Privacy Preserving Mining of Web Reviews Based on Sentiment Analysis and Fuzzy Sets

[1]Mostafa A. Nofal, [2]Sahar F. Sabbeh and [2]Khaled M. Fouad
[1]*Department of Computer Engineering, College of Computers and Information Technology, Taif University, Kingdom of Saudi Arabia*
[2]*Department of Information Systems, Faculty of Computers and Informatics, Benha University, Egypt*

Key words: Sentiment analysis, sentiments classification, privacy preserving, sentiment feature extraction, fuzzy sets

Corresponding Author:
Mostafa A. Nofal
Department of Computer Engineering, College of Computers and Information Technology, Taif University, Kingdom of Saudi Arabia

Page No.: 122-136
Volume: 19, Issue 7, 2020
ISSN: 1682-3915
Asian Journal of Information Technology
Copy Right: Medwell Publications

Abstract: Customers like online submitting unstructured reviews that has turned out a popular way to come across sentiments across the products purchased or services extradited. In each review, customer typically submit about both the positive and negative opinion of the product, although, the general sentiment toward that product may be positive or negative. The sentiment analysis tries to extract sentiments and subjectivity of customer reviews. These reviews can be beneficial for gathering sentiments of customers about products by analyzing it. However, this analysis should derive careful consideration of customer's anonymity and the privacy of the sensitive data because a privacy is a significant concern for either customers and enterprises. In this research, automatic analysis of sentiment is carried out to achieve such detailed aspects based on domain ontology. Sentiment analysis recognizes the features in the sentiment and classify the sentiments of the review for each of these features. In the proposed approach, the sentiment polarity and polarity strength are provided and computed using fuzzy set. The fuzzy set theory is just effective in processing natural languages because it measures the vagueness. The fuzzy set theory is effective in analyzing reviews which are generally in natural languages. Additionally, the proposed system takes privacy into consideration by masking data before final publishing. The evaluation of the proposed approach is based on using dataset of London restaurant's reviews on TripAdvisor. The evaluation utilizes three different classifiers MLP, SVM and NB and utilizes 5×2-fold cross validation for four evaluation measures; accuracy, precision, recall and F1.

INTRODUCTION

The massive loads of user-provided data; like user reviews are supported by Web. The user-provided data identifies the customer's sentiments associated with merchandise. This data is useful for consumers to support buying decisions and for business associations that aim at supporting the marketing decisions[1]. The marketing decision-supporting is influenced by the opinions provided by conception leaders and ordinary customers. In a marketing, a customer who requires to deal a product online, he discovers the reviews and opinions provided by other customers[2]. The restaurants are one of these marketing[3]. As in latest studies[4], nearly 70% of customers check out reviews of other customers before attending a final deal, 63% of customers are favorable to deal from Web site if it includes a product reviews. The 90% of customer's decisions of customers have modified their views and take a final decision about dealing depended on online reviews[4]. The manual investigating through the massive collection of reviews to acquire useful decisions is very sophisticated and time-consuming issues[5].

Sentiment analysis or opinion mining[6], identifies positivity or negativity scores of a text unit. Sentiment analysis[7] utilizes the Natural Language Processing (NLP) and scientific computation to automatically extract or classify sentiments from customer reviews. The sentiments and opinions analysis has disseminated through many attributes; like consumer information, marketing, books, application, websites and social. Sentiment Analysis is considered as a significant area in decision-support[8]. The main objective of sentiment analysis is processing the reviews and acquire the sentiments' scores. This processing is partitioned into four levels; document[9], sentence[10], word/term[11] or aspect. The processes of sentiment analysis are gradated to sentiment analysis evaluation and sentiment polarity detection[12].

This customer opinion data can be visible as a grey region. This data cannot always be presented into a binary value of yes or no, otherwise it alters on a greyness scale[13]. The benefits of using fuzzy logic is that linguistic values are used to phrase a set and this depends on fuzzy inference rule. The rules; like if-then, utilize a fuzzified variable. Because of the fuzzy set is perfectly effective to process natural languages, to handle the vagueness, it is effective to analyze reviews which are presented using natural languages. In issue of sentiment analysis, fuzzy logic is exploited to represent the polarity scores acquired from the data of customer reviews.

The web-based opinions or sentiments are public and are necessary to be analyzed and understood for a customer democratic process. The decision-makers are supported by public opinions to comprehend your concerned tacit issues that are of ultimate significance for them[13]. This opinion data may contain customer personal data that are private. The majority of opinions are considered sensitive; thus, opinions are released without sufficient identification raise the issues of privacy concern[14].

In this study, the proposed architecture of sentiment analysis depends on these phases; preprocessing of review text using natural language processing, semantic based masking of the user identification using the domain ontology, feature extraction from customer reviews based on keypharase extraction, feature sentiment score calculation using terms expansion based on Wordnet and sentiment's lexicon, sentiment fuzzification, sentiment classification using naive bayes and neural network algorithms.

Background and material

Privacy preserving: There are many methods of privacy preserving for data mining. These contain k-anonymity, supervised learning, unsupervised learning, association rule, distributed privacy preserving, randomization, taxonomy tree, condensation, l-diverse and cryptographic[14]. The privacy preserving for data mining methods safeguard the identification data by altering it to deface the main sensitive one to be stashed. These methods are based on the principal of privacy failure, the capacity to identify the main identification data from amended one, deficiency of information and appreciation of the data accuracy deficiency[15]. The main purpose of these methods is rendering a trade-off through accuracy and privacy. Contrariwise, privacy preserving for data mining utilizes data apportionment and horizontal or vertical distribution of partition among multiple entities[16].

Data anonymization[17] is disregarding a data that would produce sensitive information exposure. This can be accomplished by eliminating the unique identifiers and tackling quasi-identifiers that may produce a unique identification of individuals. Consequently, anonymization utilizes the methods of data suppression, generalization, permutation to alter data that can be used during supplying privacy for sensitive data.

There have been many works for anonymization methods. These methods are based on generalization, suppression[18, 19] or statistical procedures[20]. The most commonly utilized anonymization methods are k-anonymity[21], l-diversity[22] and t-closeness[23].

Fuzzy logic: Fuzzy logic, or fuzzy thinking has suggested by Zadeh[24, 25]. Zadeh deducted that a binary format cannot describe the real world, because it is complicated, they are numerous grey regions, besides data that can be identified as black and white. A binary description can be extended by fuzzy logic to describe occult variables. The approximate reasoning can be provided by fuzzy logic.

In the proposed approach, fuzzy logic is exploited for the representation of the polarity scores attached with linguistic features that identify a certain domain. The main references are provided to the fuzzy logic elements utilized in the residual of that research. The fuzzy logic elements are provided in detail by their mathematical specification in[26].

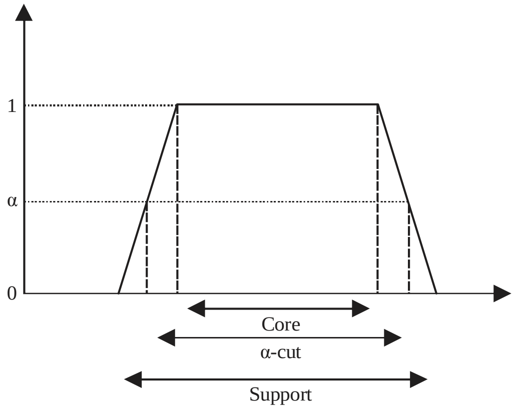

Fig. 1: The elements that form a fuzzy membership function

The values of fuzzy sets are considered as a generalization of values of crisp sets achieved by substituting the characteristic function of a set F, X_z, which appropriates values of {0, 1}; X_z (x) = 1 if $x \in F$, $X_z(x)$ = 0 otherwise by a function called membership function μ_f which can postulate any value in 0, 1. The value $\mu_f(x)$ or F(x) is the membership score of element x in F; the score is where x belongs in F. A fuzzy set is perfectly identified by its membership function. The elements that form a fuzzy membership function is shown in Fig. 1.

