Journal of Theoretical and Applied Information Technology
30th September 2018. Vol.96. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PRIVACY PRESERVING MINING OF WEB REVIEWS BASED ON SENTIMENT ANALYSIS AND FUZZY SETS

[1]MOSTAFA A. NOFAL, [2,3] SAHAR F. SABBEH, [4] KHALED M. FOUAD

[1]Computer Engineering Dept., College of Computers and Information Technology, Taif University, Kingdom of Saudi Arabia

[2] Information Systems Dept., Faculty of Computers & Information technology, University of Jeddah, KSA.

[3] Information Systems Dept., Faculty of Computers & Informatics, Benha University, Egypt, sahar.fawzy@fci.bu.edu.eg

[4] Information Systems Dept., Faculty of Computers & Informatics, Benha University, Egypt, kmfi@fci.bu.edu.eg

ABSTRACT

In the traditional Web, users are considered as information consumers. In social Web, users play a much more active role since they are now not only information consumers but also data providers. Users like online posting reviews which has become an increasingly popular way to express opinions and sentiments toward the products bought or services received. Analyzing these reviews can be helpful for collecting opinions of people about products, social events and problems and would produce useful actionable knowledge that could be of economic values to vendors and other interested parties.

Thus, due to the huge number of reviews and their unstructured nature, efficient computational methods are needed for mining and summarizing these reviews, because regular analysis of reviews does not indicate user likes and dislikes. In a review, user typically writes about both the positive and negative aspects of the object, although the general sentiment toward that object may be positive or negative. That's why sentiment analysis together with opinion mining try to extract and study of user's opinions, sentiments and subjectivity of text.

However, this analysis must come with careful consideration of user's anonymity and the privacy of their sensitive data as privacy is today an important concern for both users and enterprises.

In this research, automatic analysis of opinions (opinion mining) is performed to obtain such detailed aspects based on ontology. Opinion mining identify the features in the opinion and classify the sentiments of the opinion for each of these features. Opinion mining is a difficult task, owing to both the high semantic variability of the opinions expressed, and the diversity of the characteristics and sub-characteristics that describe the products and the multitude of opinion words used to depict them.

In the proposed approach, the opinion polarity and polarity strength are measured using fuzzy set. As the fuzzy set theory is quite effective in processing natural languages, to measure the vagueness, it will also be effective in analyzing review articles, which are generally in natural languages. Additionally, the proposed system takes privacy into consideration by anonymizing data before final publishing. Methods of generalization and micro-aggregation are utilized for anonymizing quasi-identifiers to maintain the balance between data utility and user privacy.

Keywords: *Sentiment Analysis, Sentiments Classification, Privacy Preserving, Sentiment Feature Extraction, Fuzzy Sets.*

1. 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. 90% of customer's decisions of customers have modified their views and take a final decision about dealing depended on online reviews [4].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

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, 9]. The main objective of sentiment analysis is processing the reviews and acquire the sentiments' scores. This processing is partitioned into four levels [10]; document [11], sentence [12], word/term [13] or aspect [14]. The processes of sentiment analysis are gradated to sentiment analysis evaluation and sentiment polarity detection [15].

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 [16]. 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 [16]. 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 [17].

In this paper, the proposed architecture of sentiment analysis depends on these phases; review text preprocessing 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.

The organization of this paper is as follows. In section 2, Background and material. Section 3, related work. Section 4, architecture of the proposed system. Section 5, experimental results. Finally, section 6 the conclusions and future works will be summarized.

2. BACKGROUND AND MATERIAL

2.1 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 [17]. 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 [18]. 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 [19].

Data anonymization [20] 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 [21].

There have been many works for anonymization methods. These methods are based on generalization, suppression [22, 23], or statistical procedures [24]. The most commonly utilized anonymization methods

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

are k-anonymity [25], l-diversity [26] and t-closeness [27].

2.2 Fuzzy logic

Fuzzy Logic, or fuzzy thinking has suggested by Zadeh [28, 29]. 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 [30].

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 figure 1.

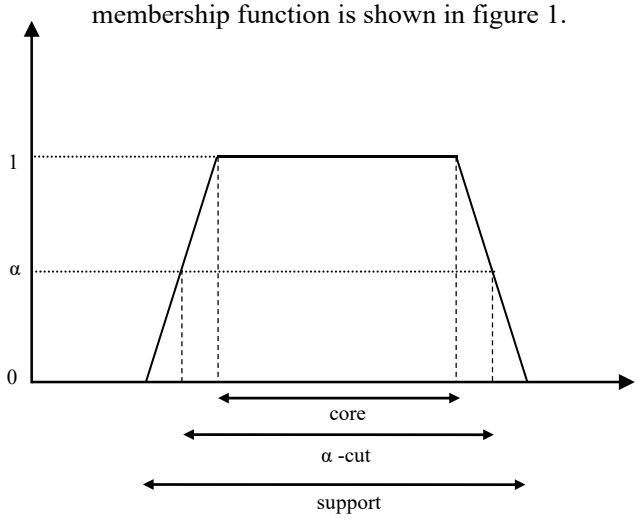

Figure 1: The Elements That Form A Fuzzy Membership Function

From a fuzzy set [31] F in figure 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 $\alpha \in (0, 1$, is named the α-cut of F, symbolized F_α.

2.3 Semantic resources

2.3.1 Ontology

Ontologies [32] are utilized in platforms that required to reuse data contents, and to be used for reasoning, contrariwise, just utilized for presenting information. They enable machines' interpretability of data content by expanding supplementary vocabulary along a formal semantics. The main purpose for ontology [33] 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.

2.3.2 WordNet

WordNet [34, 35] 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.

2.3.3 Sentiment lexicon

Many researches that are addressing the issues of sentiment analysis utilized lexicons which are exploited for the sentiment involoved in a set of terms. These terms, known as opinion terms, are used with parsing process in order to acquire the users' 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 comments' references. As well, the terms in the lexicon was gothered by hand from actual comments allowing the colloquial; non-standard, terms is acquired.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Other works have proposed lexicons that depend not on standard dictionaries, boosting colloquial language and multi-term expressions [36].

