JOURNAL OF PURE AND APPLIED MICROBIOLOGY, Aug. 2015. Vol. **9**, p. 37-45
SPECIAL ISSUE ON RECENT RESEARCH CHALLENGES IN BIO-MEDICAL APPLICATIONS

Computer Aided Content Analysis of Web Based Medical Queries: Classifying Affective and Informative Posts

R. Satheesh Kumar[1] and S.Vijayan[2]

[1]Department of Information Technology , Hindusthan College of Engineering
& Technology, Coimbatore , Tamilnadu , India.
[2]Surya Engineering College, Perundurai, Erode, Tamilnadu, India.

(Received: 12 May 2015; accepted: 30 June 2015)

Various websites today provide medical information and this information can either be affective or informative. For example the blogs in Mayo clinic or Net doctor contains both information which are facts and information which are opinions from a fellow patient, doctor or nurse who try to analyze the given query and give an opinion. Similarly a site visitor can give his opinion without any knowledge on the subject. This paper proposes a semantic based feature extraction for automatically classifying affective and informative posts. The proposed semantic based feature selection uses SentiWordNet which is a lexical resource of WordNet database extracted terms and is available for research purposes. Results achievedshow that SentiWordNet is a good resource for sentiment opinion classification.

Key words: Opinion Mining, Semantic based feature selection, SentiWordNet and sentiment classification.

Sentiment analysis/opinion mining analyzes people's opinions, evaluations, sentiments, attitudes, appraisals and emotions to entities like products, organizations, services, issues, individuals, topics, events and their attributes. It is a large problem space and have been successfully applied in various domains including medical social media data[1]. Social media data not only consists of information which are facts but also consists of opinions from the person who visits the website to the website moderator. Generally doctors and nurses are also invited to give their opinion. Generally opinions are expressed on anything, e.g., a product, service, topic, individual, organization, or an event. A general term object denotes the entity commented on. An object has a components (or parts) set and an attributes set. Each component can have sub-components and attributes set etc. Hence, an object is hierarchically decomposed depending on part-of relationship[2].Beliefs, opinion, emotions and sentiments are private states parts which cannot be observed. These are expressed in documents through subjective words which identify private states using specific dictionaries like WordNet or SentiWordNet[3]. Opinion mining analyses customer's opinions using product reviews providing information including opinions polarity.

Feature extraction is important in opinion mining as customers do not usually express product opinions totally, but separately based on individual features. Two tasks should be accomplished in feature-based opinion mining. First, product features on which reviewers expressed opinions must be identified and extracted. Second, opinion orientation or polarities must be determined[4]. Finally, opinion mining summarizes extracted features and opinions.

* To whom all correspondence should be addressed.
E-mail: satheesh9.opm@gmail.com

Feature selection is used in tasks like image classification, data mining, cluster analysis, image retrieval, and pattern recognition. It is a data analysis preprocessing technique where a subset from original data features is selected to delete noisy features. This task allows reducing computational cost and improving accuracy of the data analysis process. A feature-selection approach example tailored to sentiment analysis using syntactic relations between text attributes are Feature Subsumption Hierarchies (FSH)[6].

Feature selection reduces original feature set through removal of irrelevant features for text sentiment classification for improved classification accuracy and decreased learning algorithms[7] running time. There are three feature selection group techniques i.e. filter, wrapper and embedded. In a filter category, a features group is chosen based on a specific mathematical equation which is used with any classifier. Contrary to that, features selected in wrapper and embedded techniques are bound to specific classifiers[8]. Other than being rigid regarding classifiers, wrapper and embedded techniques need high resource allocation and longer execution time.Filter methods choose features based on discriminating criteria like Fisher's discriminant criterion, Mutual information to name a few. Filter-based methods earlier evaluated features in isolation without considering inter feature correlation.Wrapper methods use classifiers as a black box to score features subsets based on predictive power[9].

Some of the feature extraction methods are listed in the table 1[10].

A relationship's semantic features are termed as relationship polarity and relationship strength, respectively. Relationship polarity is specifically classified into 4 types: positive, negative, neutral and no-relationship. Strength feature has three values: weak, medium and strong. The "no-relationship" polarity is evident in biomedical articles[11]. No-relationship differs from the neutral polarity: "no-relationship" indicates lack of an association between biological entities in consideration, while in neutral polarity, entities are associated without orientation.