From a fuzzy set[27] F in Fig. 1, the core is the set of elements x where, F(x) = 1; the support sup(F) is the set of elements x where F(x)>0. The set of all elements x of F where, $F(x) \geq \alpha$, for a given a $\alpha \in 0$, 1, is named the α-cut of F, symbolized F_α.

Semantic resources

Ontology: Ontologies[28] are utilized in platforms that required to reuse data contents and to be used for reasoning, contrariwise, just utilized for presenting information. They enable machine's interpretability of data content by expanding supplementary vocabulary along a formal semantics. The main purpose for ontology[29] is enabling connection between computer applications in a trend that is independent of the technology of the system, information structure and the domains. Ontology involves affluent relationships between concepts and the specific domain. The structure of the ontology is constructed using mapped ontology. The ontology encompasses issues such as artificial intelligence, data structures, database, programming, etc.

WordNet: WordNet[30] is a great lexical database for English terms. It was constructed in 1986 in Princeton University. The fact is given that talkers have knowledge about tens of thousands of terms and the concepts associated with these words. It pretends reasonable to suppose efficient and economic storage and access techniques for terms and concepts. The model of Collins and Quillian had a hierarchical structure of concepts to support the inheritance. The particular knowledge to specific concepts requires to be saved and associated with such concepts. Therefore, it occupied subjects longer to emphasize a statement like "canaries have feathers" than the statement "birds have feathers". WordNet is indicated to as an ontology; indeed, some philosophers handling ontology have assessed WordNet's upper structure and commented on it.

Sentiment lexicon: Many researches that are addressing the issues of sentiment analysis utilized lexicons which are exploited for the sentiment involved in a set of terms. These terms, known as opinion terms are used with parsing process in order to acquire the user's sentiment. The lexicon also conserves a set of objective expressions, which do not provide any opinion. These objective expressions are utilized to discover the focus comment's references. As well, the terms in the lexicon were gathered by hand from actual comments allowing the colloquial; non-standard, terms are acquired. Other works have proposed lexicons that depend not on standard dictionaries, boosting colloquial language and multi-term expressions[31].

Valence Aware Dictionary for sentiment Reasoning "VADER" is an unpretentious rule-based model for generic sentiment analysis. VADER conserves the advantages of traditional sentiment lexicons; like LIWC[32], yet just as purely inspected, understood, readily applied and facilely extended. In a similar way, LIWC and VADER sentiment lexicons are gold-standard quality and have been manually evaluated and validated; human-validated. VADER differentiates itself from LIWC that it is further sensitive to sentiment terms for contexts of social media and propagates favorably.

Feature's terms expansion: In the Features Expansion (FE), the input feature term is extended and enriched by concatenating supplementary features that assemble different relationships between the main features of the two objects[33]. This feature expansion is presented in the previous work[33] but in this context, the usage is considerably different.

The core idea of FE is identifying the missing terms in reviews vector representation if it can be subrogated with semantically related term[34]. This procedure aims to enhance the process of acquiring the scores of each feature in the sentiment lexicon "VADER".

In this study, original terms in the review's vector are input and output is a set of semantically similar "synonym" related to each term in the original terms in the review's vector. This method is performed using WordNet and Word Sense Disambiguation (WSD)[35, 36].

Sentiments classification: Sentiment analysis is utilized to determine and acquire the subjective information from these user's reviews. In the sentiment analysis, the scores of each term existed in a review are determined. Subsequently, sentiments of terms should be classified to demonstrate the final user sentiment either positive, negative or neutral at various levels. Therefore, various classification methods can be utilized[37]. These methods include Linear regression and rule based approach[38].

There were systems utilized Naive Bayesian classifier with sentiment analysis for classification[39]. SVM overestimated the Bayesian classifier[40] when SVM and Bayesian classifier are compared for user's reviews classification. Additionally, many of those methods cannot capture the meaning of user's sentiment. To evaluate such sentiments, fuzzy classifiers and fuzzy set theory is efficient to check the ambiguity[41, 42].

Literature review: The opinion mining methodology[43] was proposed to exploit advantages of Semantic Web-guided solutions to improve the outcomes achieved with traditional NLP techniques and sentiment analysis procedures. The basic objectives of the proposed methodology were improving feature-based opinion mining based on ontologies at the feature selection stage and providing a method for sentiment analysis based on vector analysis-based.

The method that aimed to contextualize and enrich massive semantic based knowledge bases for opinion mining was proposed[44]. The method was effective to universal, multi-dimensional affective resources. It involved these steps; identifying ambiguous sentiment words, providing context information acquired from a specific domain training corpus and grounding this contextual information to structured knowledge sources; like ConceptNet and WordNet.

The common and common-sense knowledge were integrated together to construct a universal resource that was considered as an attempt to simulate how implicit and explicit knowledge is regulated in the humanitarian mind. This was utilized to accomplish reasoning through sentiment analysis[45].

The senti-lexicon was proposed for the sentiment analysis of reviews about the restaurants[46]. When a review document was classifyed as a positive and a negative sentiments by using a method of the supervised learning algorithm, there was a trend to increase the accuracy of positive classification higher than the accuracy of negative classification. The improved Naive Bayes algorithm is proposed to alleviate such issue.

The domain specific sentiment lexicon is presented and is applied for extracting sentiment feature[47]. The effective features for sentiment classification are extracted by using generative uni-gram mixture model based domain specific sentiment lexicon learnt by utilizing emotion text of labelled blogs and tweets.

The reduction of the vocabulary mismatch with word embedding was presented[48]. The features were expanded by using Word2vec. Word2vec tries to combine words with points in space. The spatial distance between words then idenities the similarity association between these words. Two processes are provided to achieve the word's similarity. The first process utilizes the neighboring words to foresee a word target. The second process utilizes a word to foresee the neighboring words in a sentence.

The dictionary-based classification was proposed for accurate classification of the reviews. Support vector machine algorithm is performed to improve the accuracy of the classification of neutral reviews. The quality of the product was identified based on the sentiment graph that was provided for the product's reviews.

SentiWordNet was incorporated as the labeled training corpus to extract the sentiment scores on the part of speech data. A vocabulary SentiWordNet-V with scores of reviewed sentiments, acquired from SentiWordNet, was utilized for Support Vector Machines model[49].

The sentiment analysis[50] was employed to extract required information from a blog to examine the level of customer goodwill for the services of aviation and non-aviation. The feedbacks proposed that travelers concentrate their evaluation on a limited set of services regarding food and drink and the shopping area.

The achievement of domain independent lexicons was improved integrating machine learning and a lexical based approach to identify the weight of a feature based on SentiWordNet. Support vector machine is utilized for the feature scores learning and an intelligent selection approach was exploited to enhance the classification accuracy. Considerably, the subjectivity was used to select the features and the effects of POS on feature selection were presented[51].

The metaheuristic method (CSK) was proposed based on K-means and cuckoo search. The proposed method was exploited to achieve the optimum cluster-heads in the sentimental data of Twitter[52].

Proposed architecture: The proposed architecture aims at enhancing the solution of sentiment analysis by enrich the solution by using privacy preserving to perform the anonymization of the user identification by using masking technique that is based on ontology-based generalization. The sentiment analysis is performed by using feature's extraction and using features and terms expansion based on WordNet. The fuzzy logic is used to tackle the vagueness in the sentiments' scores for each feature.