Valence Aware Dictionary for sEntiment Reasoning "VADER" is an unpretentious rule-based model for generic sentiment analysis [37]. VADER conserves the advantages of traditional sentiment lexicons; like LIWC [38], 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.

2.4 Features' 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 [39]. This feature expansion is presented in the previous work [39], 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 [40]. This procedure aims to enhance the process of acquiring the scores of each feature in the sentiment lexicon "VADER".

In this paper, original terms in the reviews' vector are input and output is a set of semantically similar "synonym" related to each term in the original terms in the reviews' vector. This method is performed using WordNet [34] and word sense disambiguation (WSD) [41, 42].

2.5 Sentiments classification

Sentiment analysis is utilized to determine and acquire the subjective information from these users' 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 [43]. These methods include Linear regression and rule based approach [44].

There were systems utilized Naïve Bayesian classifier with sentiment analysis for classification [45]. SVM overestimated the Bayesian classifier[46], when SVM and Bayesian classifier are compared for users' reviews classification. Additionally, many of those methods cannot capture the meaning of users' sentiment. To evaluate such sentiments, fuzzy classifiers and fuzzy set theory is efficient to check the ambiguity [47, 48].

3. Related Work

The opinion mining methodology [49] 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 [50]. 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 [51].

The senti-lexicon was proposed for the sentiment analysis of reviews about the restaurants [52]. 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 naïve bayes algorithm is proposed to alleviate such issue.

The domain specific sentiment lexicon is presented and is applied for extracting sentiment feature [53]. 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 [54]. The features were expanded by using word2vec. Word2vec attempts to associate words with points in space.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The spatial distance between words then describes the similarity relation between these words. Two processes are provided to achieve the words' 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 [55]. 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 [56]. The sentiment analysis [57] 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 & 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 weights learning and an intelligent selection approach was exploited to improve the classification accuracy. Considerably, the subjectivity was used to select the features and the effects of POS on feature selection were presented [58].

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 [59].

4. 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 features' 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.

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 reviews' 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.

Journal of Theoretical and Applied Information Technology
30th September 2018. Vol.96. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

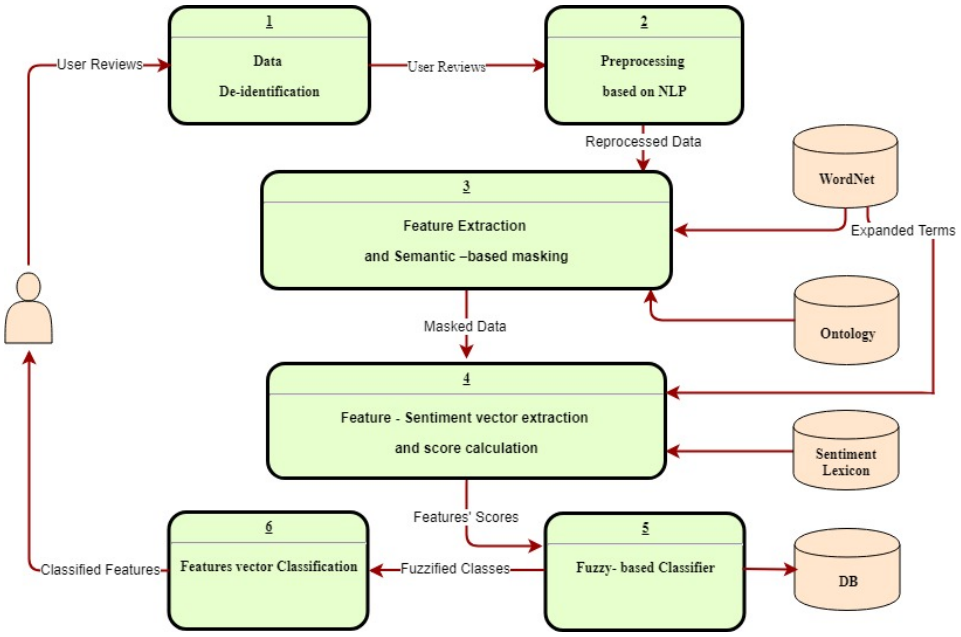

Figure 2: Key Components Of The Proposed Architecture

4.1 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 [60]. For the proposed system, de-identification was achieved by removing all the PIIs and quasi-identifiers from the data.

4.2 Natural Language processing NLP

During this step, NLP techniques are used to identify the morphologic and syntactic structure of each sentence. This step includes 1) a sentence segmentation component, 2) tokenizer, 3) POS tagging component, and 4) stop words removal components as shown in Figure 3.

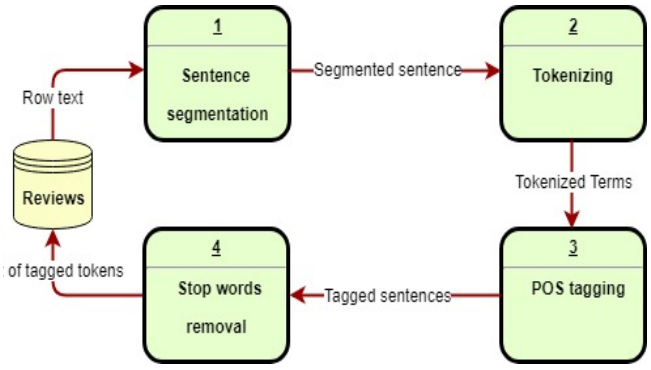

Figure 3: Main Components Of Text Preprocessing Based On NLP

4.2.1 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 takes into account the local context of the punctuation (periods, exclamation points, question marks). Question marks and exclamation 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

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

attached. Sentence segmentation step is shown in algorithm.1.

```
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)+1 Є  P 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 w Є A 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
```

Algorithm 1: Sentence Segmentation

4.2.2 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 separated 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.

```
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 w Є A 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 Є 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 w Є L 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
```

Algorithm 2: Low – Level Tokenization

4.2.3 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 [61] 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.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.2.4 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 [62]. Removing some common stop words like "don't", "not", "couldn't" can change sentiment of a sentence.

