Multimedia Tools and Applications (2022) 81:11989–12004
https://doi.org/10.1007/s11042-020-10480-w

1177: ADVANCES IN DEEP LEARNING FOR MULTIMODAL FUSION AND ALIGNMENT

Exploration of sentiment analysis and legitimate artistry for opinion mining

R. Satheesh Kumar[1] · A. Francis Saviour Devaraj[2] · M. Rajeswari[1] · E. Golden Julie[3] · Y. Harold Robinson[4] · Vimal Shanmuganathan[5]

Received: 12 May 2020 / Revised: 28 October 2020 / Accepted: 29 December 2020 /
Published online: 23 January 2021

Abstract
Sentiment analysis/opinion mining is a technique that analyzes people's opinions, evaluations, sentiments, attitudes, appraisals and emotions to entities like products, organizations, services, issues, individuals, topics, events and their attributes. It is a massive problem space. People tend to express their opinions on anything, such as, a product, service, topic, individual, organization, or an event. Here, the term object represents the entity commented on. Certain private states parts that cannot be judged and observed include the following, beliefs, opinion, emotions and sentiments. The above mentioned aspects are usually expressed in documents using certain subjective words that determine the private states with the help of unique dictionaries like the WordNet or SentiWordNet. The feature selection concept is incorporated in the following tasks such as image classification, data mining, cluster analysis, image retrieval, and pattern recognition. This is observed as a data analysis pre-processing strategy; here a subset from the original data features is thus selected for eliminating the noisy/irrelevant/redundant features. This technique essentially helps in minimizing the incurred computational expenses and helps in enhancing the accuracy level of the data analysis procedures. The Semantic features are meant to concentrate on the relationship between the signifiers such as that of the words, phrases, signs and symbols. A special of semantics called as the linguistic semantics is used for understanding human based expression in opinions and blog. A semantic based feature selection strategy has been introduced for establishing the opinion mining tasks. This introduced semantic based feature selection makes use of the SentiWordNet that is observed to be a lexical resource of the WordNet database extracted terms and is therefore used in the research tasks. Feature set is minimized with the help of the introduced semantic based approaches for the purpose of considering the individual predictive ability words and selection features.

✉ Vimal Shanmuganathan
svimalphd@gmail.com

Extended author information available on the last page of the article

Experiments were conducted with the help of the Naïve Bayes, the FLR and the AdaBoost classifiers and the obtained results were compared for understanding and judging the feature selection methods.

Keywords Opinion mining · Sentiment analysis · IMDb dataset · Stemming · Stop words · Term frequency – Inverse document frequency · AdaBoost · Fuzzy lattice reasoning · Naïve Bayes

1 Introduction

Personal Evaluation of OM seems to be a difficult task. This is because the process involves a higher number of opinions that directly reflects every single person's perception towards the issue [1]. For convenience, the OM methodshave been categorized as the Attribute-enabled and the Sentiment-enabled strategies. The primary goal is to extract suitable as well as better opinions by means of restricting the keywords or the attributes and by eliminating the wrong/misleading ideas [13]. High precision for recalling is the most significant disadvantage of this method. It arises due to the Out-of-Vocabulary attributes.

The various methods that are similar to theabove-discussed techniques used for the OM and the ML depends onthe Lexicon (a text formed by the collection of words without considering their particular relations among such words [7]). Opinion Lexicons are resources combined with sentiments and words. The major disadvantage of this method is that few wordsare sensed to be both positive and negative based on its usage. The MLclassification algorithm has been trained with an efficient corpus and annotations for recognizing the essence of the word as positive or negative, depending on the situation.Any ML-based classification mechanism requires both the trainingand the testing datasets [18]. The dataset preparation method used by the automatic classifier for distinguishing the properties of the documents and the test datasetshave been used for measuring the accuracy of the automated classifier. A lot of ML techniqueshave been used for differentiating the reviews. Some of the prominent mechanisms used for text classifications includetheSupport Vector Machines (SVM), and the tools used for Natural Language (NL) K-Nearest Neighbors, ID3, N-Gram Model, C5, Winnow Classifier, and Centroid Classifier.For most of the naturalimplementations,theNaïve Bayes classifier has been treated as the best choice as it exhibitsa better evaluation of the parameters with less training data. When it comes to the fastest application,theAdaBoost classifier should be considered. The significant advantage of this classifier is that, it can be used as a combination with the other learning algorithms. It does not require any prior information regarding the weak learners and the works on any data like numbers, text, and it can be further extended to the learning problems beyond the binary classifications.

SentiWordNet is a lexical tool based on the English WordNet which integrates the sentiment information for each of thesynset. In WordNet three numerical scores are calculated for each of the synset S: Pos(S), Neg(S) and Obj(s), describing the positivity, negativity, and objectivity of the synset. Scores from the SentiWordNet basically an English WordNet with synset-linked polarity scores offers weights to the features, i.e. words of a document leading to better classification of the sentiments in the documents. SentiWordNetsemi-automatically offers the term level informations on the opinion polarities by deriving this information from the WordNet database of English words and relations. Positive and negative scores ranging from 0 to 1 are seen in the SentiWordNet

for each of the WordNet term, which reveals its polarity. Higher scores doesindicate terms with information about the heavy opinion bias, while lower scores indicate a less subjective term.