Related work

An enhanced feature extraction/ refinement method titled FEROM which extracted correct features from review data by using grammatical properties and feature words semantic characteristics refining features by recognizing and merging similar ones was proposed by Jeong *et al*[12]. Experiments on actual online review data proved FEROM to be effective at extracting/ refining features to analyze customer review data leading to accurate opinion mining.

A feature selectors range regarding their efficiency in improving classifiers performance for sentiment analysis was evaluated by Isabella *et al* [13]. This study uses movie reviews for sentiment analysis. Computational performance improved through use of basic feature selection in most research. Sentiment analysis identified if opinion on a document topic was positive/negative. But many sentiment analysis applications were not possible due to voluminous feature amount in the corpus.

The main subtask of opinion summarization was proposed by Somprasertsri *et al*[14]. Opinion summarization addresses determination of sentiment, attitude or opinion expressed by an author in natural language regarding a specific feature. Product feature and opinion extraction tasks are important to opinion summarization as its effectiveness affects opinion orientation identification performance. It is important to identify semantic relationships between product features & opinions.

How syntactic dependency relations based features can be used to improve opinion mining performance was explored by Joshi *et al*[15]. Using a dependency relation triples transformation they are converted to "composite back-off features" which generalize better than regular lexicalized dependency relation features. Experiments comparing this approach with other methods generalizing dependency features or n-grams revealed composite back-off features utility.

Combining two approaches that intersection (AND) and union (OR) was undertaken by Thubaity *et al*[16]. The experiments revealed slight classification accuracy improvement in combining two/three feature selection methods. No classification accuracy improvement was possible when four/five feature selection methods combined. Yanjun *et al*[17] proposed combining many feature selection methods using Combinatorial Fusion Analysis (CFA). It revealed that rank-score function and rank-score graph can be used for selection of feature selection methods combination. Feature selection was evaluated by converging speed of boosting classifier based on single feature and high-dimensional and small-sized gene-expression data by Yun *et al*[18] & Feng *et al*[19]. Classification accuracy improved through the proposed methods. A new Genetic Algorithm (GA) based semi-feature selection method was proposed byHualong *et al*[20]. Experiments on public datasets showed that the suggested method outperformed traditional supervised and state-of-the-art filter based semi-feature selection algorithms.A Semantics Base Feature Vector using Part of Speech (POS) used to extract terms concept using WordNet and co-occurring and associated terms was proposed by Khan *et al*[21]. The suggested method was applied on small documents dataset proving that the suggested method outperformed TF-IDF.A new metric integrating correlation and reliability information between feature and class from Multiple Correspondence Analysis (MCA) to score features for feature selection was proposed by Zhu *et al*|[22]. Results showed that the new method outperformed statistical techniques in classification accuracy and feature subspace size.

A two-stage feature selection algorithm dependent on a feature selection method and latent semantic indexing was proposed by Meng *et al* [23]. Latent semantic indexing offsets problems due to using statistically derived concept indices and not individual words. It built a concept vector space where each term/document was represented as a vector in space. Due to too much calculation time to construct a new semantic space in this algorithm, a feature selection method was used to reduce term dimensions. Second, new reduced semantic space was built between latent semantic indexing method based terms., The proposed two-stage feature selection method performed better through applications involving spam database categorization.A new perspective for feature selection in face retrieval to increase face similarity feedback coherence was introduced by Dai *et al*[24] which narrowed semantic gap between human and machine.

A new method for semantic features selection and representation to be described by user through appropriate verbal descriptions using natural language concepts was proposed by Alattab *et al*[25]. One hundred respondents participated in selecting/ weighting human face semantic features based on each feature's level of importance. Features classes integration and classification was through Euclidian distance. The suggested human facial image retrieval was evaluated through many experiments using precision/recall methods. Results indicated high accuracies, compared to low-level features based facial image retrieval techniques.

A new feature selection process based on part-of-speech and HowNet was proposed by Liu *et al*[26]. According to text property composition, words chosen with larger information by various part-of-speech, expanded these words semantic features based on HowNet. Thus, short text had more useful features. The Test data set collected from sina micro-blog was used and it adopted micro average and macro average of F1-Measure to evaluate short text classification effects. Results showed that short text feature selected through the new method ensured a good amount of information and good classification results.