4.3 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 1) keyword extraction, 2) term expansion, and 3) textual feature anonymization as follows:

4.3.1 keyword extraction

Keywords extraction aims to identify and extract the most informative terms from a specific text [63]. 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 [64] which depends on both statistical and linguistic features of text terms. The algorithm includes three main steps:
1) preparing dictionary of distinct entries, 2) mapping dictionary entries with Wikipedia titles, and
3) ranking entries.

4.3.2 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 [65] 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 [66]. The grammatical rule favor noun phrases, which appear earlier or at the end of sentences. The later score is calculated according to the equation 5:

$$0 \leq p_0 < \frac{N_i}{2} \quad Or \quad p_0 > \left(\frac{3\ x\ N_i}{4}\right) \ldots\ldots (1)$$

Where N_i is a number of words in sentence I and p_0 is an index of first word in phrase p in the sentence.

4.3.3. 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.

4.3.4 Ranking Entries

The bulk of key words ranking algorithms depend solely on key phrase frequency. Other algorithms such as n-gram filtration technique [64, 66], calculate the influence of key phrase according to number of grammatical rules. The entries are ranked according to equation.6 [64].

$$Rank(i) = log\left(p_i\, x\, \frac{TF_i + TI_i}{L} + CF_i\right) \ldots (2)$$

Where p_i is the position of dictionary entry i. The position is calculated as $p_i = (L - L_s)$, 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, TI_i and CF_i indicate respectively the term frequency, influence weight, and Wikipedia confidence factor for dictionary entry i.

4.3.5 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 [67]. Spreading activation aims to identify the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

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 equation (4): (i) a constant C_{dd} which is distance discount that causes a node closer to the activation origin to get a higher score; (ii) the activation score of node I, and (iii) W(i,j), the weight of the link from I to j.

$$Activation_{score(j)} = C_{dd} * \sum_{i \epsilon\, Neighbor(j)} Activation_{score(i)} + W(i,j) \quad (3)$$

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

4.3.6 Feature extraction and generalization

The purpose of this step is to identify features included in the review text, mask those textual features using concept generalization to ensure anonymization. This is performed by identifying the ontology concepts that correspond to review words. Concept identification is based on the overlap of the local context of the analyzed word with every corresponding ontology entry. A domain ontology is used in order to extract the features included in the review text. Features are grouped in accordance with their semantic distance and are then attached to a main concept of the domain ontology [68]. For our restaurant domain we use ambience/atmosphere, service, food, drinks, Price, 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.

4.4 Feature - Sentiment vector extraction and score calculation

In this step, extracted sentiments and their synonymous are associated with each feature. Then a sentiment lexicon [37] 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 equation 5. The final sentiment score associated with each feature is calculated by equation.6:

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

Where; Score(s) is the final score of sentiment s, N is list of negation words, and k is score of s in the sentiment lexicon.

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

Where; Sscore(f) is the final sentiment score of feature f, S is the sentiment list associated with feature f, and Score(s) is the sentiment score of sentiment s.

4.5 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 features' 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 [69].

Journal of Theoretical and Applied Information Technology
30th September 2018. Vol.96. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.6 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 the figure 4.

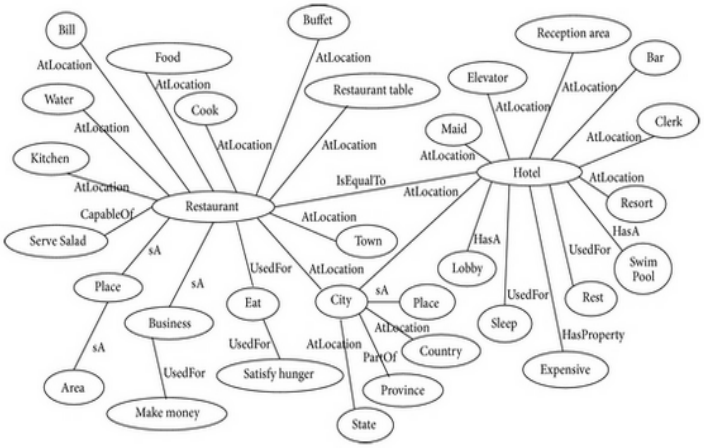

Figure.4. The Restaurants' Domain Ontology

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

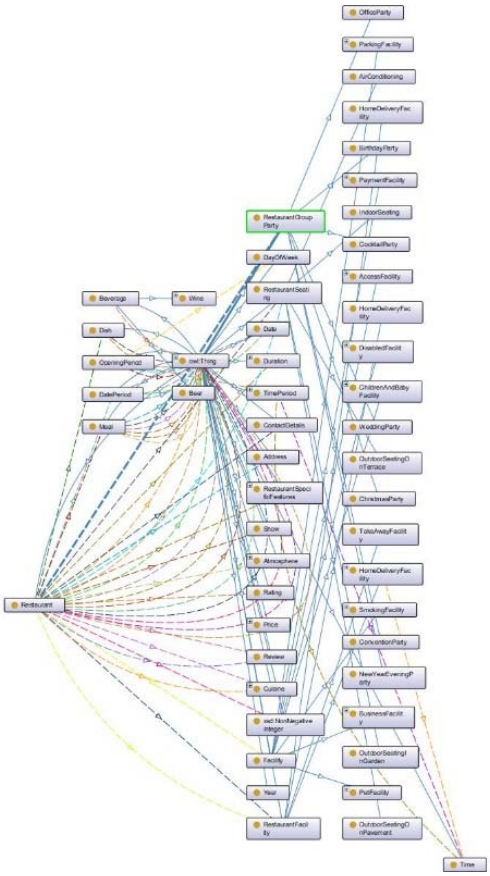

Figure.5. The Snapshot Of The Ontology In Protegee Tool

4. EXPERIMENTS DESIGN

I. METHODOLOGY

1- Dataset

The used dataset contains London restaurants' 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 to 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 in Table.2.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table.2. Dataset Metadata.