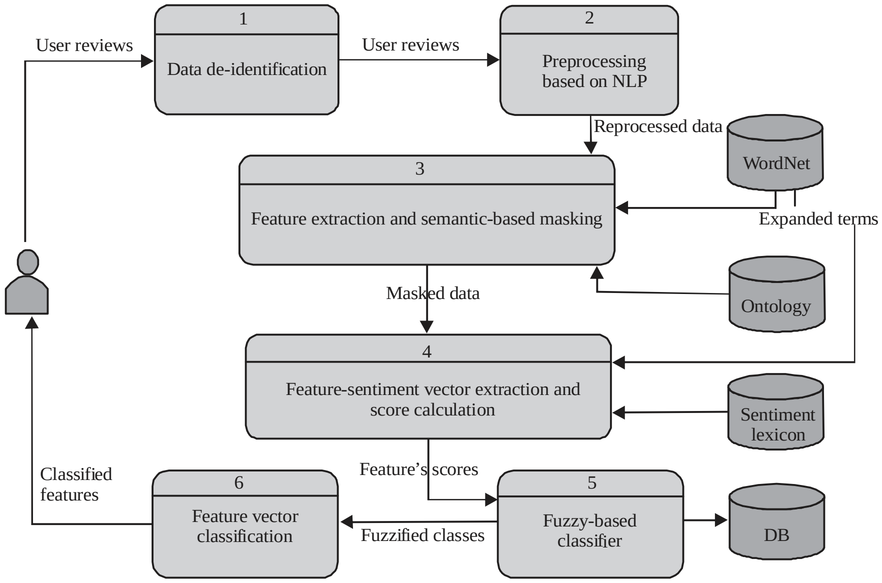

Fig. 2: Key components of the proposed architecture

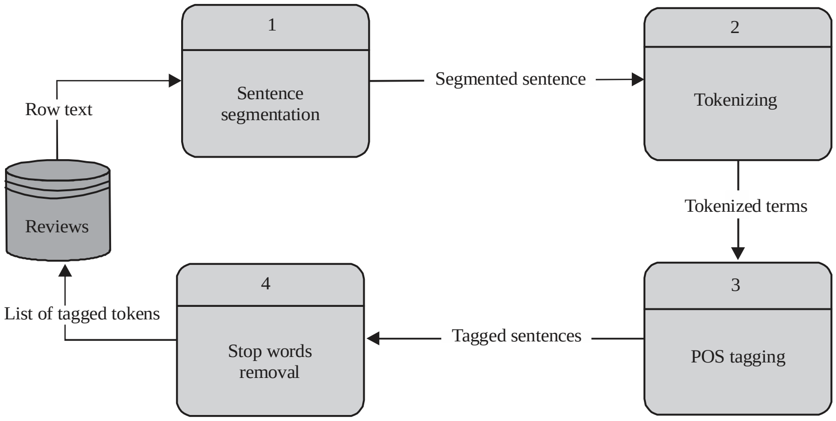

Fig. 3: Main components of text preprocessing based on NLP

In the proposed architecture, the user reviews should firstly de-identified. The NLP is exploited to preprocess the de-identified data of reviews. Using NLP to prepare the reviews terms to the next steps of sentiment analysis procedures. The features will be extracted using domain ontology of restaurants and can be expanded by using WordNet. The scores of extracted features' vector are generated by using sentiment lexicon; VADER. If review's terms may be not found in the lexicon, the expanded terms should be acquired which are extracted from the WordNet, to enrich the term vector and to enhance the procedure of acquiring the scores of each feature in the sentiment lexicon. The sentiment scores of each feature can be fuzzified to handle the vagueness in the sentiments' scores. The classification algorithm is applied to classify the sentiments based on the fuzzified sections that provide linguistic values. Figure 2 shows the key process of the proposed architecture (Fig. 3).

Data de-identification: The first step in data processing is to anonymize data to ensure data de-identification. The collected data usually include some personally and/or quasi identifiable information. Personally, Identifiable Information (PII) is any piece of data that can uniquely identify a specific person such as: name, email address, Social Security Number (SSN), telephone number, fax number, etc. Where quasi-identifiers are pieces of information that are not considered to be unique identifiers for themselves but can create one if combined with other quasi-identifiers such as: postal code, job, gender, age, birthdate, location and timestamp, etc., de-identification can be achieved by replacing identifiers with random values or recode the variables (age or age range instead of date of birth) or by simply dropping the identifying columns[53]. For the proposed system, de-identification was achieved by removing all the PIIs and quasi-identifiers from the data.

Natural Language Processing NLP: During this step, NLP techniques are used to identify the morphologic and syntactic structure of each sentence. This step includes a sentence segmentation component, tokenizer, POS tagging component and stop words removal components as shown in Fig. 3.

The sentence segmentation component: This component is responsible for determining the sentence boundaries to split a paragraph into sentences for further processing. Sentence segmentation components consider the local context of the punctuation. Question marks and interjection points are unambiguous boundaries markers unlike periods which can be ambiguous as it can be a part of an abbreviation (Mr., Av., a.m., A.S.A.P, com, etc.) that's why an abbreviation dictionary must be attached. Sentence segmentation step is shown in Algorithm 1.

Algorithm 1; Sentence segmentation:

```
Inputs: P: list of punctuation marks
         A: List of abbreviations
        W: raw text (string)
Output: N : list pf sentences
Step 1: start
Step 2: Initialize sentence list N [], i = 0, start = 0, EOS = false
Step 3: for each word w in W
        Step 3.1: i = index(w)
        Step 3.2 if exists blank-line after w then EOS = true
        Step 3.3 Elseif if i+length(w)+1ЄP and i+length(w)+1=`?” or
        i+length(w)+1= “!” then EOS=true
        Step 3.4 Elseif i+length(w)+1 = “.” Then
            Step 3.4.1 if wЄA then
                   EOS = false
        Step 3.4.2 Else
                 EOS = true
        Step 3.4.3 End if
Step 3.5 Else
        EOS = false
Step 3.6 End if
Step 3.7 if EOS = true then
         Step 3.7.1 length=i – start+length(w)
         Step 3.7.2 n = substring(W, start, length)
         Step 3.7.3 N[i] = n
         Step 3.7.4 start = i+length(w)+1
               Step 3.8 End if
Step 4 Loop
```

Tokenizer: The segmented sentences from the previous phase are received by this component which iterates over all sentences of each paragraph and identifies the basic elements/tokens of the sentence to be processed (i.e., words, phrases, symbols, etc.). The correctness of the tokenization can affect the whole text analysis process. Standard algorithms usually split tokens in text based on white spaces which is not always true as tokens are not always detached by white space characters. A boundary period at the end of a sentence does not belong to the last token while a period at the end of an abbreviation belongs to the token. Additionally, some contexts require the identification of units that do not need to be decomposed. for the proposed system we chose the low-level-tokenization algorithm which splits text into tokens according to the definition in a grammar file. The low-level tokenizer takes into consideration abbreviations and hyphenated words which can guarantee a high accurate tokenization of the text as shown in Algorithm 2.

Algorithm 2; Low-level tokenization:

```
Inputs: P: list of punctuation marks
           A: List of abbreviations
           L : list of lexical hyphen words
           S: sentence
Output: T : list of tokens
Step 1: start
Step 2: j=0
Step 3: W = split(s, “ “)//split sentence into word array based on
whitespace
Step 3: for each word w in W
        Step 3.1: i = index(w)
        Step 3.4 if i+length(w)+1 = “.” Then
             Step 3.4.1 if wЄA then
                  Token = w + “.”
              Step 3.4.2 Else
                   Token = w
              Step 3.4.3 End if
Step 3.5 Elseif i – 1 = “.” Then
          Step 3.5.1 if w ? A then
               Token = “.” + w
           Step 3.5.2 Else
              Token = w
           Step 3.5.3 End if
Step 3.6 elseif i+length(w)+1 = “-” Then
           Step 3.6.1 if wЄL then
             Token = w + “-“ + W[w+1]
           Step 3.6.2 Else
               Token = w
            Step 3.6.3 End if
   Step 3.7 End if
   Step 3.8 T[j] = token
   Step 3.9 j = j+1
Step 4 Loop
```

The Part-of-Speech (PoS) tagger: This component is responsible for marking text tokens with their corresponding type (i.e. noun, verb, adjective, etc.). In the proposed system, RDRPOSTagger[54] is used. RDRPOSTagger is based on an incremental knowledge acquisition technique where rules are modified on error. RDRPOSTagger provides a competitive accuracy compared to other POS taggers.