A comprehensive study on BoW representation choices including vocabulary size, stop word removal, weighting scheme, spatial information, feature selection and visual bi-gram was conducted by Jiang *et al*[27]. A soft-weighting method elaborated to assess significance of a visual word to an image was used for weighting scheme. Experiments showed that soft-weighting outperformed other weighting schemes like TF-IDF with a large margin. TRECVID data sets extensive experiments indicated that BoW feature with correct representation choices produced competitive concept detection performance.

Use of semantic frame-based analyzer FrameNet was proposed by Houen *et al*[28]. This was achieved by applying semantic information as a feature for a machine learning-based classifier and using semantic analysis in a bootstrapping process for the creation of a sentiment lexicon for classifier use. Many approaches to use FrameNet meta-data features were attempted, but pure FrameNet data achieved an F1-Score of 0.651, compared to a simple bag-of-tokens approach that resulted in an F1- Score of 0.819. From these findings it seems that using FrameNet as an opinion mining tool was not a rewarding research avenue.

A PageRank application, a random-walk

model formerly devised to rank Web search results to ranking WordNet synsets on how strongly they possess semantic property was presented by Esuli *et al*[29]. Semantic properties used to exemplify the approach included positivity and negativity, two properties of great importance in sentiment analysis. The idea was from observations that WordNet could be a graph where synsets were connected through binary relation "a term belonging to synset s_k occurs in gloss of synset si", based on the hypothesis that this relation could be a transmitter of semantic properties. The data for this is available with eXtended WordNet, a sense disambiguated WordNet version. Experiments supported proposed intuitions.

METHODOLOGY

In this work the polarity of the blog is selected as affective or informative. We follow the technique used by[1] for data collection. The obtained data is cleaned using stemming and stop words. Features are expanded and selected using the semantic expansion. Semantic expansion is achieved using SentiWordNet. Product of Term Frequency – Inverse Document Frequency is computed and the obtained feature vector is classified using Naïve Bayes and Ada Boost classifier. The detailed description is given in subsequent sections.

Data Collection from MAYO clinic& Netdoctor

Data were collected from Mayo Clinic a non profit medical organization. 1600 data consisting of equal number of affective and equal number of informative queries are collected from the website from various links including from sections of Question and Answer, reviews of drug, review on disease. An example of informative class used in this work and collected from Mayo clinic website is given

"Endometrial cancer is a type of cancer that begins in the uterus. The uterus is the hollow, pear-shaped pelvic organ in women where fetal development occurs. Endometrial cancer begins in the layer of cells that form the lining (endometrium) of the uterus. Endometrial cancer is sometimes called uterine cancer. Other types of cancer can form in the uterus, including uterine sarcoma, but they are much less common than endometrial cancer. Endometrial cancer is often detected at an early stage because it frequently produces abnormal vaginal bleeding, which prompts women to see their doctors. If endometrial cancer is discovered early, removing the uterus surgically often cures endometrial cancer"

A affective class document obtained from Netdoctor is given

"hello it has been going on 2 yrs, i cant sleep and i get chronic headache when i do sleep. i get up every two hours or sometimes even shorter

Table 1. Some of the Feature Selection Methods

Method	Usage
Document Freq. (DF)	The number of documents in which a particular term occurs. For example the word "doctor" may occur in 25 documents
Inverse Document Freq. (IDF)	Logarithmic technique to find rare term in documents
Inverse Collection Freq. (ICF)	Computed from terms which high collection frequency
Information Gain (IG)	The importance of a word or vector computed statistically
Word Frequency (WF)	A weighted number of occurrences of words with respect class.

Table 2. Summary of Results

Techniques	Classification Accuracy	RMSE	F Measure
TF-IDF feature extraction : Naïve Bayes Classifier	88.5	0.3124	0.886
TF-IDF feature extraction : Ada Boost	73.25	0.4528	0.7326
TF-IDF + Semantic Feature extraction : Naïve Bayes Classifier	92.5	0.2652	0.9257
TF-IDF + Semantic Feature extraction : Ada Boost	83.25	0.3614	0.8327

time. i get a migrane headache if i sleep wrong and i have to stay up for at least an hour . when i stay up the headache sometimes go away.i am just so tired, that i want to take a pill that would let me sleep for a week. i have tried alot of sleeping pills and the most hours i have gotten was 4 hours of straight sleep with a total migrane headche .the headache has been comming annoying it comes more at night and i always have a slight during the day,as long as my head is not laying down or tilted., i do suffer from nightmares.there is times when in my dream i see my room as it really is and things are happenning to me and my eyes r open.it feels so real, untill my husband wakes me up because i start yelling for him to help me in my dream or i have to pray calling jesus to help me, by using his name to break the chain,of the evil in my dream .i am so tired all i want is my mind and body to feel rested.I have tried countless medicine, perscribe and over the counter without any sucess/i need guidance, running thoughts in my head just keeps running"