Variable	Type	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%

2. Data transformation

- The variables "URL, restaurant id, restaurant_location, name, title, and
- Service, food and value are removed as they have more than 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. category" were removed for their irrelevance to the sentiment analysis problem.
- The proposed system was applied on review_text field to extract the following weighted related features: (cleanliness, menu, atmosphere, comfort, safeness, noiselevel, speed, service, cost, taste, drinks, food, and location) as presented in table.3

Table.3. 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 comfortable
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

Journal of Theoretical and Applied Information Technology
30th September 2018. Vol.96. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

a. 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 Figure.6.

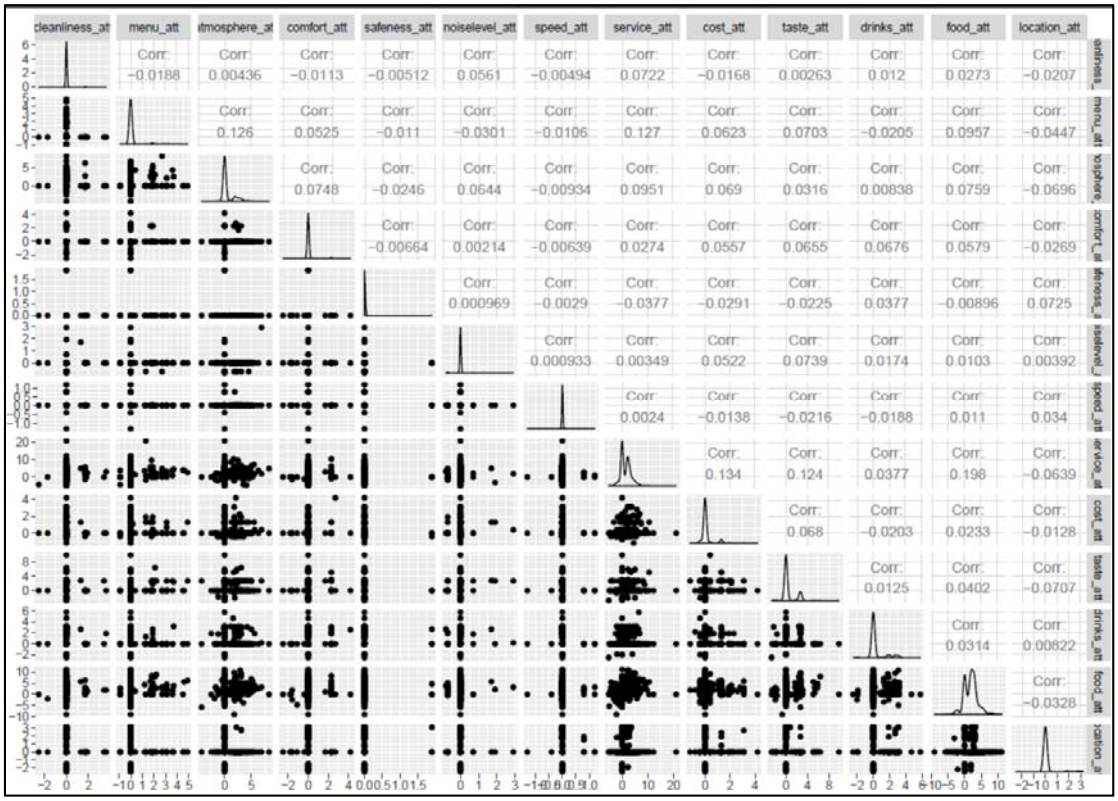

Figure.6. Pairwise Correlation Matrix.

b. 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 [70] 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 Table4. and Figure 7.

Table.4. Mean Decrease Importance Of Variables

feature	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

Journal of Theoretical and Applied Information Technology
30th September 2018. Vol.96. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

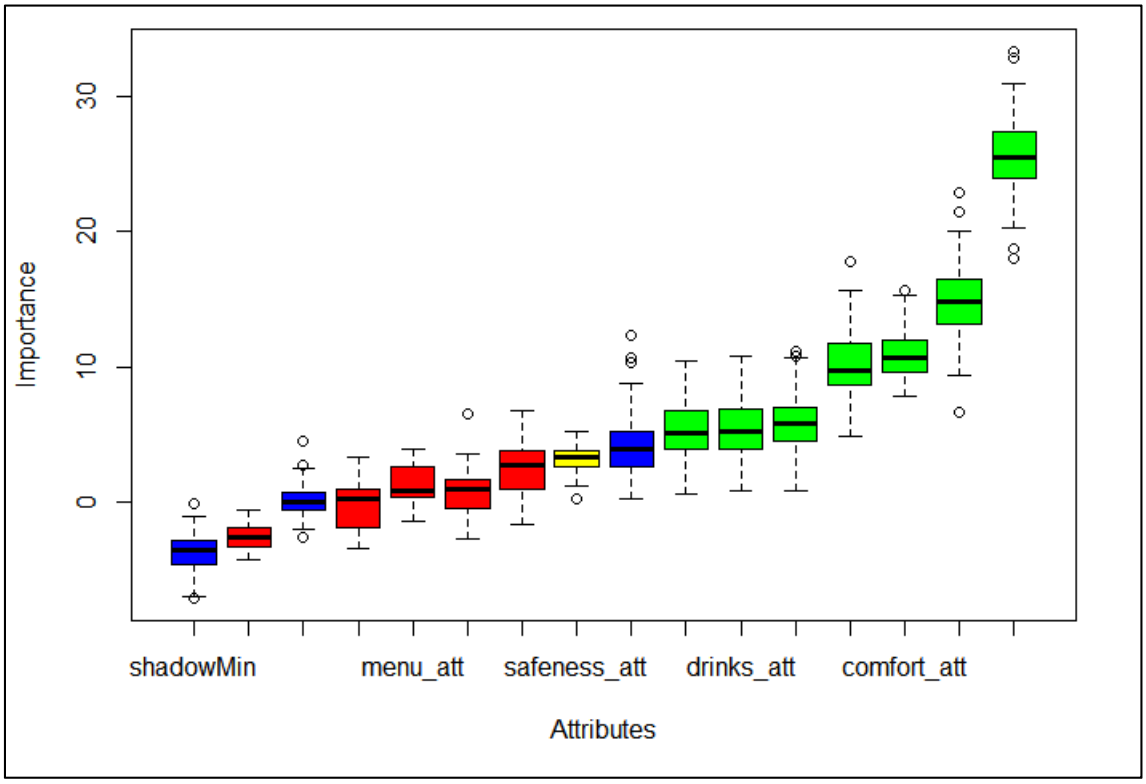

Figure.7. Mean Decrease Importance Of The Variables.

