learning techniques reduces the performance	F-Measure is above 0.6%-0.8%
[15] HDCM	STS and ISEAR	HDCM achieves good performance compared to SVM kernel function (linear and RBF) and Char SCNN	Applicable only on unseen noisy data and performs poor on small scale noisy data.	For large scale data, noisy labels give better classification performance. Accuracy - 79%, Precision - 80% Recall - 79% F-Measure - 79%
[17] CLLDA integrates CLJST and CLASUM	Chinese hotel reviews dataset	Useful for sentiment classification in different domains and different languages.	Suitable for only small datasets	Results are based on increasing number of parameters. Accuracy of CLJST - 56% Accuracy CLASUM - 76%
[19] TME	Real world datasets BBC, Digg, Myspace, Runners World, Twitter, YouTube	Mainly used for over-fitting problem	Poor performances on emotional application domains. TME achieves less classification performance on average Pearson's correlation	Accuracy is 86.06% Mean of average precision is 87% Average Pearson's correlation is 48%
[20] Polaritysim, SVM, MNB	Corpora, Restaurant, MP3, Photography	Proposed method outperforms when compared with	MNB achieves highest performance on MP3 dataset. It only classifies binary polarity	Restaurant dataset achieves highest accuracy of 91%

		SVM and MNB in out-domain corpus. Polaritysim is labor-intensive to build		
[21] Bootstrapping, corpus based polarity detection and scoring	Health reviews	Proposed method gives better performance compared with other methods	Health related lexicon is required to reduce the noise in expanded lexicon. Dynamic online updating of lexicons is required to investigate.	Precision 89% Recall 79% F-Measure 83%

3. Hybrid based approaches

A hybrid approach is a combination of machine learning based approach and lexicon based approach which gives better classification accuracy. A semantic detection based on sentiments is discussed [2]. This approach combines the dictionary based approach with the machine learning based approach. The SentiWordNet is used to determine the polarity of the words and generate the weights of the features using Chi-Square, GSS Coefficient and Odd Ratio. SVM is used for better sentiment categorization. The methodology is compared with seven benchmark datasets and it is observed that their proposed method outperforms the current state of art techniques.

A methodology to solve sparsity or short text classification in large scale web document is presented. This approach is compared with TF-IDF, Paragraph Vector, Long short term memory and other baseline algorithms. The proposed method is used for short text classification and word embedding [22].

The word sense disambiguation for sentiment classification is proposed [23]. SentiWordNet 3.0 lexicon is used for extracting words and deals with disambiguation of words. The proposed method is compared with movie and hotel domain oriented sentiment lexicons and improves the classification performance of sentiments in domain documents. This method provides three WSD techniques such as general sentiment lexicon, SentiWordNet and WSD sentiment lexicon using RBF, SVM, J48, NB and SVM-linear for determining the accuracy. The threshold is calculated as given below.

$$TF_i = \frac{w_i}{\sum_k w_k} \tag{11}$$

Where w_i denotes the frequency of words occurring in the document collection w_k. Similarity distance of two tokenization approaches are compared in the given equation.

$$TF_i = \frac{w_{i,\ i+1}}{\sum_k w_{k,\ k+1}} \tag{12}$$

Where $w_{i,\ i+1}$ is the frequency of bigram. After pre-processing of documents the first important word is word$_i$ and the second important word is word$_{i+1}$. The similarity of words and documents are measured by the cosine similarity in the following equation.

$$Score_{cosine}(P,R) = \frac{\sum_{k=1}^{n} P_k \times R_k}{\sqrt{\sum_{k=1}^{n} P_k^2} \times \sqrt{\sum_{k=1}^{n} R_k^2}} \tag{13}$$

Where P is the final sentiment vector and R is the document vector representation. The shortest mutual distance is calculated between two words based on the equation

$$WordNet\ path\ Simil(X.Y) = \frac{1}{\min(mutual\ dist_{x.y})+1} \tag{14}$$

Where, Word Net path similarity is defined as the semantic structure from the hierarchical tree. X and Y are the two words used to calculate the minimal similarity distance. The average score of words used in the document is calculated using the equation

$$Average\ Similarity_i = \frac{\sum_{i=1}^{n} Similarity\ _i}{n} \tag{15}$$

Where similarity value i of words occurring in the reviews n is calculated. A hybrid approach on sentiment analysis at sentiment level is used to estimate polarity determination. This approach improves the performance compared with Naïve Bayes and Maximum Entropy state of art techniques. The proposed approach focuses on sentiment analysis at sentence level and in polarity determination [24].