Word Stemming

Word stemming is a crude pseudo-linguistic process to delete suffices to reduce words to word stem. For example, words ‘classifier’, ‘classified’ and ‘classifying’ would be reduced to word stem ‘classify’. Subsequently, feature space dimensionality is reduced by mapping morphologically similar words onto word stems[31]. A common stemming algorithm is the Porter developed suffix stripper. Arabic language has two different morphological analysis techniques; stemming and light-stemming.

Stop Words

Stop words refer a set of terms/words with no inherent useful information. Stop words are problematic in key concepts and words identification from textual sources when they are not removed by their presence regarding

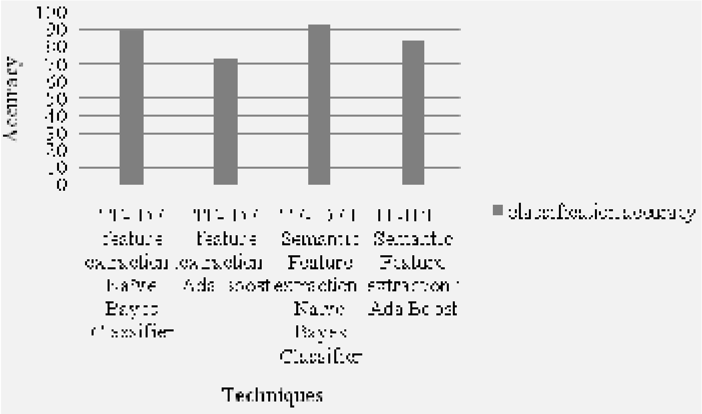

Fig. 1. Classification Accuracy

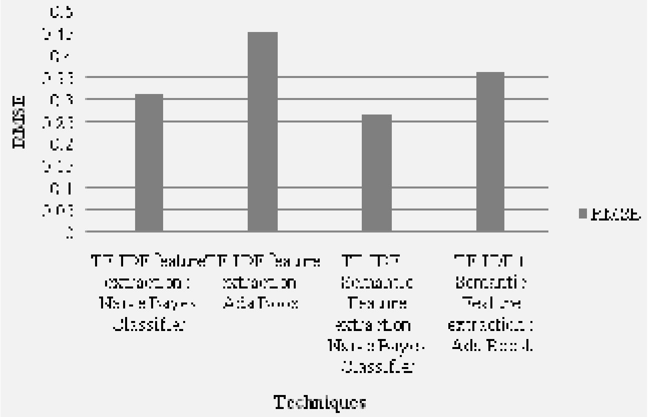

Fig. 2. RMSE

frequency and occurrence in textual sources[32]. There is no definite stop words list used by all tools. Some tools avoid their removal to support phrase search.

A words group can be chosen as stop words for a purpose. These are some of the most common, short function words - for some search machines – like is, at, which, and on. Here stop words cause problems when locating phrases which include them, specifically in names like 'The Who', 'The', or 'Take That'. Other engines remove some common words—including lexical words like "want"—from a query to improve performance.

The output obtained after stemming and stop words for the two examples are

"Endometr cancer type cancerbegins uterus hollow pearshaped pelvic organ women etal development occurscells form lining sometimes called uterine cancer including uterine sarcoma common endomet cancer often detected stage because frequent produces abnormal vagin bleeding prompts women doctors endomet cancer discovered early removing uterus surgic often cures endomet cancer"

"hello going 2 yrs cant sleep chronic headac when sleep every two hours sometim even shorter time migrane headac sleep wrong stay least hour when stay headac sometim away just tiredpill that would let sleep weektried sleeping pills most hours gotten 4 hours straight sleep total migrane headc headac been comming annoying comes more night always slight during long head lay down tilted suffer nightmare times when room really things happenn eyes open feels so real, untill my husband wakes because start yelling help dream pray calling jesus help using evil dream tired a mind body feel rested have tried countless medicine,perscribe over counter without anysucessneed guidance running thoughts head just keeps running"

Inverse document frequency (IDF)

IDF is a numerical statistic reflecting the importance of a word to a document in a given set of documents. IDF value increases proportionally to the many times a word appears in a document, but is offset by word frequency in the corpus helping to control the fact that some words are more common than others.