Stop words removal: This component is responsible for removing the common words that have no significance in the text analysis task. Stop-words carry no meaning in natural language such as articles, prepositions and conjunctions are natural candidates for a list of stop-words. For the sake of this study a customized version of the stop words list has been used. As, using a generic list of stop words can have a negative impact on sentiment analysis performance[55]. Removing some common stop words like “don’t”, “not”, “couldn’t” can change sentiment of a sentence.

Feature extraction and semantic-based masking: The main role of this component is to identify and extract keywords, anonymize and mask textual features by using semantic-based generalization. This component performs three main tasks keyword extraction, term expansion and textual feature anonymization as follows:

Keyword extraction: Keywords extraction aims to identify and extract the most informative terms from a specific text. In the proposed system, we used an unsupervised approach for keywords extraction from reviews text. The proposed system depends on the keywords extraction approach in[56] which depends on both statistical and linguistic features of text terms. The algorithm includes three main steps: preparing dictionary of distinct entries, mapping dictionary entries with Wikipedia titles and ranking entries.

Preparing dictionary of distinct items: In this step, a hieratical n-gram dictionary of distinct terms together with their co-occurrence frequency with other terms is built. The algorithm utilizes LZ78 compression technique[57] to handle words generated from previous stages. The tokenized text from previous stage is used to construct a bigram dictionary. If the pattern does not have an index in the dictionary, it should be added with a frequency value of “one”; otherwise the frequency of pattern is incremented by “one”. Each entry in the dictionary is assigned two different scores. The first core is the frequency of occurrence where, the second is the influence weight which is a frequency times calculated according to a grammatical rule by Kumar and Srinathan[58]. The grammatical rule favor noun phrases which appear earlier or at the end of sentences. The later score is calculated according to Eq. 1:

$$0 \le p_0 < \frac{N_i}{2} \text{ or } p_0 > \left(\frac{3 \times N_i}{4}\right) \qquad (1)$$

where, N_i is a number of terms in sentence I and 0 is an index of first term in phrase p in the sentence.

Mapping dictionary entries with, Wikipedia titles: Wikipedia titles are extracted and assigned for each dictionary entry. Additionally, a confidence value equal to 1 is assigned to that entry to indicate that this entry is considered as a verified Wikipedia concept; otherwise it will be assigned to value of zero.

Ranking entries: The bulk of key words ranking algorithms depend solely on key phrase frequency. Other algorithms such as n-gram filtration technique[56, 58], calculate the influence of key phrase according to number of grammatical rules. The entries are ranked according to Eq. 2[56]:

$$\text{Rank(i)} = \log\left(p_i \times \frac{TF_i + Tl_i}{L} + CF_i\right) \qquad (2)$$

where, p_i is the position of dictionary entry i. The position is computed by $p_i = (L-L_s)$ where, L is a total number of lines in document and L_s is the first sentence where a dictionary entry i occurs. TF_i, Tl_i and CF_i indicate, respectively the term frequency, influence weight and confidence factor of Wikipedia for dictionary item i.

Term expansion: A challenging task is detecting sentiments in user-generated content as text may include some terms that are not commonly used or even terms that are ambiguous. Thus, in order to best identify the sentiments in text, we perform semantic expansion of lexical terms using WordNet ontology. Terms that are semantically close to the main key terms are identified using WordNet which can be used to obtain a list of synonymous words by an iterative process given an initial set of terms and after calculating the spreading activation[59]. Spreading activation aims to identify the activation origin node which represent the concept of the given term. Next, nodes one link away are activated, then nodes that are two links away and so on. During this iterative process, activation score of node (j) is calculated based on three factors as in Eq. 3: a constant C_{dd} is a dimension discount that causes a node closer to the activation origin to get a higher score; the activation score of node I and W(i,j), the weight of the link from I-j:

$$\text{Activation}_{\text{score}(j)} = C_{dd}{}^{*} \sum_{i \in \text{neighbor}(j)} \text{Activation}_{\text{score}(i)} + W(i,\ j) \qquad (3)$$

The top N words with the higher activation scores are then are selected as the expanded terms.

Feature extraction and generalization: The purpose of this step is identifying features included in the review text, mask those textual features using concept generalization to ensure anonymization. This is performed by identifying the concepts of the ontology that identified to review terms. Concept identification is performed using the intersection of the local context of the analyzed term with every identical ontology entry. A domain ontology is utilized for extracting the features involved in the review text. Features are assorted in congruence with their semantic measure and are assigned to a basic concept of the domain ontology[60]. For our restaurant domain we use ambience/atmosphere, service, food, drinks, Price,

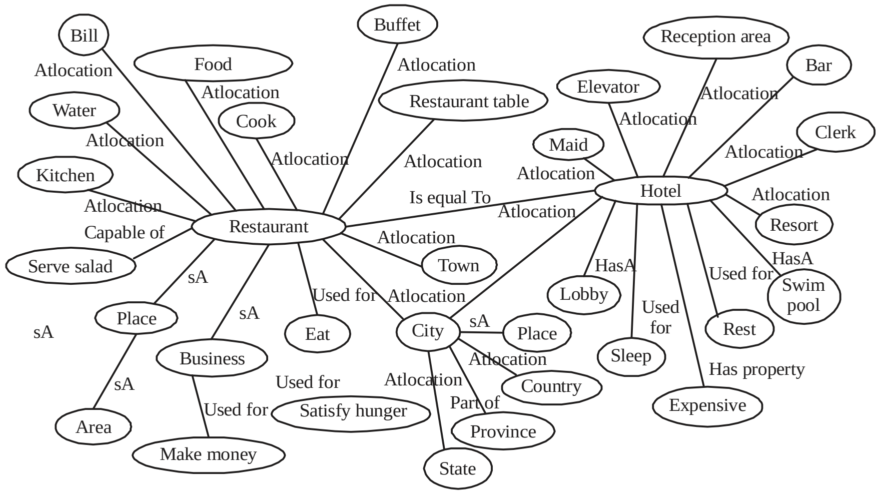

Fig. 4: The restaurant's domain ontology

comfort and noise level. The synonymous extracted from WordNet are used to find individuals of top-level class that have a matching concept. When a concept is found, we include all its types as features. For example, when we find the concept of "steak", top level concepts are also including such as meat and food.

Feature-sentiment vector extraction and score calculation: In this step, extracted sentiments and their synonymous are associated with each feature. Then a sentiment lexicon is used to retrieve each sentiment score. A final score is calculated to each specific feature and used to define its membership to a certain sentiment level in the next stage. The proposed performs negation handling as sentiments extracted are associated with some adverbs that represent positive or negative sentiment (i.e., don't, not, never, etc.). The system changes the orientation of the sentiment score by reversing the sign of the score, as if a positive sentiment is proceeded by a negation word, score is converted to the negative and vice versa as in Eq. 4. The final sentiment score associated with each feature is calculated by Eq. 5:

$$\text{score(s)} = \begin{cases} \text{k} & \text{if } \text{s-1} \notin \text{N} \\ \text{-k} & \text{if } \text{s-1} \in \text{N} \end{cases} \tag{4}$$

Where:

Score(s) = The final score of sentiment
s, N = List of negation words
k = Score of s in the sentiment lexicon

$$\text{SScore(f)} = \sum_{\text{s in S}} \text{score(s)} \tag{5}$$

Where:

Sscore(f) = The final sentiment score of feature
f, S = The sentiment list associated with feature f
Score(s) = The sentiment score of sentiments

Fuzzy logic-based classifier: Fuzzy logic techniques have advantages for tackling issues of ambiguity and imprecision for terms utilized in natural language. The fuzzy based technique is applied in the proposed solution over the extracted set of the feature's sentiment scores to obtain the fuzzified features' sentiment scores. After the features' sentiment scores are determined for each attribute, the generated scores are assessed over the different developed rules. This process requires to check and compare each attribute in each review, the combinations of the terms and the scores of current reviews.

