Content-Specific Unigrams and Syntactic Phrases to Enhance Senti Word Net Based Sentiment Classification

Muhammad Latif, Usman Qamar, and Abdul Wahab Muzaffar

***Abstract*—Sentiment classification intelligently detects the polarity of documents by ascertaining polar values encapsulated in the document to classify them into positive and negative sentiments. Machine learning classifier completely relies on the feature set orientations. SentiWordNet is a lexical resource where each term is associated with numerical scores for subjective and objective sentiment information. SentiWordNet based sentiment classifier uses sentiment features generated from 7% subjective terms available in the resource. Sentiment features bear generic orientation for multiple domains but lacks comprehensive coverage e.g. Text unit with null or few sentiment features reflects ambiguous or null sentiments. Use of content specific unigrams and syntactic phrases along with sentiment features ensures consistency in the classification while enhancing the performance paradigm. Model proposed in this research is validated on sentiment and polarity datasets. Results of this research, completely out performs previous approaches and methods.**

***Index Terms*—Content specific features, lexicon based classification, sentiment classification, Senti word net.**

I. Introduction

Business communities are keen to utilize sentiment analysis for the purpose of business intelligence and identifying consumer behavior. Huge data is available over the internet in the form of reviews, blogs discussions, emails and tweets. This data creates an opportunity to improve corporate decision making.

The goal of sentiment analysis is to detect subjective information from text and has been largely divided into two categories; sentiment classification and semantic orientation. In semantic-orientation, the polarity of a given text is known through sentiment bearing lexicons either with the use of rule based or un-supervised approach. In this technique it may use a corpus to identify sentiment bearing words and phrases or an available dictionary or a lexicon resource.

On the contrary, sentiment classification classifies the text into positive, negative or neutral classes on the basis of various independent and combinatory features as articulated in the literature. Different machine learning algorithms e.g. Naïve Bayes, Support Vector Machine and Max Entropy were also employed in standalone and combinatory perspectives.

English language lexicon resource SentiWordNet is designed specifically to assist sentiment analysis tasks. Previously, it was mostly used to determine the semantic orientation of sentence and subsequently the whole document. However, in sentiment classification, sentiment features were derived from the 7% subjective terms available in SentiWordNet.

It was identified that text usually contained few or null sentiment features and ambiguous or null polarity characterized by the classifier e.g. a negative review taken from kitchen domain of multi domain dataset[1]:

"I've had my share of dutch ovens in my time, and I have to say that this is the foulest one yet. I thought I smelled a good deal when I got it, but boy was I mistaken"

In the above example only one subjective term "good" was identified as a sentiment feature which was insufficient to determine the negative polar orientation of review. Ultimately, inadequate performance observed when only sentiment feature was considered.

Sentiment features as new compositional dimension with various features were experimented in the literature to enhance the performance of sentiment classification process. Analysis revealed that even the use of this compositional approach towards features did not improve the performance remarkably. Therefore better representation of features was required to reinforce the SentiWordNet based sentiment classification.

A machine learning based framework is proposed with the following objectives:

1) Content specific new feature dimension based on syntactic constituency relation.
2) Find the best representative feature combination from content specific and sentiment features.
3) Reduce dimensionality and computation by features selection to further increase the performance.

Organization of the paper is, Section II describes the related studies, Section III explains the proposed framework, Section IV covers the experimental setup, Section V shows the results with discussion and in Section VI and Section VII induces the conclusion and future work.

II. Related Work

Different studies have been considered to investigate how various independent and combinatory features was articulated in the literature. Some of them used the SentiWordNet and some were experimented on the same datasets.

The machine-learning based sentiment-classification was experimented in [1]. Different sets of features comprised of adjectives, unigrams, bigrams, POS and position information were processed and evaluated with feature frequency and mostly with feature presence. They performed comparison of

Manuscript received October 15, 2014; revised February 25, 2015.

The authors are with the National University of Sciences and Technology, Islamabad, Pakistan (e-mail: latif_awang001@yahoo.com, usmanq@ceme.nust.edu.pk, wahab.muz@ceme.nust.edu.pk).

[1] http://www.cs.jhu.edu/~mdredze/datasets/ sentiment/index2.html

Max Entropy, Naïve Bayes and SVM to classify movie reviews.

Semantic orientation based unsupervised approach was studied by [2]. Learning algorithm started with extraction of phrasal lexicon from reviews using POS based 5-patterns. Followed by the polarity of phrases learned using point wise mutual information-PMI i.e. co-occurrence with "excellent" and "poor". Finally, a review was rated by averaging the polarities of phrases. In these experiments, they used phrases instead of words.

Study on ensemble of feature and classifier was conducted by [3]. Three types of ensemble techniques were used for both features and classifiers. They considered two types of features, POS based (unigrams of adjectives, nouns, adverbs & verbs) and Word Relation (unigrams, bigram and dependency Tree). Naive Bayes, maximum entropy, and SVM were used for ensemble of classifiers.

However, SVM was used as base-line classifier with linear kernel for ensembles of features on product reviews and movies reviews.

A hybrid approach of both semantic orientation and machine learning classification was used in [4]. They proposed new feature dimension as sentiment features extracted from the subjective terms available in SentiWordNet. In their experiments, three types of features were considered; content free (lexical, syntactical and structural)-F1, content specific (unigrams and bigrams)-F2 and sentiment -F3. Different combination of features was experimented with information gain feature selection. Performance results proved that sentiment features is a better addition in feature set for sentiment classification.

To overcome the SentiWordNet limitation of only 7% subjective words, objective words (after revised score) were used in [5]. The scores of objective words were reassigned using the relevance of objective word with the semantic orientation of sentences. SVM classifier was used on movie review dataset with both original SentiWordNet and revised SentiWordNet.