It is usual to replace simple word frequency with weighted frequency prior to computing cosine and other statistics. Weighted frequency statistic is TF-IDF (term frequency – inverse document frequency) statistic which computes a weight for every term reflecting its importance. TF-IDF is a common text categorization tasks[33] metric, but its use in sentiment analysis is limited and has not been used as a unigram feature weight. TF-IDF comprises of term frequency and inverse document frequency. The former is found by counting the times a term occurs in a document, and IDF is got through dividing total documents by documents a given word appears in. When these values are multiplied it results in a score that is highest for words appearing frequently in few documents, and low for terms appearing regularly in all documents, permitting location of terms important in a document. TF- IDF is computed by[34]:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \qquad \text{...(1)}$$

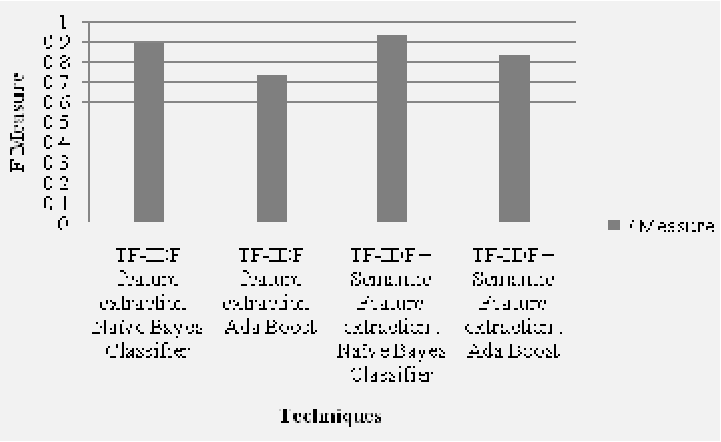

Fig. 3. F Measure

Where, $w_{i,j}$ is weight for term i in document j, N the number of documents in a collection, $tf_{i,j}$ is term frequency of term i in document j and df_i is document frequency of term i in a collection.

Naive Bayes Classification

Naive Bayes classifier is based on Bayes' theorem and assigns class $c^* = \arg\max_c P(c \mid d)$ to document d. Its underlying probability model is described as "independent feature model".

"Naive" attribute is due to the fact that model assumes all features are independent, when in real problems they never are. Despite serious basic assumptions violations and the classifier simplistic design it turns out they suit problems involving normal distributions, common in real-world problems[36]. The bayes classification is a supervised learning technique and classification's statistical technique. This assumes an underlying probabilistic model allowing the capture of the model's uncertainty in a principled way through determining outcomes probabilities

The Naive Bayes (NB) classifier uses Bayes' rule[35] as in equation (2):

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)} \qquad ...(2)$$

Where, P(d) plays no role in c* selection. To estimate term $P(d \mid c)$, Naive Bayes decomposes it by assuming f_i are conditionally independent given d's class as in equation (3):

$$P_{NB}(c \mid d) = \frac{P(c)(\prod_{i=1}^{m} P(f_i \mid c)^{n_i(d)})}{P(d)} \qquad ...(3)$$

Where f_i the feature vector.

Ada Boost Classifier

Boosting is a method to improve accuracy of any learning algorithm through combining weak or base learners[38]. A known boosting algorithm is AdaBoost which takes a training instances set $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ as input, where each x_i belongs to instance space X, and label y_i is in some label space Y.

AdaBoost algorithm for binary classification [39] is a known boosting algorithm whose base procedure is a classifier with values in {0, 1}, for example, a classification tree. AdaBoost ensures an efficient mechanism for feature selection [40]. If a weak classifiers set is just a binary features (called boosting stumps) set each boosting round adds a single feature to current features set. AdaBoost is an iterative process where each round chooses a weak classifier, $h_t()$ that minimizes as in equation (4):

$$Z_t = \sum_i D_t(i) \exp(-y_i h_t(x_i)) \qquad ...(4)$$

The Ada Boost classifier algorithm steps are,

•Select subset of training sets from dataset $D_i(i) = 1/m$

•For

1. Train selected weak learners with
2. Select weighting criteria for each training algorithm $\alpha_t \in R$
3. Update the distribution by (5):

$$D_{t+1}(i) = \frac{D_t(i)\, e^{-\alpha_t y_i h_t(x_i)}}{Z_t} \qquad ...(5)$$

Where Z_t is a normalization.