The input of the identified membership functions are numeric values or vectors which are crisp. The membership functions involve the real concepts of the linguistic terms. The primary sentiment value for items in the sentiment synset list of membership functions has acquired from sentiment lexicon.

In the proposed solution, the fuzzy sections are determined as four linguistic values; low, fair, medium and high. In this course, the membership functions can be determined and achieved using the certain fuzzy sections. Since the nature of the input of the identified linguistic terms will frequently provide indiscriminate sentiment combinations that suited a Gaussian distribution. The Gaussian function is used to determine the membership functions[61].

Restaurant domain ontology: The domain of the applied ontology is carefully done so it will fit the restaurants that are privileges hence they involve to specific protocols. The restaurants' domain and ranges are indicated, as well as the sub classes as illustrated in Fig. 4.

Based on the Semantic Web, a class is a collection of resources with similar properties. In the proposed

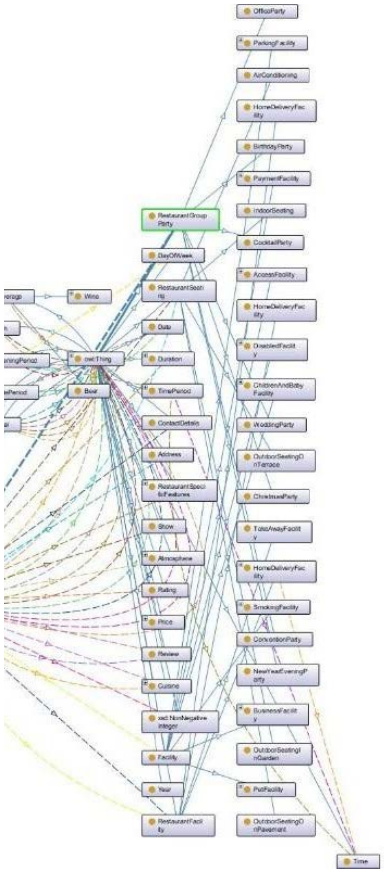

Fig. 5: The snapshot of the ontology in protegee tool

architecture, the ontology has various classes, which include; for example, Staff, Expenses, Inventory, Minu, Booking, Customer, Takeaway, Address etc. These classes have subclasses of their own and some of them have other subclasses. In this ontology, the Staff class involves the restaurant employees. Staff has subclasses, regrouping all the types of employees. The Customer class involves customer attributes; like, name, phone number and email. The Booking class involves the reservation attributes. The restaurant's domain ontology can be downloaded from the Web site link (https://www.disi.unige.it/person/LocoroA/download/wilfontologies/restaurant.owl). Figure 5 shows the snapshot of the ontology in Protegee tool.

MATERIALS AND METHODS

Experiments design

Dataset: The used dataset contains London restaurant's reviews on TripAdvisor. The dataset contains 19999 reviews represented by 16 variables and one response variable. The target variable labels each restaurant to be of level from 1-5. The dataset contains 2773 instances with no label which were disregarded. The remainder 17223 instances include approximately 44% rated 5, 31% rated 4, 12% rated 3, 7% rated 2 and 6% rated 1. The dataset contains missing data as shown dataset metadata in Table 1.

Data transformation:

- The variables "URL, restaurant id, restaurant_location, name, title and category" were removed for their irrelevance to the sentiment analysis problem
- Service, food and value are removed as they have >50% missing values
- De-identification was achieved by removing all personally identifiable information: author_name, author_URl and author_location and quasi-identifiers: "restaurant name, visited on, review date
- The proposed system was applied on review_text field to extract the following weighted related features: (cleanliness, menu, atmosphere, comfort, safeness, noise level, speed, service, cost, taste, drinks, food and location) as presented in Table 2

Explanatory data analysis: This step aims mainly to discover patterns or correlation between variables. The pair-wise correlation among variables indicated low or no correlation among all of the variables as shown in Fig. 6.

Variable selection: In this step, the most informative features were selected to reduce dimensionality before model training. Features were evaluated and ranked using the model in[62] which uses an iterative permutation process to measure the effect of each feature on the label. The features are then ranked based on their mean decrease importance based on which, features are either confirmed or refused. After the iterative process, 7 attributes were confirmed: comfort, cost, drinks, food, location, taste and service while 5 attributes were rejected: atmosphere, cleanliness, menu, noiselevel, speed as shown in Table 3 and Fig. 7.

Performance measures: The performance of the selected model's predictive power is evaluated based on accuracy, precision, recall and F-measure (F1).

Accuracy: Indicates the ability of the model to classify reviews to their accurate rate. Accuracy os calculated for each rate label individually. It's the proportion of True Positive (TP) and True Negative (TN) in all evaluated reviews:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \qquad (6)$$

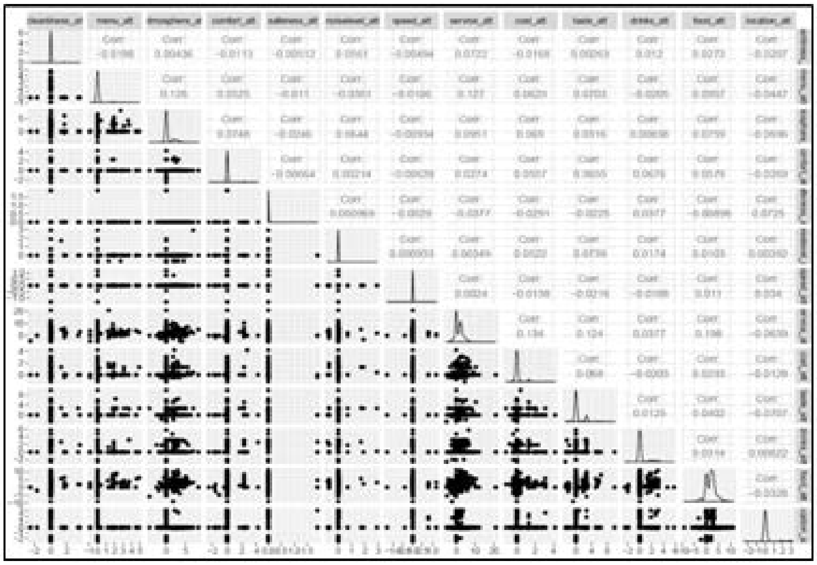

Fig. 6: Pairwise correlation matrix

Table 1: Dataset metadata

Variables	Types	Description	Missing data (%)
Uique_id	Nominal	Id for each review	0
url	Nominal	Review URL	0
restaurant_id	Nominal	Unique id for each restaurant	0
restaurant_location	Nominal	Location of the reviewed restaurant	0
Name	Nominal	Restaurant name	0
Category	Categorical	Review type (restaurants, hotels…etc)	0
Title	Nominal	Title of the review	11
Review_date	Date	The date on which review was written	11.5
Review_text	Nominal	The textual content of user review	11
Author	Nominal	Reviewer name	11.3
Author_URL	Nominal	User URL	12.65
Location	Nominal	Author location	27.5
Visited_on	Date	The date of the visit to the restaurant	16.6
Rating	Ordinal	Label the restaurant rate on scale from 1 to 5	13.8
Food	Ordinal	Food rating	55.88
Value	Ordinal	Rating of the value of the experience	54.88
Service	Ordinal	Rating of the service	54.3