2- Performance measures

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

a) 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:

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

where:

TP: is the total number of reviews correctly classified to be of rate R

FP: is the total number of reviews incorrectly classified to be of rate R.

TN: is the total number of reviews correctly classified not to be of rate R.

FN: is the total number of reviews incorrectly classified not to be of rate R.

b) 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 classified instances, while recall is the fraction of reviews correctly classified over the total number of reviews in the rate R.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

c) F-measure

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

$$F_1 = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \quad (9)$$

3- MODEL TRAINING AND VALIDATION

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The selected model were first trained using the dataset which was split into 80% for training and validation and 20% for testing. For model training and validation, 5 x 2-fold cross validation was applied as recommended by [71]. 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.5.

Table.5. Parameters Values.

Model	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

4- 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.6.

Table.6. Performance Evaluation Of The Models Before And After Fuzzy

	Before Fuzzification				After Fuzzification			
Model	**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

Results presented in Table.6 show that data fuzzification enhances the predictive power of all the used classification models. Results show that 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 figure.8.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

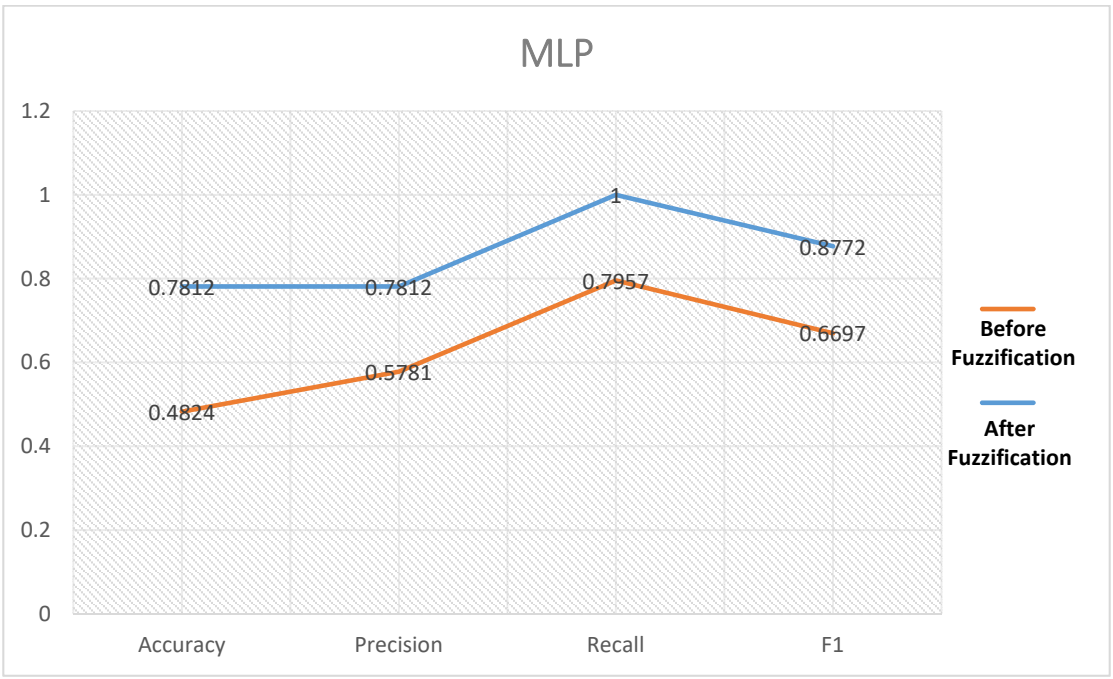

Figure.4 (A) Performance Of MLP

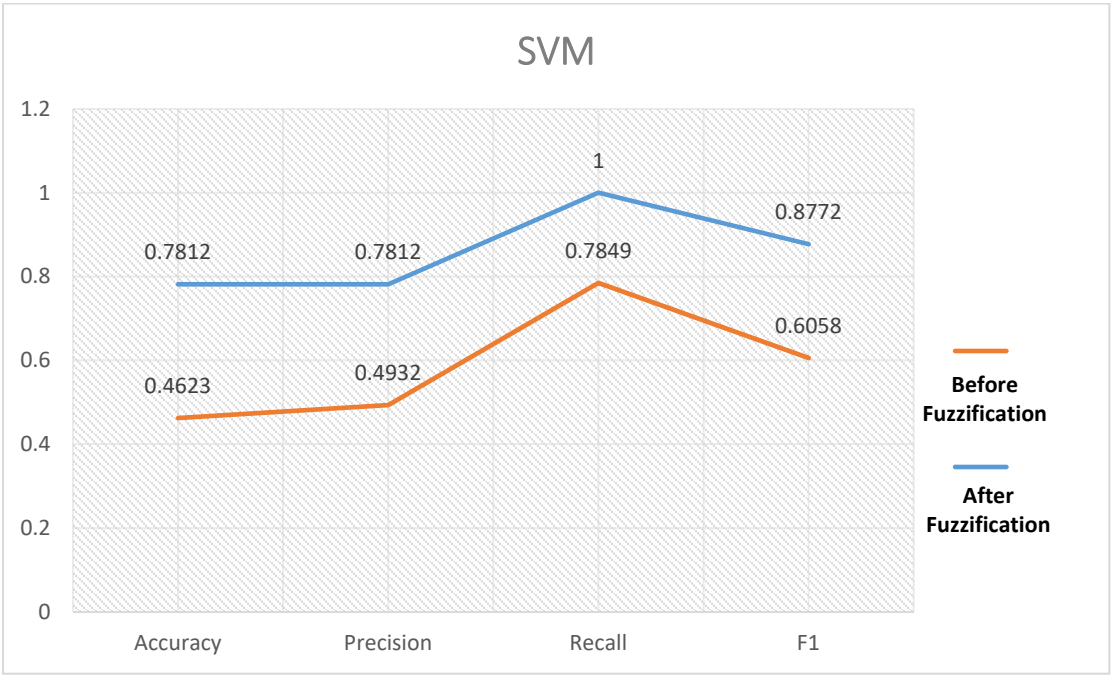

Figure.4(B) Performance Of SVM

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

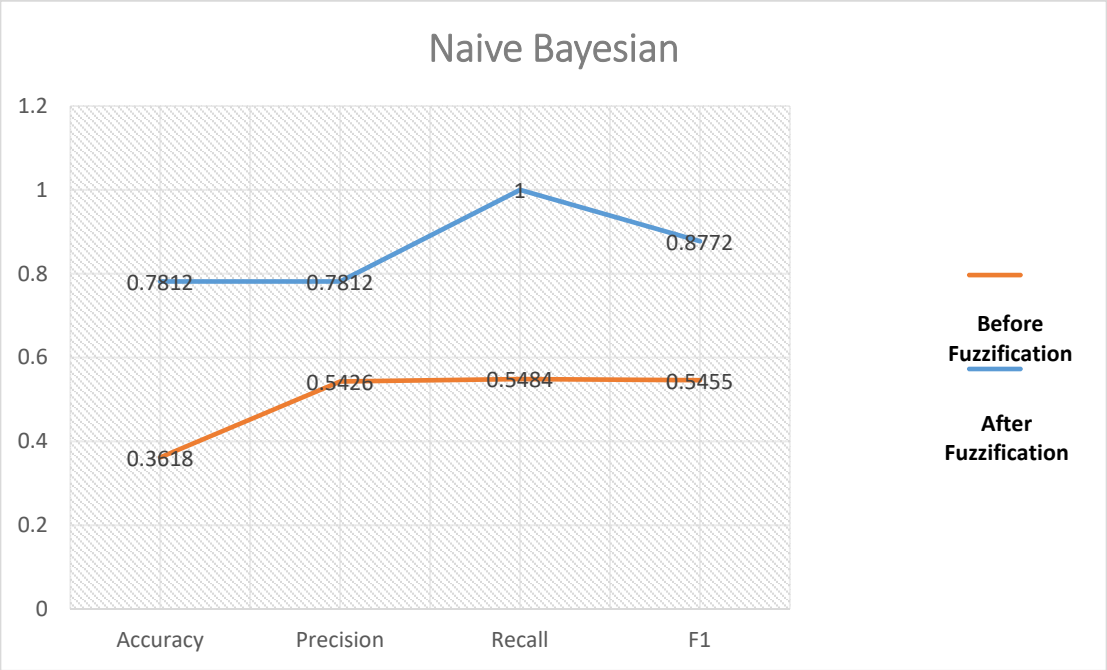

Figure.4 (C) Performance Of NV

6. CONCLUSIONS AND FUTURE WORKS

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

[1] Kirange, K. & Ratnadeep R. (2016). Aspect and emotion classification of restaurant and laptop reviews using SVM. International Journal of Current Research, 8, (03), 28352-28356.