The FLR classifierhas been used for incremental learning and has been found to be capableof tuning the generalization beyond a hyper box, granular computing, dealing with the loss of information, and implementation beyond $R^N$for a common lattice data domain. Thus, these three Naïve Bayes, FLR, and an AdaBoost classification technique used for the OM has been investigated in this paper.

2 Literature review

2.1 Opinion mining

The system proposed by Jeyapriya&Selvi (2015) [14] is based on a phrase-level examination of the client reviews. Phrase-level OM is also known as an aspect-based OM that is used for extracting the most relevant aspects of an item and for predicting an aspect orientation from the item reviews.In customer product reviews and mining opinions on whether the review is positive / negative, the projected system has implemented an aspect extraction technique using the frequent item-set mining. In customer reviews it has been defined as the sentiment orientation of each of the dimensionswith the help of supervised learning algorithms.

Mukhtar et al., (2018) [22] proposed and defined a new type of tree called the tree of opinion. As a result, the flexible OM model was created based on the opinion tree, here the coarse-grained, medium-sized and fine-grained (three different granularities) OM has been realized as a unified flexible model. The flexible OM procedure is set for public opinions on the internet. Finally, an experiment on how to build the tree of opinions was completed and the overall opinions were constructed in the form of a tree as a hot topic on the internet.

Chinsha& Joseph (2015) [8] focused on the OM aspect levels, they have proposed a new approach with reference to the syntactic dependency, aggregate score of opinion words, SentiWordNet and OM aspect tables. The restaurant reviews were worked out in their work. The data collection for the restaurant reviews were obtained from the web and were labeled manually. On the annotated test range, the new approach achieved an accuracy of 78.04%. The method was contrasted with one that used Part-Of - Speech tagger for the extraction of features; results showed that the new method on the annotated test set provided 6% more precision than that of the existing ones.

2.2 Web based queries

Pappas et al., (2012) [23] proposed an agent-focused crawling system for the retrieval of the subject and the genre based web documents. A set of focused crawler agents tend to explore the specific web paths in the parallel topicswith the help of the dynamic seed Uniform Resource Identifiers (URIs), these belong to certain web genres and are collected from the web search engines, starting from a simple topic query. The agents here have used an internal framework for measuring the unvisited web pages with the subject and the genre-relevance ratings. The authors conducted an experimental study for testing the actions of the agents for various subject questions, they have further demonstrated the advantages and capabilities of the system.

The functions of a framework designed for the conduct analysis of the e-commerce customers defined by Dziczkowski et al., (2013) [10] allowed for user identifications and customer behavior extractionsfor communicating with the customers of the website. General Web Usage Mining approaches were presented with proposals for extending the database with information from e-commerce site for queries. The system analyzed and measured the opinions on the approach that was based on the natural language linguistics and the statistical treatments. Three different methods were used for identifying the opinions from the forum of clients, and the application of the linguistic information based on the two new methods depends on the emotions / opinions of the clients mentioned in the comments of the forum.

Abu-Salih et al. (2018) [2] had proposed a new OM-based approach. Customers evaluated the indexes of various Websites quantitatively with the help of reviews. The authors used the Mutual Reinforcement Approch (MRA) for improving the accuracy of the mining results. Unlike most of the OM approaches, this method focusses on the explicit factor-opinion pairs, MRA effectively mined the implicit pairs with respect to the pre-constructed associatives. The study and assessmentsthus performed have portrayed that the results of the proposed approach were consistent and reliable in comparison with the traditional approaches.

2.3 Survey on opinion mining of blogs

Kao & Lin (2010) [15] planned to develop a sentiment analysis system suitable for Chinese reviews that extracted user-interested features and detected semantine-oriented opinions depending on certain features / opinions in a particular category. The authors presented integral results to users. Experiments showed that dependency between features / opinions was effectively calculated by the system. Examination of analysis sentiments confirmed the applicability of the proposed process.

Alaoui et al., (2018) [3] discussed the topic of opinion question answering to answer opinion questions about goods using the opinions of reviewers. The authors approach, called the Aspect-based Opinion Question Addressing (AQA), supports addressing opinion-based questions by enhancing the shortcomings of the existing techniques.In the pre-processing phase, AQA adopted an OM technique for defining and estimating the efficiency of the target aspects. Goal aspects are characteristics or components which a review focusses on. The authors conducted experiments on a real life dataset, Epinions.comand, which demonstrated the AQA's improved efficacy with respect to its accuracy of the retrieved answers.

Li et al., (2018) [17] demonstrated an OM system which extracted the opinions and views from consumers / customers and analyzed them to provide concrete market flow with validated statistical data. To provide these features, the program used an OM-based grouping, clustering, and lingual awareness.

2.4 Survey on feature selection methods in opinion mining

All feature reduction methods improved the classifier's performance, as demonstrated by Alsaffar& Omar (2014) [4]. Support Vector Machine (SVM) approach ensured the highest accuracy in feature selection in comparison with the other classification approaches, such as the Principal Component Analysis (PCA) and the CHI square, in classifying the Malay feelings. In feature selection, SVM recorded an experimental accuracy of 87%.