An aspect based extraction in opinion mining and Deep Convolutional Neural Network is used to extract the opinionated words from the sentence either as aspect or not aspect word [25]. This approach is compared with existing state of art techniques which achieves highest accuracy and classification.

The stress and relaxation detection in the context of transportation is explored. This approach is used with human annotators, supervised and unsupervised versions and machine learning algorithms. The proposed method is used for sentiment classification in stress and relaxation domain [26].

A methodology proposed for corpus and lexicon based approaches mainly is used to create the text documents. This approach is used for classifying sentiments and polarity [27]. A genetic algorithm is proposed for optimization problem and is used for finding lexicons in the opinionated text. Meta level, Bing liu and n-gram features are used to extract opinionated text. The proposed approach is compared with 13 other approaches and is observed to improve the performance in terms of accuracy and f1-measure. It is 5.53% better in SOMD dataset, 2.19% on HCR dataset 1.37 on OMD dataset and 0.34 on STS Dataset.

An enhanced sentiment analysis and polarity determination for opinion mining are presented [28]. The framework is evaluated and compared with seven well known benchmark datasets which gives better results. The proposed method is used for sentimental analysis and polarity classification. Table 3. Presents the overview of various techniques on hybrid based approaches and the performance metrics is discussed in detail.

Table 3: Various Method on Hybrid Based Approaches

Method	Domain	Advantages	Limitations	Results
[2] SentiWordNet, Chi-Square, GSS Coefficient, Odds Ratio, SVM	LMR, CMR, Books, DVD, Apparel, Health, Video	Outperforms the current state of art techniques	Pre-processing technique is required	Average feature presence Accuracy is 81%, F-measure is 83% Average feature frequency Accuracy is 78% F-Measure is 80% and CMR dataset Accuracy is 86%
[22] NLP includes short text classification+ word embedding clustering, CNN	Google Snippets and TREC	Achieves better classification performance	Requires supervised down sampling method, task specific embedding learning, embedding affinity measurement to solve the data sparsity	Word2Vec word embedding achieves highest performances. 85.5% on Google Snippets dataset. Glove word embedding received highest results 96.8% on TREC
[23] Pre-processing of WOM documents, tokenization, WSD, WSD-based sentiment lexicon SVM-Linear, SVM-RBF, j48 decision tree and NB	IMDB and Hotel reviews	Achieves improved performance in domain documents	Not suitable for multi-linguistic domains. Needs powerful supervised learning techniques to the WSD	DSWN- Method 1 achieves accuracy 75.35% and t-test 72.39% using SVM classifier on IMDb DSWN-Method 2 achieves accuracy 78.29% and t-test 77.87% using SVM classifier on Hotel review dataset
[24] Sentiment/opinion lexicon, Semantic rules, Fuzzy Sets, HSC, HAC	Movie review, twitter dataset A, twitter dataset B	Significantly improve the performance.	Ambiguous in dealing with idioms, jargon, argot and leads to misclassification. Focuses only on opinions and not imagery cases. Precision rate is low.	HSC classifier achieves highest performance on Twitter A dataset compared to the other datasets. Accuracy 88.02% Precision 84.24% HAC classifier achieves polarity classification above 80% in all cases
[25] 7-layer deep convolution neural network, Linguistic patterns	Different Product based reviews, SemEval dataset	Does not require any feature engineering. It takes less time and cost. SemEval dataset achieves highest performance on Amazon embedding	Precision and recall are less in unlabeled data	F-Score on laptop data is 80.68%, Restaurant -85.70%, Linguistic patterns on SemEval 2014 dataset of Laptop achieves Recall 78.35%, Precision 86.72%, F-Score 82.32% Restaurant achieves Recall 86.10%, Precision 88.27% , F-Score 87.17%
[26] Tensistrength uses lexicon rule based approach, generic machine learning classifiers	Social Media tweet messages	Tensi strength gives equal performances compared to human coders. MAD gives outstanding performance compared to all other techniques	ML algorithm provides better performance compared to Tensi Strength. Requires to improve classification performance related to stress in other domain.	Mean absolute deviation of stress and relation is 53% and 31% compared to machine learning algorithms
[27] Genetic algorithm, sentiment lexicon, meta-level features, Bing Liu's lexicon and n-gram features	Sanders, OMD, Strict OMD, HCR (Health Care Reform), SemEval, Stanford	Integration of lexicon and corpus based approaches give better results	Creation of lexicon is time consuming. Sentiment score of different domain is different. It is difficult to compute using proposed method	Six datasets achieves overall accuracy of 80%. Four datasets achieves F-measure of 80%
[28] SWN-V	Cornell movie review dataset, Cornell, Apparel, Books, DVD, Health, Video, LMR	Efficiently deals with data unavailability, data sparsity, domain dependence and contextual information problems	Requires information gain and gain theory to increase the performance.	LMR dataset achieves highest performance. Accuracy 85.76% Recall 87.47% F-Measure 86% Video dataset achieves highest precision rate of 84.62%