[2] TC, C. & Joseph, S. (2014). Aspect based Opinion Mining from Restaurant Reviews.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

International Journal of Computer Applications (0975 – 8887) Advanced Computing and Communication Techniques for High Performance Applications (ICACCTHPA-2014).

[3] Perera, I. & Caldera, H. (2017). Aspect based opinion mining on restaurant reviews. 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA). DOI: 10.1109/CIAPP.2017.8167276.

[4] Ling, P., Geng, C., Menghou, Z. & Chunya, L. (2014). What Do Seller Manipulations of Online Product Reviews Mean to Consumers? (HKIBS Working Paper Series 070-1314) Hong Kong Institute of Business Studies, Lingnan University, Hong Kong.

[5] Nithin Y. & Poornalatha G. (2018). Feature Based Opinion Mining for Restaurant Reviews. In: Thampi S., Krishnan S., Corchado Rodriguez J., Das S., Wozniak M., Al-Jumeily D. (eds) Advances in Signal Processing and Intelligent Recognition Systems. SIRS 2017. Advances in Intelligent Systems and Computing, vol 678. Springer, Cham.

[6] Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2 (1-2).

[7] Basant, A., Namita, M., Pooja, B. & Sonal G. (2015). Sentiment Analysis Using Common-Sense and Context Information. Hindawi Publishing Corporation Computational Intelligence and Neuroscience.

[8] Tawunrat, C. & Jeremy, E. (2015). Chapter Information Science and Applications, Simple Approaches of Sentiment Analysis via Ensemble Learning, Volume 339 of the series Lecture Notes in Electrical Engineering, DISCIPLINES Computer Science, Engineering SUBDISCIPLINESAI, Information Systems and Applications-Computational Intelligence and Complexity.

[9] Matthew, J.K., Spencer, G. & Andrea, Z. (2015). Potential applications of sentiment analysis in educational research and practice – Is SITE the friendliest conference? In: Slykhuis, D., Marks, G. (Eds.), Proceedings of Society for Information Technology & Teacher Education International Conference 2015. Association for the Advancement of Computing in Education (AACE), Chesapeake, VA.

[10] Thomas, B. (2013). What Consumers Think About Brands on Social Media, and What Bunesses Need to do About it Report, Keep Social Honest.

[11] Ainur, Y., Yisong, Y. & Claire, C. (2010). Multi-level structured models for document-level sentiment classification. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. MIT, Massachusetts, Association for Computational Linguistics, USA, pp. 1046–1056.