OM on Thai restaurant reviews using K-Means clustering and Markov Random Field (MRF) feature selection proposed by Claypo&Jaiyen (2015) [9] started with pre-processing of the text for breaking the reviews into words and for removing the stop words, followed by the text transformation for creating keywords and for generating vectors for inputs. Selection of MRF features was done to select relevant features from many of the extracted features. K-Means was then employed to cluster it into positive / negative reviews. Results showed that selection of MRF features effectively reduced the features in a computational time-decreasing data. Baccianella et al., (2014) [6] introduced six novel feature selection methods explicitly designed for ordinal classifications. It was tested on two product review data sets against 3 literature approaches using two SV regression-tradition learning algorithms. Results showed that all the six metrics outperformed the three baseline techniques on both the data sets and the learning algorithms (were more stable than others by order of magnitude).

2.5 Survey on semantic based feature selection in opinion mining

Application and comparison of three classification techniques over a text corpus composed of reviews of commercial products to detect opinions about them was focused on by Mazzonello et al., (2013) [20]. This domain is about "perfumes", and user opinions in the corpus were written in Italian. The new approach was data-driven: a selection procedure for terms of Term Frequency / Inverse Document Frequency (TFIDF) was used to produceefficient computations, improved classification outcomes and for managing issues related to specific classification procedures were adopted.

Weichselbraunet al.(2014) [27] proposed a new approach for contextualizing and enriching the broad semantine knowledge bases for the OM based on Internet intelligence systems and high-throughput big data applications. The approach refers to traditional lexicons of emotion and multidimensional affective tools such as SenticNet. Quantitative assessment showed major changes by using an enriched version of SenticNet for polarity classification. Gold standard data from crowds with a qualitative evaluation shed light on the strengths / weaknesses of the concepts and the enrichment process.

An innovative OM methodology using a new Semantic Web-guided solutions to enhance the results with traditional natural language processing techniques and sentiment analysis processes was proposed by Peñalver-Martinez et al., (2014) [24] aimed at (1) enhancing the feature-based OM using ontologies at the feature selection stage, and (2) providing a new vector analysis method for sentimental analysis. Compared to the other traditional methods, the technique implemented / tested in a real-world movie review-themed scenario yielded very positive results. Table 1 discussed the various Opinion mining tools used at different levels.

Various tools implemented for extracting the opinions from user-created information have been explained below [5].

- A Review Seer Tool was used for site combinationwith various opinions automatically. Naïve Bayes Classifier picks both the positive and the negativeviews and avails a score for the extracted features.
- A Web Fountain uses beginningspecific Base Noun Phase heuristic strategy for obtaining the required features that assists in the construction of a primary interface for the web.
- A Red Opal tool supports the users in specifying the product's opinion orientations according to their features. Results were displayed using a web interface.

Table 1 Opinion mining at different levels

Classification of Opinion mining at different levels	Assumptions made at different levels	Tasks associated with different levels
1. Opinion Mining at Sentence level.	1. A sentence has one opinion posted by one opinion holder; in many cases, that could not be true. 2. In a document the second sentence boundary is established.	**Task 1:** identifying a sentence as subjective/opinionated **Classes**: objective and subjective (opinionated) **Task 2:** opinion classification of a sentence. **Classes:** positive, negative and neutral.
2. Opinion Mining at Document level.	1. A paper focuses on one topic and has one opinion holder posting opinions. 2. Not relevant for blog / forum post because in these sources there are multiple opinions about different items.	**Task 1:** opinion classification of reviews **Classes:** positive, negative, and neutral
3. Opinion Mining at Feature level.	1. Data source focuses on features of an object posted by the holder of a single opinion. 2. Not relevant to blog / forum post as multiple opinions on different items may exist.	**Task 1**: Identify/extract object features that were commented on by an opinion holder. **Task 2**: Determine whether opinions on features are positive/negative/neutral. **Task 3:** Group feature synonyms.

- An Opinion Observer is a mining system used for distinguishing the opinions gathered from the Internet for the user — aWordNet Exploring method usedfor assigning prior polarity.

3 Methodology

3.1 Proposed work

Feature selection's purpose is reduce the features by deleting those which are irrelevant while maintaining/enhancing the classification accuracy. Many search algorithms used for feature selections are available in literature. A semantic based feature selection for OM is proposed in this research. The objective of this research work is to evaluate the efficiency of the various types of classifiers in classifying the movie/web based medical query posts.

3.2 Methodology

In this work, the accessible IMDb dataset is used for categorizinga review as either positive or negative. Noise is removed using stop words and stemming procedures. Featuresare eliminated with the help of the Inverse Document Frequency mechanism along with Naïve Bayes, AdaBoost, and Fuzzy Logic Reasoning classifiers, respectively.

3.2.1 IMDb dataset and medical service dataset

The movie reviews data set comprises of 2000 movie reviews [25]: 1000 eachof positive and negative reviews evaluated by the classification algorithms. The previous version of the data

set had used 700 each ofpositive as well as negative reviews [11]. Reports gathered from an Internet Movie Database (IMDb) archives were segregated as either positive or negative classesautomatically derived from the ratings. The resultant dataset solely containsof reviews where the stars are used for representing the ranks of a movie. The entire searching process utilizes 200 positive or 200 negative opinions subsets.