4. Comparative analysis and discussions

Tables 1, 2 and 3 describe the approaches, data set, advantages, limitations and observations of various approaches. Domain type indicates the data collected from various domains. Observations describe the classification and categorization performance such as Accuracy, Precision, Recall, F-Measure/Score, Average Mean, Auto-Pearson Correlation and Ratio of time.

Lexicon based approach consists of dictionary and corpus based approach. Comparative analysis of a dictionary based approach is presented in table 1. The higher classification performance is presented and highlighted. The methodology is proposed by [11] using Senticircle and lexicon based approach achieves highest accuracy in terms of performance with subjectivity classification of 81%, polarity detection of 85% F-measure of 84% and the ratio of time as10ms compared on OMR, HCR and STS-Gold data sets. A methodology termed temporal semantic relations, which integrate the connection entity, lexical syntactic patterns, context sentences, context graph and context communities achieving highest precision of 98% on Yahoo's movie star data set are presented [10].

An ensemble approach which integrates the text pre-processing, text normalization, semantic indexing techniques achieves highest f-score 96% on UMICH data set [9]. The comparison of corpus based approach is presented in table 2. The methodology proposed by [20] polaritysim, SVM, MNB is performed on restaurant data set which achieves highest accuracy of 91%. A methodology such as bootstrapping, corpus based polarity detection and scoring, which are performed on health care review data set achieve precision of

89%, recall of 79% and F-Measure of 83% [21]. A methodology termed TME, which achieves highest classification accuracy as 86.06%, Mean of average precision as 87% and Average Pearson's correlation as 48% are discussed in [19].

Hybrid based approach is a combination of a lexicon based and machine learning based approach. The comparison of hybrid based approach is presented in table 3. A methodology called Sentiment/opinion lexicon, Semantic rules, Fuzzy Sets, HSC, HAC, which achieves better classification accuracy - 88.02% and Precision - 84.24% compared to that of twitter data set is discussed [24]. The methodology proposed by [28] SWN-Vocabulary, which is performed on LMR data set achieves highest performance accuracy - 85.76%, recall - 87.47%, F-Measure - 86%. Video data set achieves highest precision rate of 84.62%. A methodology called 7-layer deep convolution neural network and Linguistic patterns achieves highest classification performance [25]. The SemEval (2014) data set with restaurant data achieves recall - 86.10%, precision - 88.27% and F-Score- 87.17%. A methodology termed as word embedding clustering and CNN achieves highest performance on classification of short tweets (96%) on TREC data set. By observing the comparison of three approaches the combination of machine learning and lexicon based approaches achieve highest classification performances [22].

5. Open issues and research gaps

The following challenges are presented with a perspective to focus subsequently by researchers.

Lexicon based approach: Senticircle and lexicon based approach is proposed by [11]. In order to reduce misclassification, optimization rules to reduce irrelevant, unwanted tweets and computational time to re-examine can be incorporated. It requires to increase the classification of precision with ensemble methods and different dictionaries [13]. An approach based on bootstrapping, corpus based polarity detection and scoring for creating a health-related sentiment lexicon are presented [21]. However, it needs specific health-related domains to reduce the noise in the expanded lexicon. Online dynamic updating is required in the web repositories. An information retrieval approach for word-level polarity in within and out of domain are explained [20]. It requires determining of fine grained polarity categories too. A methodology is proposed for sentiment propagation only. In order to improve the classification performance, it requires linguistic patterns on different domain [6]. A methodology based on word level and sentence level sentiment analysis on building lexicon level sentiment classification is discussed. As re-embedding takes time, ternary classification is required [8].