Table 2: Extracted features metadata

Variable	Type	Description
service_att	Numeric	Score of user sentiments of the restaurant services
taste_att	Numeric	Sentiment score of food taste
comfort_att	Numeric	Sentiment score to indicate to what degree the restaurant was comfortabl
food_att	Numeric	Sentiment score of the food quality
location_att	Numeric	Sentiment score of the restaurant location
drinks_att	Numeric	Sentiment score of the rinks quality
cost_att	Numeric	Score to indicate user sentiment of the cost
safeness_att	Numeric	Sentiment score of safety of the restaurant
atmosphere_att	Numeric	Score of the user sentiments of the restaurant atmosphere
menu_att	Numeric	Sentiment score of the menu
speed_att	Numeric	Score of the service speed
cleanliness_att	Numeric	Score of the cleanness of the restaurant
noiselevel_att	Numeric	Sentiment score of the noise level around the place

Where:

TP = The total number of reviews correctly classified to be of rate R

FP = The total number of reviews incorrectly classified to be of rate R

TN = The total number of reviews correctly classified not to be of rate R

FN = The total number of reviews incorrectly classified not to be of rate R

Precision and recall: Precision and recall can give a better insight in the performance as they do not assume equal misclassification costs. Precision indicates is the fraction of reviews correctly classified among all

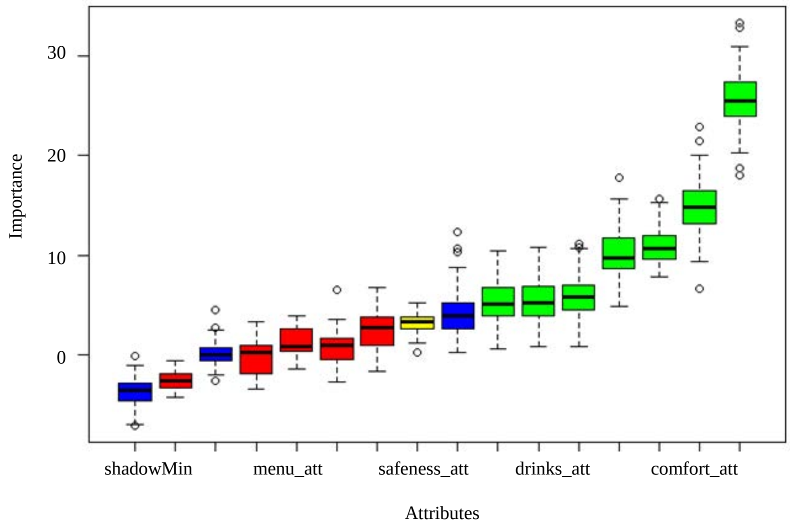

Fig. 7: Mean decrease importance of the variables

Table 3: Mean decrease importance of variables

Variables	meanImp	Decision
service_att	25.60621	Confirmed
taste_att	14.82524	Confirmed
comfort_att	10.91944	Confirmed
food_att	10.15451	Confirmed
location_att	5.842265	Confirmed
drinks_att	5.4439	Confirmed
cost_att	5.332758	Confirmed
safeness_att	3.164489	Rejected
atmosphere_att	2.636999	Rejected
menu_att	1.316091	Rejected
speed_att	0.902282	Rejected
cleanliness_att	-0.27399	Rejected
noiselevel_att	-2.51113	Rejected

Table 4: Parameters values

Models	Parameters	Tuning values after fuzzification	Tuning values before fuzzification
MLP	Learning function	Std_Backpropagation	Std_Backpropagation
	Maximum iterations(maxit)	100	100
	Initial weight matrix (initFunc)	Randomized_Weights	Randomized_Weights
	number of units in the hidden layer(size)	[1, 3 , 5]	[1, 3 , 5]
SVM	δ	0.04727892	0.3180782
	C (cost of penalty)	[0.25, 0.50 , 1.00]	[0.25, 0.50 , 1.00]
NB	FL	0	0
	Userkernel	Yes	Yes
	adjust	1	1

classified instances while recall is the fraction of reviews correctly classified over the total number of reviews in the rate R:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{8}$$

F-measure: F-measure (F1) is calculated based on a combination of both precision and recall providing a better evaluation of predictive performance:

$$F_1 = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}\times\text{Recall}} \tag{9}$$

Model training and validation: The selected model was first trained using the dataset which was split into 80% for training and validation and 20% for testing.

For model training and validation, 5×2-fold cross validation was applied as recommended by Dietterich[63] and Kursa *et al*.[64] Initial parameters are tuned via grid search during the training stage. The optimal parameter values are selected based on cross validated accuracy as shown in Table 4.

Table 5: Performance evaluation of the models before and after fuzzy

	Before fuzzification				After fuzzification			
Models	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
MLP	0.4824	0.5781	0.7957	0.6697	0.7812	0.7812	1.0000	0.8772
SVM	0.4623	0.4932	0.7849	0.6058	0.7812	0.7812	1.0000	0.8772
NB	0.3618	0.5426	0.5484	0.5455	0.7812	0.7812	1.0000	0.8772

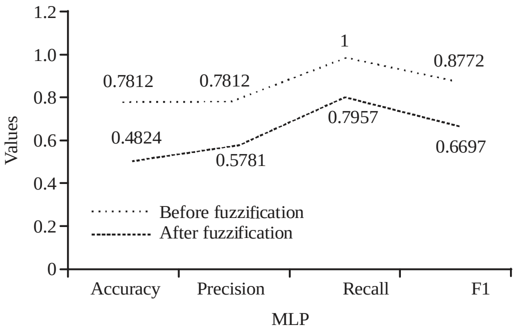

Fig. 8: Performance of MLP

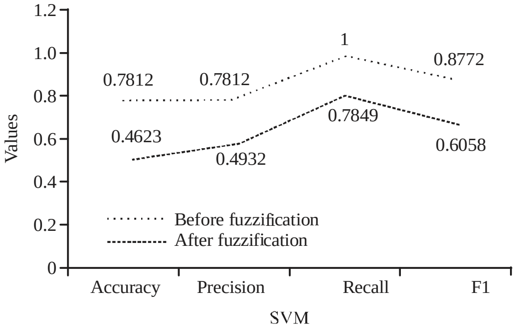

Fig. 9: Performance of SVM

RESULTS AND DISCUSSION

The experiment was performed using an acer machine with 64-bit Windows 10 OS, Intel®Core™ i7-7500U CPU @ 2.70GHZ and 8 GB Memory using R language. In order to test the performance of the selected models, unlabeled 20% of the dataset was used as an input to the trained models for performance evaluation. Results of testing are used to compare the models based on predictive performance in terms of the selected metrics as shown in Table 5.

The results presented in Table 5 show that data fuzzification enhances the predictive power of all the used classification models. Results show that multilayer perceptron MLP achieved high performance compared to the other used models followed by SVM while NB comes at the last of the list. Results indicate that fuzzification, increases the predictive power of the chosen models by approximately 30% in terms of accuracy and 21% in precision, recall and f -measure. The performances of each model before and after fuzzification is shown in Fig. 8-10.

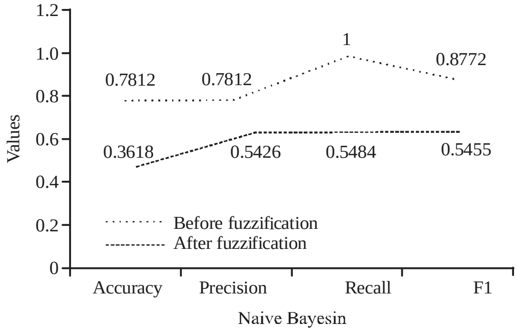

Fig. 10: Performance of NV

CONCLUSION

Sentiment analysis has the capability to determine the scores of the positivity or negativity of a review text. Sentiment analysis exploits the Natural Language Processing (NLP) and computational methods to extract or classify sentiments from unstructured customer reviews.