IMDb.com, the internet movie database, collects movie data obtained from fans and studios. It claims itself asthe most significant web movie database maintained by Amazon. Additional data on IMDb.com is found online, which also includes the process of data collection. IMDb makes raw data available, but classified asa text file with each file's format possessing minute differences. A data file created by collecting the entire volume of data is to write a ruby script for extracting and preserving the required information in a database. This informationexported to CSV would then be importedin the form of several programs. Though IMDb data is available for download and personal use, it is appropriately protected by copyright laws [19]. Though we have transformed the IMDb data to the RDF format, we have failed in acquiring the required permissionsin publishing the same, and hence, our implementation has no information from IMDb even though the external links to IMDb pageshave been included based on the requirements.

For the medical query dataset, data collected from sites like Mayo Clinic Query and Answer (Q&A), medical weblogs, and reviews for a medical service dataset have been considered. It has medical weblogs posts on topics of medicine/health care based on the author's medical weblogs that are differentiated into blogs, preferably entered by the health care professionals and the patients respectively. Data has been collected from Mayo Clinic, a nonprofit medical organization. One thousand and six hundred data contentsconsisting ofan equal number of valid and same number of informative queries have been received from the website and the various links, including sections of Question and Answers, reviews of drugs, and from thestudies on diseases. An example of an educational class data used in this work has been collected from the Mayo clinic website and thus presented for perusal.

Endometrial Cancer develops in women's uterus. It is a hollow and a pear-shaped pelvic organ, where the fetal develops. This cancer slowly expands from the starting of a cell's internal layer that leads to the formation of a lining (endometrium) in the uterus. It is thus referredtoas the "Uterine Cancer. Various forms of cancer can also originate in the uterus but observed in a very smaller number of cases. Endometrial Cancer can be generally recognized at an initial phase as it causes abnormal vaginal bleeding. If endometrial cancer is identified at an early stage, surgery can be done for curing the same.

3.2.2 Pre-processing and feature extraction

Stemming Word Stemming is treated as a pseudo-linguistic procedure that eliminates suffixes to minimize the words to a word stem. For example, 'classifier', 'classified', and 'classifying' are preferably reduced as a word stem 'classify.'The dimensionality of feature space is controlled by mapping the morphologically identical words to the word stems [21]. A stemming algorithm is thus observed as the Porter developed suffix stripper. Arabic language comprises oftwo different morphological analysis techniques, namely, the stemming and the light-stemming techniques. While stemming minimizes a word to its stem, light-stemming deletes common affixes from a word.

Stop words Stop words are considered as a set of terms/words without useful information. Stop words are problematic in critical concepts and word identification from textual sources when not removed by their presence as regards frequency/occurrence in textual sources [26]. Stop wordsare observed as words that are strainedfor processing NLdata (text) in computing. There is no exact specific stop word list used by the tools. Some avoid their removal to support phrase searches. A word group was chosen to stop words for a purpose. These are frequently used words - for search machines – "like is, at, which, and on.

Term frequency IDF is anarithmatical statistic reflection that depends upon a word's significance in the respective document within a collection/corpus. It is considered as the combination of image processing and mining basedon the weighting factors. IDF value improves correspondingly to the times a phraseperforms in a text and is offset by word frequency in a quantity controlling the other words.

Simple word frequency can be replaced by the weighted frequency before the calculation of the cosine and the various statistics. The weighted frequency statistic is the Term Frequency-Inverse Documents Frequency that calculates a weight for a term that suitably reflects its significance. TF-IDF is a standard text categorization task metric, where its use in sentiment analysis is limited and not used as a unigram feature weight [28]. TF-IDF comprises of expression frequencies and contrary document frequencies. The former located by counting the times a term occurs in a document, and IDF achieved the same by dividing the total texts with reference to thereports offered bythe word. When such values multiplied, it resulted in highest score wordsmaterializing in a few documents frequently and lowest score words for terms appearing frequently in records, thus permitting the location of terms that are important in a document.

TF*IDF comes from IDF with a heuristic perception that a query term is not a better discriminator and so be issued with minimum weights than the one occurring in limited documents [26]. The Eq. (1) is a classical term weighting TF*IDF formula:

$$wt_{i,j} = tfr_{i,j} \log\left(\frac{N}{df_i}\right) \tag{1}$$

Where, $wt_{i,j}$ corresponds to the weight of the term i in the document j, N is the total number of documents in a collection, $tfr_{i,j}$ is the term frequency of the term i in the document j, and df_i is the document frequency fori in a collection.Consider the positive movie review "The critics have less than kind to 'Sample People' - so I had expectations that the film somewhat of a dud when I saw it. Many of the criticisms of the film are correct; it's a little derivative and quite a messy movie - but that's part of its charm. It's quite brave for an Australian film - it's noisy, colorful and never dull. It contains strong performances from Nathan Page, Ben Mendelsohn, Kylie Minogue, and David Field; and a brilliant soundtrack of Australian artists covering classic Australian songs. The film's production design is excellent - it looks like a Gregg Araki film, and the editing and cinematography are relentlessly brash. It's imperative that people go and support films such as this - a low budget Aussie indie pic because lack of support from critics and lack of distribution and publicity means that it remains unseen by the young adult demographic it is intended. 'Sample People' (brilliant title) conveysright as any film I have seen this year."

After stop words, the terms eliminated are:

The, to, so, i, of, a, it, and, for, an, is, go, as, by.