[12] Noura, F., Elie, C., Rawad, A.A. & Hazem, H. (2010). Sentence-level and document-level sentiment mining for arabic texts. In: Proceeding IEEE International Conference on Data Mining Workshops.

[13] Nikos, E., Angeliki, L., Georgios, P. & Konstantinos, C. (2011). ELS: a word-level method for entity-level sentiment analysis. In: WIMS '11 Proceedings of the International Conference on Web Intelli- gence, Mining and Semantics.

[14] Haochen, Z. & Fei, S. (2015). Aspect-level sentiment analysis based on a generalized probabilistic topic and syntax model. In: Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference. Association for the Advancement of Artificial Intelligence.

[15] Khairullah, K., Baharum, B., Aumagzeb, K. & Ashraf, U. (2014). Mining opinion components from unstructured reviews: a review. J. King Saud Univ. Comput. Inform. Sci. 26 (3), 258–275.

[16] Sattar, A. Li, J., Ding, X., Liu, J. & Vincent, M. (2013). A general framework for privacy preserving data publishing. Knowledge-Based Systems 54 (2013) 276–287.

[17] S. Aldeen, Y., Salleh, M., & Razzaque, M. (2015). A comprehensive review on privacy preserving data mining. SpringerPlus, 2015, 4:694, DOI: 10.1186/s40064-015-1481-x.

[18] Xu Z. & Yi X. (2011). Classification of privacy-preserving distributed data mining protocols. In: Sixth international conference on digital information management, pp 337–342. http://doi.org/10.1109/ICDIM.2011.6093356.

[19] Ciriani V, Vimercati SDC, Foresti S. & Samarati P. (2008). k-anonymous data mining: a survey. In: Privacy-preserving data min- ing. Springer, New York, pp 105–136.

[20] EL Emam, K. & Dankar, F. (2008). Protecting Privacy Using k-Anonymity. Journal of the American Medical Informatics Association Volume 15 Number 5.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[21] Balaji R. (2013). The Complete Book of Data Anonymization. From: Planning to Implementation. Auerbach Publications, Boston, MA, USA.

[22] Junqiang L. & Ke W. (2010). Anonymizing transaction data by integrating suppression and generalization. 14th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part I, PAKDD'10, pages 171 {180, Berlin, Heidelberg, Springer-Verlag.

[23] Mahesh, R. & Meyyappan, T. (2013). Anonymization technique through record elimination to preserve privacy of published data. 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, pages 328-332. IEEE.

[24] Anbazhagan, K., SUGUMAR, D, Mahendran, M. & Natarajan, R. (2012). An Efficient Approach for Statistical Anonymization Techniques for Privacy Preserving Data Mining, International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 7.

[25] Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertain. Fuzziness Knowledge-Based Systems. 10(5): 571-588.

[26] Ding, X. Liu, B. & Yu, P. (2008). A Holistic Lexicon-Based Approach to Opinion Mining. 2008 International Conference on Web Search and Data Mining, Pages 231-240.

[27] Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In Data Engineering, ICDE 2007. IEEE 23rd International Conference on, pages 106-115.

[28] Zadeh, L.A. 1965. Fuzzy Sets. Information and Control. Vol. 8. pp. 338-353.

[29] Enache, I.C. (2015). Fuzzy Logic marketing models for sustainable development. Series V: Economic Sciences. Vol. 8. Iss. 57. No. 1-2015 Ferrara, E., Varol, O., Davis. C., Menczer, F., & Flammini, A. 2016. The rise of social bots. Communications of the ACM. Vol. 59. No. 7. pp. 96-104.

[30] Dragoni, M., Tettamanzi, A. & Pereira, C., (2015). Propagating and aggregating fuzzy polarities for concept-level sentiment analysis, Cogn. Comput. 7 (2) (April 2015) 186–197, DOI: 10.1007/s12559-014-9308-6.

[31] Dubois, D. & Prade, H. (2001). Possibility theory, probability theory and multiple-valued logics: a clarification, Ann. Math. Artif. Intell. 32(1–4). 35–66, https://doi.org/10.1023/A:1016740830286.

[32] Aida, V., Karina, G., David, S. & Montserrat, B., (2010), Using ontologies for structuring organizational knowledge in Home Care assistance, international journal of medical informatics 79 (2010) 370–387, Elsevier Ireland Ltd.

[33] Using, S., Ahmad, R., & Taib, S. (2010). Ontology of programming resources for semantic searching of programming related materials on the Web. ISBN: 978-1-4244-6716-7110.

[34] Christiane, F. (2010). WordNet. Theory and Applications of Ontology: Computer Applications, DOI 10.1007/978-90-481-8847-5_10, Springer Science+Business.

[35] Stephanie, C. & Narayanan, K. (2004). Semantic Feature Selection Using WordNet. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04).

[36] Velikovich, L., Blair-Goldensohn, S., Hannan, K., & McDonald, R. (2010). The viability of web-derived polarity lexicons. In The 2010 annual conference of the North American chapter of the ACL (pp. 777–785). Los Angeles, California.

[37] Hutto, C. & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. Eighth International Conference on Weblogs and Social Media (ICWSM-14).

[38] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric proper-ties of LIWC2007. Austin, TX: LIWC.net.

[39] López-Iñesta, E., Grimaldo, F. & Arevalillo-Herráez, M. (2017). Combining feature extraction and expansion to improve classification based similarity learning. Pattern Recognition Letters 93 (2017) 95–103.

[40] Setiawan, E., Widyantoro, D. & Surendro, K. (2016). Feature Expansion using Word Embedding for Tweet Topic Classification. 2016 10th International Conference on Telecommunication Systems Services and Applications (TSSA). IEEE.

[41] Lee, W. & Mit, E. (2011). Word Sense Disambiguation by Using Domain Knowledge. International Conference on Semantic Technology and Information Retrieval. 978-1-61284-353-7/11, IEEE.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[42] Fouad, K., Khalifa, A., Nagdy, N. & Harb, H. (2012). Web-based Semantic and Personalized Information Retrieval. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 3.