In the proposed architecture, the sentiment analysis enhancement is based on exploiting many methods. These methods are feature extraction using keyphrase extraction to extract the features as keyphrase from a short review. During the sentiment scores are acquired from the reviews, the review term associated to a feature is expanded using WordNet. The expansion of the term enriches the term mapping with the sentiment lexicon. In the proposed architecture, the fuzzy set approach is exploited to enhance the classification by applying the fuzzification for each extracted feature. The fuzzification has the ability to substitute each attribute numerical value to linguistic value.

Furthermore, the proposed system exploit advantages of privacy by masking the private data to anonymize the sensitive customer data. The masking method of generalization based on domain ontology are exploited to anonymize quasi-identifiers to preserve the balance between data utility and customer privacy. The experimental results provided in this work showed that data fuzzification improves the predictive result of all the used classification models. Results showed that achieved MLP is high performance compared to the other utilized models followed by SVM while NB comes at the last of the list. Results indicate that fuzzification, increases the predictive power of the chosen models by approximately 30% in terms of accuracy and 21% in precision, recall and f-measure.

In next trends, the enhancement approach for fully automated feature extraction from the text is required to improve the sentiment feature extraction from text. Also, the required enhancement in the future is improving the feature selection and classier of the sentiment results.

REFERENCES

01. Kirange, K. and R. Ratnadeep, 2016. Aspect and emotion classification of restaurant and laptop reviews using SVM. Int. J. Curr. Res., 8: 28352-28356.
02. Chinsha, T.C. and S. Joseph, 2014. Aspect based opinion mining from restaurant reviews. Intl. J. Comput. Appl., 1: 1-4.
03. Perera, I.K.C.U. and H.A. Caldera, 2017. Aspect based opinion mining on restaurant reviews. Proceedings of the 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), September 8-11, 2017, IEEE, Beijing, China, pp: 542-546.
04. Ling, P., C. Geng, Z. Menghou and L. Chunya, 2014. What do seller manipulations of online product reviews mean to consumers?. HKIBS Working Paper Series 070-1314, Hong Kong Institute of Business Studies, Hong Kong.
05. Nithin, Y.R. and G. Poornalatha, 2017. Feature Based Opinion Mining for Restaurant Reviews. In: Signal Processing and Intelligent Recognition Systems, Thampi, S., S. Krishnan, C.J. Rodriguez, S. Das, M. Wozniak and D. Al-Jumeily (Eds.). Springer, Cham, pp: 305-318.
06. Pang, B. and L. Lee, 2008. Opinion mining and sentiment analysis. Found. Trends Inform. Retrieval, 2: 1-135.
07. Agarwal, B., N. Mittal, P. Bansal and S. Garg, 2015. Sentiment analysis using common-sense and context information. Comput. Intell. Neurosci., 2015: 1-9.
08. Koehler, M., S. Greenhalgh and A. Zellner, 2015. Potential applications of sentiment analysis in educational research and practice-is SITE the friendliest conference?. Proceedings of SITE 2015 Society for Information Technology & Teacher Education International Conference, March 02, 2015, Association for the Advancement of Computing in Education (AACE), Waynesville, pp: 1348-1354.
09. Yessenalina, A., Y. Yue and C. Cardie, 2010. Multi-level structured models for document-level sentiment classification. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, October 9-11, 2010, Association for Computational Linguistics, Cambridge, Massachusetts, pp: 1046-1056.
10. Farra, N., E. Challita, R.A. Assi and H. Hajj, 2010. Sentence-level and document-level sentiment mining for Arabic texts. Proceedings of the 2010 IEEE International Conference on Data Mining Workshops, December 13, 2010, IEEE, Sydney, Australia, pp: 1114-1119.
11. Engonopoulos, N., A. Lazaridou, G. Paliouras and K. Chandrinos, 2011. ELS: A word-level method for entity-level sentiment analysis. Proceedings of the International Conference on Web Intelligence, Mining and Semantics, May 25-27, 2011, ACM, Sogndal, Norway, pp: 1-9.
12. Khan, K., B. Baharudin, A. Khan and A. Ullah, 2014. Mining opinion components from unstructured reviews: A review. J. King Saud Univ. Comput. Inf. Sci., 26: 258-275.
13. Sattar, A.S., J. Li, X. Ding, J. Liu and M. Vincent, 2013. A general framework for privacy preserving data publishing. Knowl. Based Syst., 54: 276-287.
14. Aldeen, Y.A.A.S., M. Salleh and M.A. Razzaque, 2015. A comprehensive review on privacy preserving data mining. SpringerPlus, 4: 694-730.
15. Zhuojia, X. and Y. Xun, 2011. Classification of privacy-preserving distributed data mining protocols, Proceedings of the IEEE 6th International Conference on Digital Information Management, September 26-28, 2011, Melbourne, Australia..
16. Ciriani, V., S.D.C. Vimercati, S. Foresti and P. Samarati, 2008. K-Anonymous Data Mining: A Survey. In: Privacy-Preserving Data Mining, Aggarwal, C.C. and P.S. Yu (Eds.). Springer, Boston, USA., pp: 105-136.
17. Emam, K.E. and F.K. Dankar, 2008. Protecting privacy using k-anonymity. J. Am. Med. Inf. Assoc., 15: 627-637.
18. Liu, J. and K. Wang, 2010. Anonymizing transaction data by integrating suppression and generalization. Proceedings of the 14th Pacific Asia Conference on Knowledge Discovery and Data Mining, June, 2010, Hyderabad, India, pp: 171-180.
19. Mahesh, R. and T. Meyyappan, 2013. Anonymization technique through record elimination to preserve privacy of published data. Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22, 2013, IEEE, Salem, India, pp: 328-332.
20. Anbazhagan, K., D. Sugumar, M. Mahendran and R. Natarajan, 2012. An efficient approach for statistical anonymization techniques for privacy preserving data mining. Int. J. Adv. Res. Comput. Commun. Eng., 1: 482-485.
21. Sweeney, L., 2002. Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertainty Fuzziness Knowledge-Base Syst., 10: 571-588.