D_t $t = 1,....,T$ •For final outcome, vote H(x) is a weighted sumgiven by equation (6):

$$H(x) = sign(f(x)) = sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \qquad ...(6)$$

RESULTS AND DISCUSSION

The experiments conducted using TF-IDF feature extraction: Naïve Bayes Classifier, Ada Boost and TF-IDF + Semantic Feature extraction: Naïve Bayes Classifier, Ada Boost. The table 2 is the summary of results.

The proposed semantic feature extraction with TF-IDF and Naïve Bayes improves the classification accuracy by 4.52% when compared with the feature extraction using TF-IDF and by 13.65% for Ada Boost classifier

The proposed semantic feature extraction with TF-IDF and Naïve Bayes decreases the RMSE by a percentage of 15.11 compared to feature extraction with TF-IDF and by 20.19% for Ada Boost classifier.

The proposed semantic feature extraction with TF-IDF and Naïve Bayes improves the F

Measure in the percentage of 4.48 for the feature extraction with TF-IDF and by 13.66% for Ada Boost classifier.

CONCLUSION

This paper proposed extracting words and select words based on importance using TF-IDF with semantic based feature selection. Feature set is reduced through use of proposed semantic based procedure to consider individual predictive ability words and select features. Experiments were undertaken with Naïve Bayes and Ada Boost classifiers and results were compared to judge the feature selection methods. The results from experiments prove that the proposed semantic feature based selection improves the efficiency of the classifiers.

REFERENCES

1. Denecke, K., & Nejdl, W. How valuable is medical social media data? Content analysis of the medical web. *Information Sciences,* 2009; **179**(12), 1870-1880.
2. Liu, B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies,* 2012; **5**(1), 1-167.
3. Samsudin, N., Puteh, M., Hamdan, A. R., & Nazri, M. Z. A. Mining Opinion in Online Messages. *International Journal*, 2013.
4. Jeong, H., Shin, D., & Choi, J. Ferom: Feature extraction and refinement for opinion mining. *ETRI Journal,* 2011; **33**(5).
5. Schiezaro, M., & Pedrini, H. Data feature selection based on Artificial Bee Colony algorithm. *EURASIP Journal on Image and Video Processing,* **2013; 1**: 1-8.
6. Abbasi, A. Intelligent Feature Selection for Opinion Classification. *IEEE Intelligent Systems,* 2010.
7. Sharma, A., & Dey, S. Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis. *IJCA Special Issue on Advanced Computing and Communication Technologies for HPC Applications*, 2012; **3**.
8. Samsudin, N., Puteh, M., Hamdan, A. R., & Nazri, M. Z. A. Immune Based Feature Selection for Opinion Mining. *In Proceedings of the World Congress on Engineering,* 2013; **3**.
9. Ladha, L., & Deepa, T. Feature selection methods and algorithms. *International Journal on Computer Science and Engineering,* 2011; **3**(5), 1787-1797.
10. Neumayer, R., Mayer, R., & Nørvåg, K. Combination of feature selection methods for text categorisation. In Advances in Information Retrieval (pp. 763-766). Springer Berlin Heidelberg 2011.
11. Swaminathan, R., Sharma, A., & Yang, H. (2010, July). Opinion mining for biomedical text data: Feature space design and feature selection. In Proceedings of the 9th International Workshop on Data Mining in Bioinformatics.
12. Jeong, H., Shin, D., & Choi, J. Ferom: Feature extraction and refinement for opinion mining. *ETRI Journal,* 2011; **33**(5).
13. Isabella, J., & Suresh, R. Analysis and evaluation of Feature selectors in opinion mining. *Indian Journal of Computer Science and Engineering,* **3**.
14. Somprasertsri, G., & Lalitrojwong, P. Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization. *J. UCS,* 2010; **16**(6), 938-955.
15. Joshi, M., & Penstein-Rosé, C. (2009, August). Generalizing dependency features for opinion mining. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (pp. 313-316). Association for Computational Linguistics.
16. Al-Thubaity, A., Abanumay, N., Al-Jerayyed, S., Alrukban, A., & Mannaa, Z. (2013, July). The Effect of Combining Different Feature Selection Methods on Arabic Text Classification. In Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2013 14th ACIS International Conference on (pp. 211-216). IEEE.
17. Li, Y., Hsu, D. F., & Chung, S. M. (2009, November). Combining multiple feature selection methods for text categorization by using rank-score characteristics. In Tools with Artificial Intelligence, 2009. ICTAI'09. 21st International Conference on (pp. 508-517). IEEE.
18. Sha, Y., Yu, X., & Zhang, G. A feature selection algorithm based on boosting for road detection. In Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on 2008; **2**: pp. 257-261
19. Yang, F., & Mao, K. Z. Robust feature selection for microarray data based on multicriterion fusion. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB),* 2011; **8**(4): 1080-1092.
20. Bu, H., Zheng, S., & Xia, J. (2009, August). Genetic algorithm based Semi-feature selection method. In Bioinformatics, Systems Biology and