[43] Vyas, V. & Uma, V. (2018). An Extensive study of Sentiment Analysis tools and Binary Classification of tweets using Rapid Miner. Procedia Computer Science 125 (2018) 329–335.

[44] T. Yao & L. Li "A Kernel-based Sentiment Classification Approach for Chinese Sentences" World Congress on Computer Science and Information Engineering in March 31 2009-April 2.

[45] Nguyen, H., Xuan, A., Cuong L. & Nguyen, L. (2012). Linguistic Features for Subjectivity classification. International Conference on Asian Language Processing International Conference on Asian Language Processing in 2012.

[46] Su, X., Gao, G. & Tian, Y. (2010). A Framework to Answer Questions of Opinion Type. Seventh Web Information Systems and Applications Conference in 2010.

[47] Jusoh, S. & Alfawareh, H. (2013). Applying fuzzy sets for opinion mining. Computer Applications Technology (ICCAT), vol., no., pp.1,5, 20-22 Jan. 2013, doi: 10.1109/ICCAT.2013.6521965

[48] Dalal, M. & Mukesh, A. (2014). Opinion Mining from Online User Reviews Using Fuzzy Linguistic Hedges. Journal of Applied Computational Intelligence and Soft Computing. Vol 2014.

[49] Peñalver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Rodríguez-García, M., Moreno, V., Fraga, A. & Sánchez-Cervantes, J. (2014). Feature-based opinion mining through ontologies. Expert Systems with Applications, Volume 41, Issue 13, 1 October 2014, Pages 5995-6008.

[50] Weichselbraun, A., Gindl, S. & Scharl, A. (2014). Enriching semantic knowledge bases for opinion mining in big data applications. Knowledge-Based Systems, Volume 69, October 2014, Pages 78-85.

[51] Cambria, E., Song, Y. & Wang, H. (2014). Semantic Multidimensional Scaling for Open-Domain Sentiment Analysis. IEEE Intelligent Systems (Volume: 29, Issue: 2, Mar.-Apr. 2014).

[52] Kang, H., Yoo, S., Han, D. (2012). Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems with Applications 39 (2012) 6000–6010.

[53] Bandhakavi, A., Wiratunga, N., Padmanabhan, D. & Massie, S. (2017). Lexicon based feature extraction for emotion text classification. Pattern Recognition Letters 93 (2017) 133–142.

[54] Setiawan, E. Widyantoro, D. & Surendro, K. (2016). Feature Expansion using Word Embedding for Tweet Topic Classification. 2016 10th International Conference on Telecommunication Systems Services and Applications (TSSA). IEEE.

[55] Abinaya.R., Aishwaryaa.P. & Baavana. S. (2016). Automatic Sentiment Analysis of User Reviews. 2016 IEEE International Conference on Technological Innovations in ICT For Agriculture and Rural Development (TIAR 2016). IEEE.

[56] Khan, F., Qamar, U. & Bashir, S. (2016). eSAP: A decision support framework for enhanced sentiment analysis and polarity classification. Information Sciences 367–368 (2016) 862–873.

[57] Gitto, S. & Mancuso, P. (2017). Improving airport services using sentiment analysis of the websites. Tourism Management Perspectives 22 (2017) 132–136.

[58] Khan, F., Qamar, U., Bashir, S. (2016). S WIMS: Semi-supervise d subjective feature weighting and intelligent model selection for sentiment analysis. Knowledge-Based Systems 100 (2016) 97–111.

[59] Pandey, A., Rajpoot, D. & Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. Information Processing and Management 53 (2017) 764–779.

[60] Kogure, I., Shimoyama, T & Tsuda, H. (2016). De-identification and Encryption Technologies to Protect Personal Information. FUJITSU SCIENTIFIC & TECHNICAL JOURNAL. 28-36.

[61] Nguyen, D., Nguyen, D., Pham, D. & Pham, S. (2016). A Robust Transformation-Based Learning Approach Using Ripple Down Rules for Part-of-Speech Tagging", AI Communications, 29 (3). pp. 409-422.

[62] Saif, H., Fernandez, M., He, Y. & Alani, H. "On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter", LREC 2014: 810-817.

[63] Jean-Louis, L., Zouaq, A., Gagnon, M., Ensan, F. (2014). An Assessment of Online Semantic Annotatorsfor the Keyword

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Extraction Task. PRICAI 2014, LNAI 8862, pp. 548–560.

[64] Amer, E. and Fouad K. (2017). AKEA: An Arabic Keyphrase ExtractionAlgorithm, A.E. Hassanien et al. (eds.), Proceedings of the International Conferenceon Advanced Intelligent Systems and Informatics 2016, Advances in Intelligent Systems and Computing. PP:137-146. Springer International Publishing.

[65] Pu, m. (2006). Fundamental data Compression, ELSEVIER, 1st edition.

[66] Kumar, N., Srinathan,K. (2008). Automatic keyphrase extraction from scientific documents using N-gram filtration technique. pp. 199-208, proceeding of the eighth ACM symposium on Document engineering.

[67] Hsu, M., Tsai, M., & Chen, H. (2006). Query Expansion with ConceptNet and WordNet: An Intrinsic Comparison. AIRS 2006: Information Retrieval Technology pp 1-13.

[68] Schouten, K., Frasincar, F. & Jong, F. (2014). Ontology-Enhanced Aspect-Based Sentiment Analysis. International Conference on Web Engineering, ICWE 2017: Web Engineering pp 302-320.

[69] Bing, L. & Chan, K. (2014). A Fuzzy Logic Approach for Opinion Mining on Large Scale Twitter Data. 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing.

[70] Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Comput, 10(7):1895–1923.

[71] Kursa, M., Jankowski, A. & Rudnicki, W. (2010). Boruta – A System for Feature Selection", Fundamenta Informaticae, Volume 101 Issue 4, December 2010, Pages 271-285.