22. Ding, X., B. Liu and P.S. Yu, 2008. A holistic lexicon-based approach to opinion mining. Proceedings of the 2008 International Conference on Web Search and Data Mining, February 11-12, 2008, ACM, New York, pp: 231-240.
23. Li, N., T. Li and S. Venkatasubramanian, 2007. T-closeness: Privacy beyond k-anonymity and l-diversity. Proceedings of the IEEE 23rd International Conference on Data Engineering ICDE 2007, April 15-20, 2007, IEEE, Istanbul, Turkey, ISBN: 1-4244-0802-4, pp: 106-115.
24. Zadeh, L.A., 1965. Fuzzy sets. Inform. Control, 8: 338-353.
25. Enache, I.C., 2015. Fuzzy logic marketing models for sustainable development. Bull. Transilvania Univ. Brasov. Econ. Sci. Ser., Vol. 8,
26. Ferrara, E., O. Varol, C. Davis, F. Menczer and A. Flammini, 2016. The rise of social bots. Commun. ACM., 59: 96-104.
27. Dragoni, M., A.G. Tettamanzi and D.C.C. Pereira, 2014. Propagating and aggregating fuzzy polarities for concept-level sentiment analysis. Cogn. Comput., 7: 186-197.
28. Dubois, D. and H. Prade, 2002. Possibility theory, probability theory and multiple-valued logics: A clarification. Ann. Math. Artif. Intell., 32: 35-66.
29. Valls, A., K. Gibert, D. Sanchez and M. Batet, 2010. Using ontologies for structuring organizational knowledge in home care assistance. Int. J. Med. Inform., 79: 370-387.
30. Christiane, F., 2010. WordNet. In: Theory and Applications of Ontology: Computer Applications, Roberto, P., H. Michael and K. Achilles (Eds.)., Springer, Netherlands, pp: 231-243.
31. Chua, S. and N. Kulathuramaiyer, 2005. Semantic feature selection using WordNet. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), September 20-24, 2004, IEEE, Beijing, China, pp: 166-172.
32. Hutto, C.J. and E. Gilbert, 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the 8th AAAI International Conference on Weblogs and Social Media (ICWSM'14), June 1-4, 2014, Association for the Advancement of Artificial Intelligence, Palo Alto, California?, USA., pp: 216-225.
33. Durak, E.S. and M. Durak, 2011. The development and psychometric properties of the turkish version of the stress appraisal measure. Eur. J. Psychol. Assess., 29: 64-71.
34. Lopez-Inesta, E., F. Grimaldo and M. Arevalillo-Herraez, 2016. Combining feature extraction and expansion to improve classification based similarity learning. Pattern Recog. Lett., 93: 95-103.
35. Setiawan, E.B., D.H. Widyantoro and K. Surendro, 2016. Feature expansion using word embedding for tweet topic classification. Proceedings of the 2016 10th International Conference on Telecommunication Systems Services and Applications (TSSA), October 6-7, 2016, IEEE, Denpasar, Indonesia, pp: 1-5.
36. Lee, W.J. and E. Mit, 2011. Word sense disambiguation by using domain knowledge. Proceedings of the 2011 International Conference on Semantic Technology and Information Retrieval, June 28-29, 2011, IEEE, Putrajaya, Malaysia, pp: 237-242.
37. Fouad, K.M., A.R. Khalifa, N.M. Nagdy and H.M. Harb, 2012. Web-based semantic and personalized information retrieval. Int. J. Comput. Sci. Issues (IJCSI.), 9: 266-276.
38. Vyas, V. and V. Uma, 2018. An extensive study of sentiment analysis tools and binary classification of tweets using rapid miner. Proc. Comput. Sci., 125: 329-335.
39. Yao, T. and L. Li, 2009. A kernel-based sentiment classification approach for Chinese sentences. Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering Vol. 5, March 31- April 2, 2009, IEEE, Los Angeles, USA., pp: 513-518.
40. Xuan, H.N.T. and A.C. Le, 2013. Linguistic features for subjectivity classification. Proceedings of the 2012 International Conference on Asian Language Processing, November 13-15, 2012, IEEE, Hanoi, Vietnam, pp: 17-20.
41. Su, X., G. Gao and Y. Tian, 2010. A framework to answer questions of opinion type. Proceedings of the 2010 7th Web Information Systems and Applications Conference, August 20-22, 2010, IEEE, Hohhot, Mongolia, pp: 166-169.
42. Jusoh, S. and H.M. Alfawareh, 2013. Applying fuzzy sets for opinion mining. Proceedings of the 2013 International Conference on Computer Applications Technology (ICCAT), January 20-22, 2013, IEEE, Sousse, Tunisia, pp: 1-5.
43. Dalal, M.K. and M.A. Zaveri, 2014. Opinion mining from online user reviews using fuzzy linguistic hedges. Applied Comput. Intell. Soft Comput., Vol. 2014, 10.1155/2014/735942
44. Penalver-Martinez, I., F. Garcia-Sanchez, R. Valencia-Garcia, M.A. Rodriguez-Garcia, V. Moreno, A. Fraga and J.L. Sanchez-Cervantes, 2014. Feature-based opinion mining through ontologies. Expert Syst. Appl., 41: 5995-6008.
45. Weichselbraun, A., S. Gindl and A. Scharl, 2014. Enriching semantic knowledge bases for opinion mining in big data applications. Knowl. Based Syst., 69: 78-85.

46. Cambria, E., Y. Song, H. Wang and N. Howard, 2013. Semantic multidimensional scaling for open-domain sentiment analysis. IEEE. Intell. Syst., 29: 44-51.
47. Kang, H., S.J. Yoo and D. Han, 2012. Senti-lexicon and improved Naive Bayes algorithms for sentiment analysis of restaurant reviews. Exp. Syst. Appl., 39: 6000-6010.
48. Bandhakavi, A., N. Wiratunga, D. Padmanabhan and S. Massie, 2016. Lexicon based feature extraction for emotion text classification. Pattern Recog. Lett., 93: 133-142.
49. Abinaya, R., P. Aishwaryaa, S. Baavana and N.T. Selvi, 2016. Automatic sentiment analysis of user reviews. Proceedings of the 2016 IEEE Conference on Technological Innovations in ICT for Agriculture and Rural Development (TIAR), July 15-16, 2016, IEEE, Chennai, India, pp: 158-162.
50. Khan, F.H., U. Qamar and S. Bashir, 2016. eSAP: A decision support framework for enhanced sentiment analysis and polarity classification. Inf. Sci., 367-368: 862-873.
51. Gitto, S. and P. Mancuso, 2017. Improving airport services using sentiment analysis of the websites. Tourism Manage. Perspect., 22: 132-136.
52. Khan, F.H., U. Qamar and S. Bashir, 2016. SWIMS: Semi-supervised subjective feature weighting and intelligent model selection for sentiment analysis. Knowl. Based Syst., 100: 97-111.
53. Pandey, A.C., D.S. Rajpoot and M. Saraswat, 2017. Twitter sentiment analysis using hybrid cuckoo search method. Inf. Process. Manage., 53: 764-779.
54. Ito, K., J. Kogure, T. Shimoyama and H. Tsuda, 2016. De-identification and encryption technologies to protect personal information. Fujitsu Sci. Tech. J., 52: 28-36.
55. Nguyen, D.Q., D.Q. Nguyen, D.D. Pham and S.B. Pham, 2016. A robust transformation-based learning approach using ripple down rules for part-of-speech tagging. AI Commun., 29: 409-422.
56. Jean-Louis, L., A. Zouaq, M. Gagnon and F. Ensan, 2014. An assessment of online semantic annotators for the keyword extraction task. Proceedings of the Pacific Rim International Conference on Artificial Intelligence, December 1-5, 2014, Springer, Cham, 548-560.
57. Amer, E. and K. Foad, 2016. Akea: An Arabic keyphrase extraction algorithm. Proceedings of the International Conference on Advanced Intelligent Systems and Informatics, October 26-28, 2016, Springer, Singapore, 137-146.
58. Pu, I.M., 2005. Fundamental Data Compression. Butterworth-Heinemann, Oxford, England, UK., ISBN-13:978-0-7506-6310-6, Pages: 246.
59. Kumar, N. and K. Srinathan, 2008. Automatic keyphrase extraction from scientific documents using N-gram filtration technique. Proceedings of the 8th ACM Symposium on Document Engineering, September 16-19, 2008, Sao Paulo, Brazil, pp: 199-208.
60. Hsu, M.H., M.F. Tsai and H.H. Chen, 2006. Query expansion with conceptnet and wordnet: An intrinsic comparison. Proceedings of the Asia Information Retrieval Symposium, October 16-18, 2006, Springer, Singapore, 1-13.
61. Schouten, K., F. Frasincar and F. De Jong, 2017. Ontology-enhanced aspect-based sentiment analysis. Proceedings of the International Conference on Web Engineering, June 5-8, 2017, Springer, Rome, Italy, 302-320.
62. Bing, L. and K.C. Chan, 2015. A fuzzy logic approach for opinion mining on large scale twitter data. Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, December 8-11, 2014, IEEE, London, UK., pp: 652-657.
63. Dietterich, T.G., 2002. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput., 10: 1895-1923.
64. Kursa, M.B., A. Jankowski and W.R. Rudnicki, 2019. Boruta-a system for feature selection. Fundamenta Inf., 101: 271-285.