Hybrid based approach: The proposed methodology requires lexicon based method to increase the speed [15]. It needs to improve the performance on using POS tagging, automatic real time tools, and prototypes [24]. A SWN-V for sentiment classification is proposed in [2]. In order to increase the classification performance it requires information gain and gain theory. The word embedding clustering and CNN for short tweet sentiment classification are presented [22]. It requires supervised down sampling method, task specific embedding, learning and embedding affinity measurement to solve the data sparsity. A methodology called sentiment/opinion lexicon, Semantic rules, Fuzzy Sets, HSC, HAC for sentiment classification [24]. It is required to concentrate on ambiguous data that deals with idioms, jargon, argot and reduces misclassification. An approach for sentiment classification in micro blogs is proposed. Genetic algorithm, sentiment lexicon, meta-level features, Bing Liu's lexicon and n-gram features are used in the framework [27]. It requires concentrating on creation of the lexicon to reduce the time-consuming and sentiment score in other domains.

6. Conclusion

With web 2.0, online users are increasing day to day facilitating users with more sophisticated data to express their opinions on online social media. Upon exhaustive analysis of the literature carried out, many motivating features meeting the current state of the art approach in text mining such as sentiment analysis, categorization of text documents, entity and tweet recognition, short text; polarity determination in document, sentence and aspect level analysis is presented. A broad perspective on various approaches and techniques presents in the current literature providing an overview of the advantages and limitations in the existing methods is discussed. It is observed that hybrid based approach has more scope with reasonable results.

References

[1] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Inf. Fusion*, 2017. https://doi.org/10.1016/j.inffus.2016.10.004.

[2] F. Khan, U. Qamar, and S. Bashir, "Lexicon based semantic detection of sentiments using expected likelihood estimate smoothed odds ratio," *Artif. Intell. Rev.*, 2017. https://doi.org/10.1007/s10462-016-9496-4.

[3] B. Liu, "Sentiment analysis and opinion mining," *Synth. Lect. Hum. Lang. Technol.*, 2012.

[4] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications," *Knowledge-Based Syst.*, 2015.

[5] C. Li, J. Yang, and S. Park, "Text categorization algorithms using semantic approaches, corpus-based thesaurus and WordNet," *Expert Syst. Appl.*, 2012. https://doi.org/10.1016/j.eswa.2011.07.070.

[6] M. Fernández-Gavilanes and T. Álvarez-López, "Unsupervised method for sentiment analysis in online texts," *Expert Syst. with*, 2016. https://doi.org/10.1016/j.eswa.2016.03.031.

[7] R. Shah, Y. Yu, A. Verma, S. Tang, and A. Shaikh, "Leveraging multimodal information for event summarization and concept-level sentiment analysis," *Knowledge-Based*, 2016.

[8] D. Tang, F. Wei, B. Qin, N. Yang, T. Liu, and M. Zhou, "Sentiment Embeddings with Applications to Sentiment Analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 2, pp. 496–509, Feb. 2016. https://doi.org/10.1109/TKDE.2015.2489653.

[9] J. Lochter, R. Zanetti, D. Reller, and T. Almeida, "Short text opinion detection using ensemble of classifiers and semantic indexing," *Expert Syst. with*, 2016. https://doi.org/10.1016/j.eswa.2016.06.025.

[10] Z. Xu, X. Luo, S. Zhang, X. Wei, L. Mei, and C. Hu, "Mining temporal explicit and implicit semantic relations between entities using web search engines," *Futur. Gener. Comput.*, 2014.

[11] H. Saif, Y. He, M. Fernandez, and H. Alani, "Contextual semantics for sentiment analysis of Twitter," *Process. Manag.*, 2016.

[12] Y. Gutiérrez, S. Vázquez, and A. Montoyo, "A semantic framework for textual data enrichment," *Expert Syst. Appl.*, 2016. https://doi.org/10.1016/j.eswa.2016.03.048.

[13] A. Felbermayr and A. Nanopoulos, "The Role of Emotions for the Perceived Usefulness in Online Customer Reviews," *J. Interact. Mark.*, vol. 36, pp. 60–76, Nov. 2016. https://doi.org/10.1016/j.intmar.2016.05.004.

[14] S. Chakraverty and M. Saraswat, "Review based emotion profiles for cross domain recommendation," *Multimed. Tools Appl.*, 2017.

[15] Y. Wang, Y. Rao, X. Zhan, H. Chen, and M. Luo, "Sentiment and emotion classification over noisy labels," *Knowledge-Based Syst.*, 2016.

[16] E. Iglesias, A. Vieira, and L. Borrajo, "An HMM-based over-sampling technique to improve text classification," *Expert Syst. Appl.*, 2013. https://doi.org/10.1016/j.eswa.2013.07.036.