After stemming the output is given by

"Critic has been less than kind sample people had to expect that film will be somewhat dud when saw mani critic correct little derive quit Messi but part charm brave Australian noisy colour never bore contain strong perform from Nathan page ben mendelsohnliminogudavid field brilliant soundtrack artist cover classic song product design look like greggaraki edit cinematographirelentlessli brash imper support film such the low budget aussiindi pic because lack distribute public mean remain unseen young adult demographics intend title good ani seen year" the Table 2 shows the various term frequencies commonly used:

3.3 Classifiers

In this section Naïve Bayes, a popular probability-based classifier, AdaBoost, an ensemble-based classifier, and a fuzzy-based classifier,FLR described in detail.

3.3.1 Naïve Bayes

It is treated as a variant of Bayes Theorem,which depends on the probabilistic manner of classifier containsdependent assumptions as well as personal assumptions [12, 28]. Hence specified as an autonomousFeatureModel. The Naïve Bayes classifier feels that the features of a class cannot be related to any other features. The classifier's probability

Table 2 Commonly used Term Frequencies

Term	Frequency	Term	Frequency	Term	Frequency	Term	Frequency
adult	1	Indi	1	that	1	subject	1
araki	1	intend	1	titl	1	super	1
artist	1	kind	1	unseen	1	sure	1
aussi	1	Kyli	1	when	1	take	1
australian	1	Lack	1	will	1	term	1
becaus	1	Less	1	year	1	that	1
been	1	Like	1	young	1	there	1
bore	1	Little	1	nikon	1	these	1
brilliant	1	mean	1	onli	1	time	2
budget	1	mendelsohn	1	oper	1	took	1
charm	1	messi	1	optic	1	total	1
cinematographi	1	minogu	1	option	1	troubl	2
colour	1	never	1	part	1	twiddl	1
contain	1	noisi	1	person	1	unfamiliar	1
correct	1	page	1	photo	1	uniqu	1
critic	2	peopl	1	pictur	2	wast	1
david	1	perform	1	pretti	1	week	1
demograph	1	product	1	product	1	well	1
design	1	Quit	1	resolut	1	when	1
distribut	1	relentlessli	1	review	1	while	1
edit	1	remain	1	ring	1	wide	1
excel	1	sampl	1	room	1	wish	1
from	1	soundtrack	1	shot	1	zoom	2
good	1	strong	1	shutter	1		
gregg	1	such	1	spare	1		
have	1	support	1	speed	1		
imper	1	than	1	start	1		

model is conditional over a variable C1 with limited outcomes number or *classes*, conditional on many feature variables Fe_1 through Fe_n.

$$p(C1|Fe_1, ..., Fe_n) \tag{2}$$

The problem is that when features are more in number or depend upon the values of specific functions, such a model depending on the probability tables would be treated as infeasible. Applying Bayes theorem as in Eq. (3):

$$p(C1|Fe_1, ..., Fe_n) = \frac{p(C1)p(Fe_1, \ldots, Fe_n|C1)}{p(Fe_1, \ldots, Fe_n)} \tag{3}$$

Psuedo Code of Naive Bayes Classifier.

1. for q = 1…w // loop for each mining models element
2. $\mu[q] = 0$; // initialization of mining models elements
3. end for;
4. for j = 1…m // loop for each row
5. $\mu[d[j,p]]$++; // increment number of row for value $x_{j,p}$ of object x_j;
6. for k = 1 …p-1 // loop for each column
7. $\mu[\varphi(k\text{-}1)+(d[j, k]\text{-}1)\cdot\varphi(0)+ d[j, p]]$++; // increment number of rows with value $x_{j,k}$ // and value $x_{j,p}$, where $\varphi(k)=s+\sum_{q=1}^{k}(|T_q|\cdot s)$
8. end for;
9. end for;

3.3.2 AdaBoost

This classifier integratesthe sequence of the weighted classifiers that can preferably be imposed on learning various data aspects for generating an aggregated classifier that evaluates the better result of its in-built high probability [29]. The algorithm's necessary steps are.

Step 1: weights are initialized using wt_i=1/n

Step 2: form = 1 to M do

Step 3: fix y_1 = ⟦ht⟧_m (x_1) as the initial weighted classifier using wt_iand d

Step 4: assume

Wt(ht_m)= ∑_(i=1)^N▒⟦Wt_i I{y_iht_m (x_i)=-1}&α_m=log((1-Wt-ht)/(Wt-ht)) ⟧

Step 5: wt_i=wt_i e^(α_m)

Step 6: End

This algorithm is a repetitive process that integrates a greater number oflowest classifiers to estimated Bayes classifier $Cl*(\boldsymbol{x})$. AdaBoost, beginning with an un-weight classifier. The weight increases when a training data point is misclassified. The second classifier having updated its weights would not bethe same. So, boosting

Table 3 Summary of Results using IMDB dataset

	Naïve Bayes	FLRC	AdaBoost
Classification Accuracy	0.8064	0.8115	0.8242
Precision for Negative opinion	0.8295	0.8344	0.8406
Precision for Positive opinion	0.771	0.7765	0.7984
Recall for Negative Opinion	0.8473	0.8505	0.8677
Recall for positive opinion	0.7469	0.7547	0.7609
F Degree for negative opinion	0.7588	0.7654	0.7792
F Degree for positive opinion	0.8383	0.8424	0.8539

the weights of the misclassified training datashould be done as aniterative process that generates 500/1000 classifiers. A classifier that is appropriately provided with a value.