Intelligent Computing, 2009. IJCBS'09. International Joint Conference on (pp. 521-524). IEEE.

21. Khan, A., Baharudin, B., & Khan, K. (2010, June). Semantic based features selection and weighting method for text classification. In Information Technology (ITSim), 2010 International Symposium in (Vol. 2, pp. 850-855). IEEE.

22. Zhu, Q., Lin, L., Shyu, M. L., & Chen, S. C. (2010, September). Feature selection using correlation and reliability based scoring metric for video semantic detection. In Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on (pp. 462-469). IEEE.

23. Meng, J., Lin, H., & Yu, Y. A two-stage feature selection method for text categorization. *Computers & Mathematics with Applications*, 2011; **62**(7): 2793-2800.

24. Dai, W., Fang, Y., & Hu, B. Feature selection in interactive face retrieval. In Image and Signal Processing (CISP), 2011 4th International Congress on, 2011; **3**: 1358-1362

25. Alattab, A. A., & Kareem, S. A. (2013, January). Semantic Features Selection and Representation for Facial Image Retrieval System. In Intelligent Systems Modelling & Simulation (ISMS), 2013 4th International Conference on (pp. 299-304). IEEE.

26. Liu, Z., Yu, W., Chen, W., Wang, S., & Wu, F. (2010, December). Short text feature selection for micro-blog mining. In Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on (pp. 1-4). IEEE.

27. Jiang, Y. G., Yang, J., Ngo, C. W., & Hauptmann, A. G. Representations of keypoint-based semantic concept detection: A comprehensive study. Multimedia, *IEEE Transactions on*, 2010; **12**(1): 42-53.

28. S Houen. Opinion Mining with Semantic Analysis 2011.

29. Esuli, A., Sebastiani, F.: Page Ranking WordNet synsets: An application to Opinion Mining. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, vol. 45, pp. 424–431. Association for Computational Linguistics (2007)

30. Denecke and Nejdl "How valuable is medical social media data? Content analysis of medical web", 2009.

31. Mountassir, A., Benbrahim, H., & Berrada, I. (2013). A Novel Model For Text Document Representation: Application On Opinion Mining Datasets. International Journal of Computer Science.

32. Choy, M. (2012). Effective Listings of Function Stop words for Twitter. arXiv preprint arXiv:1205.6396.

33. O'Keefe, T., & Koprinska, I. (2009). Feature selection and weighting methods in sentiment analysis. ADCS 2009, 67.

34. Zhang, W., Yoshida, T., & Tang, X. A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, 2011; **38**(3): 2758-2765.

35. Sindhu C., Dr. S. ChandraKala., A Survey On Opinion Mining And Sentiment Polarity Classification. *International Journal of Emerging Technology and Advanced Engineering*, 2013; **3**(1).

36. Mahendran, A., Duraiswamy, A., Reddy, A., & Gonsalves, C. Opinion Mining For Text Classification.

37. Sindhu C., Dr. S. ChandraKala., Opinion mining and sentiment classification: a survey. *ICTACT Journal On Soft Computing*, 2012; **3**(1).

38. Xiao, M., & Guo, Y. Multi-View AdaBoost for Multilingual Subjectivity Analysis. *In COLING*, 2012: pp. 2851-2866.

39. Bühlmann, P., & Hothorn, T. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 2007; 477-505.

40. Viola, P., & Jones, M. Fast and robust classification using asymmetric adaboost and a detector cascade. *Advances in Neural Information Processing System*, 2001; **14**.