[17] Z. Lin, X. Jin, X. Xu, Y. Wang, and X. Cheng, "An unsupervised cross-lingual topic model framework for sentiment classification," *Audio, Speech, ...*, 2016.

[18] H. J. Escalante *et al.*, "Term-weighting learning via genetic programming for text classification," *Knowledge-Based Syst.*, vol. 83, pp. 176–189, Jul. 2015. https://doi.org/10.1016/j.knosys.2015.03.025.

[19] Y. Rao, H. Xie, J. Li, F. Jin, F. Wang, and Q. Li, "Social emotion classification of short text via topic-level maximum entropy model," *Inf. Manag.*, 2016. https://doi.org/10.1016/j.im.2016.04.005.

[20] O. Vechtomova, "Disambiguating context-dependent polarity of words: An information retrieval approach," *Inf. Process. Manag.*, 2017. https://doi.org/10.1016/j.ipm.2017.03.007.

[21] M. Asghar, S. Ahmad, M. Qasim, S. Zahra, and F. Kundi, "SentiHealth: creating health-related sentiment lexicon using hybrid approach," *Springerplus*, 2016. https://doi.org/10.1186/s40064-016-2809-x.

[22] P. Wang, B. Xu, J. Xu, G. Tian, C. Liu, and H. Hao, “Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification,” *Neurocomputing*, 2016. https://doi.org/10.1016/j.neucom.2015.09.096.

[23] C. Hung and S. Chen, “Word sense disambiguation based sentiment lexicons for sentiment classification,” *Knowledge-Based Syst.*, 2016.

[24] O. Appel, F. Chiclana, J. Carter, and H. Fujita, “A hybrid approach to the sentiment analysis problem at the sentence level,” *Knowledge-Based Syst.*, vol. 108, pp. 110–124, Sep. 2016. https://doi.org/10.1016/j.knosys.2016.05.040.

[25] S. Poria, E. Cambria, and A. Gelbukh, “Aspect extraction for opinion mining with a deep convolutional neural network,” *Knowledge-Based Syst.*, 2016.

[26] M. Thelwall, “TensiStrength: Stress and relaxation magnitude detection for social media texts,” *Inf. Process. Manag.*, 2017. https://doi.org/10.1016/j.ipm.2016.06.009.

[27] H. Keshavarz and M. S. Abadeh, “ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs,” *Knowledge-Based Syst.*, vol. 122, pp. 1–16, Apr. 2017. https://doi.org/10.1016/j.knosys.2017.01.028.

[28] F. Khan, U. Qamar, and S. Bashir, “eSAP: A decision support framework for enhanced sentiment analysis and polarity classification,” *Inf. Sci. (Ny).*, 2016.

Appendix

Table 4: Acronyms and Abbreviations

WN-Word Net	BPNN-Back Propagation Neural Network	WOM-Word Of Mouth	HCR-Health Care Reform
WND-Word Net Domains	MRBP-Mobility Neuron Rectified BPNN	WSD-Word Sense Disambiguation	STS-Gold- Stanford Sentiment Gold
WNA-Word Net Affect	LPEBP-Learning Phase Evaluation BPNN	RBF-Radial Basis Function	YFCC100M-Yahoo Flickr Creative Common 100M
SUMO-Suggested Upper Merged Ontology	TME-Topic Level Maximum Entropy	SVM-Support Vector Machine	STS-Stanford Twitter Sentiment
SC-Semantic Class	COSHMM-Content based Over Sampling Hidden Markov Model	NB-Naïve Bayes	ISEAR-International Survey on Emotion Antecedent & Reaction
XWN-Extended Word NET	HDCM-Hidden De-noising Classification Model	NLP-Natural Language Processing	CBT-Corpus based Thesaurus
CLLDA-Cross Lingual Latent Dirichlet Allocation	MAD-Mean Absolute Deviation	CNN-Convolution Neural Networks	DSWN-Domain Oriented Senti Word Net
CLJST-Cross Lingual Joint Sentiment Topic	POS-Parts of Speech	SWN-V – Senti Word Net Vocabulary	CDR-Cross Domain Recommendations
CLASUM-Cross Lingual Aspect & Sentiment Unification Model	HSC-Hybrid Standard Classification	IMDB-Internet Movie Dataset	TF-IDF-Term Frequency Inverse Document Frequency
KNN-K-Nearest Neighborhood	HAC-Hybrid Advanced Classification	OMD-Obama McCain Debate	LSTM-Long Short Term Memory