3.3.3 Fuzzy lattice reasoning (FLR)classifier

This classifier considers the rules from the corresponding training data with a continuous incrementofa rule's diagonal size until it reaches the maximum threshold $D_{crit.}$ FLR, is regarded as the leader-follower classifier [16] that rapidly learns with a distinct pass of training data. The input of data illustration order plays aprominent role. This classifier can formulate the learning initiatives without the help of a priori knowledge.But,in further stages, it is necessary to provide the initial rules set. The entire provisionslearnedin advance can remain asan unknownmaterial for identifying via on-line in the learning procedure. Simultaneously, this classifier trains with the additional training information without having any impact onthe previously learned statistics. Therefore, the process is performed by holding the FLRClassifier that is supplied with the latest dataset obtained as a result of improvising the existing rules or by generating new rules. It appears asasingle parameter for tuning, and it should be of the highest threshold size D_{crit}, obtained bybalancingits corresponding granularity in learning.

Various stylesinFLR has been explained below:

- Inassimilationcondition, rule generation may becomeactive, directly with the replacement of a hyperbox$A_J$withmore magnificenthyperbox$A_i \vee A_J$.
- $A_l \rightarrow Cl$, $l = 1,\ldots L$ specifies a fuzzy element $k(x \leq A_l)$ as adomestic of hyperboles,then the hyperbox A_l represents the central point.
- FLRcontains semantics in two different types: Occam razor semantics, as discussedin the earlier phases, and the non-numeric data like graphs retained as a constituent lattice.

Table 4 Summary of Results using Medical Service Dataset

	NB	FLRC	AdaBoost
Classification Accuracy	0.8386	0.8506	0.8554
Precision for Negative opinion	0.7892	0.806	0.8272
Precision for Positive opinion	0.8863	0.8925	0.8795
Recall for Negative Opinion	0.8703	0.8757	0.8541
Recall for positive opinion	0.813	0.8304	0.8565
F Degree for negative opinion	0.8481	0.8603	0.8678
F Degree for positive opinion	0.8278	0.8394	0.8404

- This classifier focused on the missing data in the respective constituent lattice L_i by replacingthe missing datum with appropriate lattice interval [a,b].

FLR Training Algorithm

S0. The first input ($a0$, $C0$) is memorized. At an instant, there are c Known Classes $C1, \ldots, Cc$memorized in memory, initially $c = 0$.
S1. Present next input (ai,Ck), $i = 1, \ldots, m$ to initial "set" family of rules.
S2. If no rules are "set" then
Store input (ai, CK),
$c = c + 1$,
Go to S1.
Else
Compute $k(a0, ai)$, $i = 1, \ldots, c$ of the "set" rules.
S3. Competition among "set" rules:
Winner is rule (aJ, CJ) so that $J = \text{argmax}\{k(a0, ai)\}$, $i = 1, \ldots, c$.
S4. The Assimilation Condition:
Both $Z(ai \vee aJ) \leq \rho$and$Ci = CJ$.
S5. If Assimilation Condition is satisfied then
Replace $aJ$by$a0 \vee aJ$.
Else
"reset" the winner (aJ,CJ), Go to S2.

4 Results and discussion

Various applications were performed to classify the sentiment procedures by making use of the medical services data set and the IMDb data set. Both data sets have positive instances and negative instances. While implementing the classification

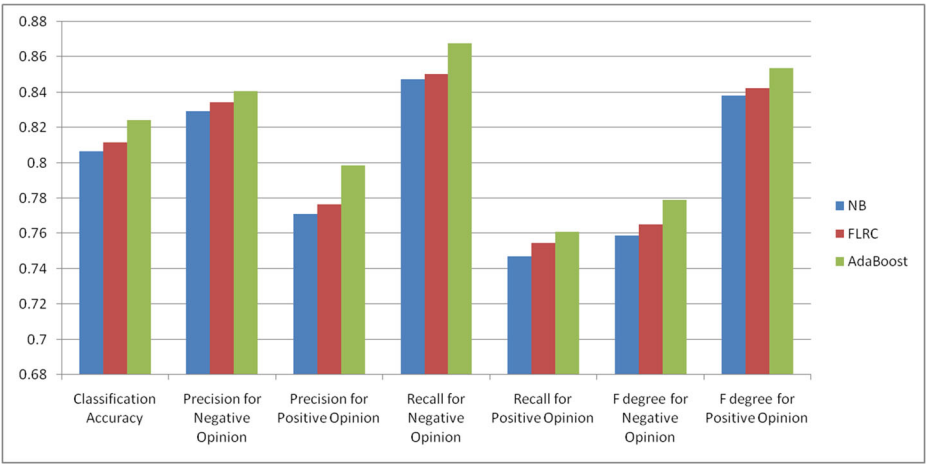

Fig. 1 Implementation of classification in IMDb data set

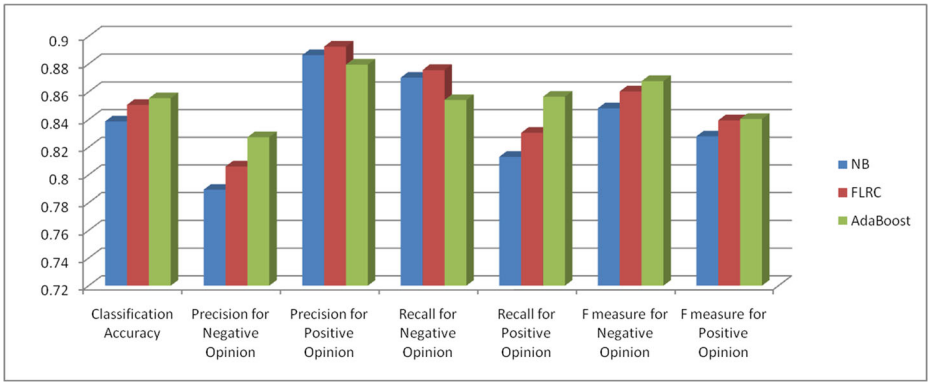

Fig. 2 Implementation of classification in Medicalservices data set

procedure, accuracy is observed to be the primary performance metric for analyzing the positive and negative opinions (Tables 3 and 4).

Fig. 1 demonstrates the implementation of classification in IMDb data set for performing the classification with several metrics like accuracy, precision, recall, F measure respectively. All the performance metricscomprises ofboth the positive and the negative opinions. All the three algorithms are analyzed from the data set and the results proved that AdaBoost algorithm performed better than the other 2 algorithms. Figure 2 also proved the same result while performing the classification on a Medical services data set.

Electronic media get information and advice on medical matters. Internet health information ranges from personal experience in medical condition and patient discussion groups to peer reviewed journal articles and tools for supporting clinical decision making. A study on how American consumers are searching for health-related information shows that the Web is a highly utilized resource for information about health. But it is difficult to locate the best source of knowledge to meet a particular need for information, as relevant information is hidden in web pages or social media data such as blogs and Q&A portals.

This research work suggests a set of semantic related features for OM where the analysis focuses on the expressed sentiment. By extracting and classifying features from reviews, sentiment is classified as positive / negative. Opinion on film review is analyzed / classified as positive / negative. Extraction of features from reviews is via Inverse document frequency and reviews classified using Naive Bayes, AdaBoost, and FLRC classifier. Results showed that the Naïve Bayes has achieved the best rating. To strengthen classification, more supervised learning-based research needs to be undertaken.

5 Conclusion

Sentiment classification is a binary-classification procedure that makes use of the structured reviews for testing, training, or identifying the suitable features andfor scoring the methodologieswith the help of the information retrieval mechanisms for recognizing the review status as positive/negative. In this work, the accessible IMDb movie review dataset has been investigated together with the medical query reviews obtained from the various websites on the internet for devising the user opinions. Thefeature vectors that were generated from the studies were essentially trained with the help of three classifiers, namely, the Naïve

Bayes, the FLR, and the AdaBoost. Classification accuracy of about 82% was achieved for the movie review dataset and about 85% was achieved for the medical query dataset. As far as the practical applications are concerned, an accuracy of 85% appears to be insufficient. An increased level of investigation and study is required for enhancing the accuracy and performance levels of the classification mechanisms.

Compliance with ethical standards

Conflict of interest The authors declare that they do not have any conflict of interests. This research does not involve any human or animal participation. All authors have checked and agreed on the submission.

References

1. Abd Samad Hasan B, Burairah H, GedePramudyaAnanta I, Zeniarja J (2013) Opinion mining of movie review using the hybrid method of support vector machine and particle swarm optimization. J Procedia Eng (Elsevier) 53:453–462
2. Abu-Salih B, Wongthongtham P, Chan KY (2018) Twitter mining for ontology-based domain discovery incorporating machine learning. J Knowl Mangement 22(5):949–981
3. Alaoui E, Gahi Y, Messoussi R, Chaabi Y, Todoskoff A, Kobi A (2018) A novel adaptable approach for sentiment analysis on big social data. J Big Data 5(1)
4. Alsaffar A, Omar N (2014) Study on feature selection and machine learning algorithms for Malay sentiment classification. In information technology and multimedia (ICIMU), 2014 international conference on (pp 270-275). IEEE
5. Angulakshmi G, Chezian MR (2014) An analysis on opinion mining: techniques and tools. Int J Advanced Res Comput Commun Eng 3(7):7483–7487
6. Baccianella S, Esuli A, Sebastiani F (2010) SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC), vol 10, pp 2200–2204
7. Bilal M, Israr H, Shahid M, Khan A (2016) Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, decision tree and KNN classification techniques. J King Saud Univ Comput Inf Sci 28:330–344
8. Chinsha TC, Joseph S (2015) A syntactic approach for aspect based opinion mining. 2015 IEEE international conference on semantic computing (ICSC), Anaheim, CA, pp 24-31. IEEE
9. Claypo N, Jaiyen S (2015) Opinion mining for thai restaurant reviews using K-means clustering and MRF feature selection. 7th International Conference on Knowledge and Smart Technology (KST), Chonburi, (pp 105-108). IEEE
10. Dziczkowski G, Wegrzyn-Wolska K, Bougueroua L (2013) An opinion mining approach for web user identification and clients' behaviour analysis. Fifth International Conference on Computational Aspects of Social Networks, Fargo, ND, (pp 79–84). IEEE
11. Fernández-Gavilanes M, Álvarez-López T, Juncal-Martínez J, Costa-Montenegro E, González-Castaño FJ (2016) Unsupervised method for sentiment analysis in online texts. Expert Syst Appl 58:57–75
12. Ficamos P, Liu Y, Chen W (2017) A naive Bayes and maximum entropy approach to sentiment analysis: capturing domain-specific data in Weibo. In: IEEE international conference on big data and smart computing (BigComp), Jeju, pp 336–339
13. JeevanandamJotheeswaran D, Kumaraswamy Y (2013) Opinion mining using decision tree-based feature selection through Manhattan hierarchical cluster measure. J Theor Appl Inf Technol 58(1):72–80
14. Jeyapriya A, Selvi CS (2015) Extracting aspects and mining opinions in product reviews using supervised learning algorithm. In Electronics and communication systems (ICECS), 2015 2nd international conference on (pp 548–552). IEEE
15. Kao HY, Lin ZY (2010) A categorized sentiment analysis of chinese reviews by mining dependency in product features and opinions from blogs. In web intelligence and intelligent agent technology (WI-IAT), 2010 IEEE/WIC/ACM international conference on vol 1, pp 456-459. IEEE
16. Khezeli YJ, Nezamabadi-pour H (2012) Fuzzy lattice reasoning for pattern classification using a new positive valuation function. Advances Fuzzy Syst 2012:14
17. Li Z, Fan Y, Liu W (2018) Image sentiment prediction based on textual descriptions with adjective noun pairs. Multimed Tools Appl 77(1):1115–1132

18. Li Z, Fan Y, Jiang B (2019) A survey on sentiment analysis and opinion mining for social multimedia. Multimed Tools Appl 78:6939–6967. https://doi.org/10.1007/s11042-018-6445-z
19. Manek AS, Shenoy PD, Mohan MC, Venougopal KR (2017) Aspect term extraction for sentiment analysis in large movie reviews using Gini index feature selection method and SVM classifier. World Wide Web 20(2):135–154
20. Mazzonello V, Gaglio S, Augello A, Pilato G (2013) A study on classification methods applied to sentiment analysis. In semantic computing (ICSC), 2013 IEEE seventh international conference on pp 426-431. IEEE
21. Mountassir A, Berrada I, Benbrahim H (2013) Representing text documents in training document spaces: A novel model for document representation. J Theoretic Appl Inform Technol 56(1)
22. Mukhtar N, Khan MA, Chiragh N (2018) Lexicon-based approach outperforms Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains. Telemat Informatics 35(8):2173–2183
23. Pappas N, Katsimpras G, Stamatatos E (2012) An agent-based focused crawling framework for topic-and genre-related web document discovery. In tools with artificial intelligence (ICTAI), 2012 IEEE 24th international conference on vol 1, pp 508-515. IEEE
24. Peñalver-Martinez I, Garcia-Sanchez F, Valencia-Garcia R, Rodríguez-García MÁ, Moreno V, Fraga A, Sánchez-Cervantes JL (2014) Feature-based opinion mining through ontologies. Expert Syst Appl 41(13):5995–6008. https://doi.org/10.1016/j.eswa.2014.03.022
25. Riaz S, Fatima M, Kamran M, Nasir MW (2017) Opinion mining on large scale data using sentiment analysis and k-means clustering. Cluster Computing 20:1–16
26. Wagh R, Punde P (2018) survey on sentiment analysis using twitter dataset, Proceedings of the 2nd International conference on Electronics Communication and Aerospace Technology (ICECA 2018), pp 208–211
27. Weichselbraun A, Gindl S, Scharl A (2014) Enriching semantic knowledge bases for opinion mining in big data applications. Knowl-Based Syst 69:78–85. https://doi.org/10.1016/j.knosys.2014.04.039
28. Xu L, Lin J, Wang L, Yin C, Wang J (2017) Deep convolutional neural network-based approach for aspect-based sentiment analysis. Advanced Sci Technol Lett 143:199–204
29. Zhang W, Yoshida T, Tang X (2011) A comparative study of TF* IDF, LSI and multi-words for text classification. Expert Syst Appl 38(3):2758–2765

Affiliations

R. Satheesh Kumar [1] · A. Francis Saviour Devaraj [2] · M. Rajeswari [1] · E. Golden Julie [3] · Y. Harold Robinson [4] · Vimal Shanmuganathan [5]

R. Satheesh Kumar
satheeshkumar@sahrdaya.ac.in

A. Francis Saviour Devaraj
saviodev@gmail.com

M. Rajeswari
rajeswarim@sahrdaya.ac.in

E. Golden Julie
goldenjuliephd@gmail.com

Y. Harold Robinson
yhrobinphd@gmail.com

[1] Department of Computer Science and Engineering, Sahrdaya College of Engineering and Technology, Kodakara, Kerala, India

[2] Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Srivilliputhur, India

[3] Department of Computer Science and Engineering, Anna University Regional Campus, Tirunelveli, India

[4] School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India

[5] Department of IT, National Engineering College, Kovilpatti, Tamil Nadu, India