TABLE I: SUMMARY ON THE DIFFERENT NATURE OF FEATURES USED FOR SENTIMENT CLASSIFICATION

Study	Features
Pang *et al*. 2002 [1]	Unigrams, Bigrams, POS and Position
Turney 2002 [2]	Phrases based on POS patterns
Xia *et al*. 2011 [3]	POS(Adjectives, Adverbs, Nouns and Verbs) and Word relations (Unigrams, Bigrams and Dependencies)
Dang *et al*.2010 [4]	SentiWordNet subjective sentiment features , Content-specific unigrams & bigrams and Content-free features
Hung *et al*. 2013 [5]	SentiWordNet subjective terms and Objective (after revised score)terms
Agarwal *et al*. 2013 [6]	Unigrams, Bigrams and Bi-tagged phrases (based on POS patterns)

Bi-tagged phrases based on nine POS patterns were introduced in [6]. Bi-tagged phrase features were articulated as a new dimension towards sentiment classification instead of only reliance on bigrams and combination with unigrams. This new dimension of feature alongside pre existential features was experimented through supervised classification algorithms. Information gain was used to reduce the dimensionality and noise. Experiments were conducted on movies reviews and results were presented in F-measure.

Different features considered in the literature are summarized in Table I.

III. PROPOSED FRAME WORK

SentiWordNet is a tool for sentiment classification, used to devise a set of sentiment features from textual documents. It is noteworthy that text usually contained few or null sentiment features, also reveal that sole reliance on sentiment features result in limited performance for classification. Eventually there was a need to enhance feature sets required for sentiment classification.

Phrase Structure Grammar / Syntactic Phrases: It is the sentence structure of text in which a sentence is viewed in terms of the constituency relation [7] as shown in Fig. 1. Sentence encapsulates phrases that are the combination of words those act as a single POS in a sentence.

The noun and verb phrases are considered as the important constituent phrases in each sentence. Noun phrases are the important key words as object and subject while carrying the most important information. Verb is the skeleton of any sentence but mostly verbs are objective in meaning. However when verbs are combined with their dependents they form verb phrases that are more meaningful.

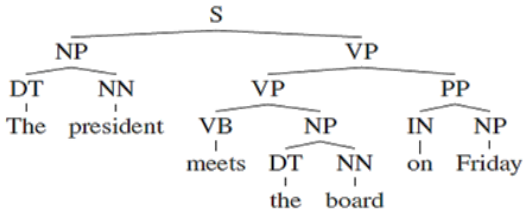

Fig. 1. Phrase structure (Example from: [8]).

Unigram appeared in text classification were used to represent the actual content in BOW model [1]. Unigrams also mostly utilized in sentiment analysis features due to its benchmarked performance in literature all across.

Currently most frequent verb phrases, noun phrases along the unigrams is proposed as features to better represent the content of text. The limitation of SentiWordNet could be reduced by considering the proposed features.

The previously quoted example review articulated in Table II in the light of proposed method. Features specially "the foulest one" and "mistaken" are the negative feature used to correctly classify it as negative.

TABLE II: ADDITION OF NEW FEATURE TYPE AND FEATURE WORDS

Feature Type	Features Words
Sentiment Feature	"good"
Unigrams	"get", "think", "say", "time", "boy", "smell", "oven", "mistaken", "deal", "dutch", "share"
Syntactic Phrases (NPs & VPs)	"get", "think", "say", "time", "boy", "smell", "oven", "have to say", "my time", "a good deal", "dutch ovens", "the foulest one"

Once a feature set is generated it can be used as starting point to train supervised learning methods of sentiment classification. The proposed system mainly consists of following three phases as shown in Fig. 2.

1) Text Pre-processor
2) Feature Extractor
3) Sentiment Classifier

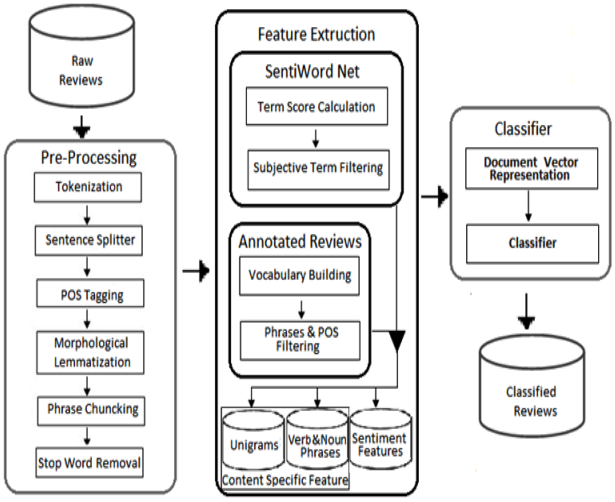

Fig. 2. Proposed model.

A. Data Sets of Reviews

1) The multi domain sentiment data set

The multi-domain sentiment dataset of product reviews (book, DVD, electronic, and kitchen appliances) taken from amazon was first used by [9]. The data set contains 1000 negative and 1000 positive labeled reviews for each domain. Each review have the information consisting of a rating (0-5 stars), a reviewer name and location, a product name, a review title and date, and the review text. Reviews with ratings > 3 were labeled positive; those with rating < 3 were labeled negative.

2) The polarity data set[2]

The polarity dataset v2.0 is a set of film review documents available for research in sentiment analysis and opinion mining. It was first introduced as a research data set for sentiment classification presented in [1] with 700 positive and 700 negative reviews. Revised dataset presented in 2004 and comprises of 1000 positive and 1000 negative labeled film reviews extracted from the Internet Movie Database Archive.

B. The Text Preprocessing

Purpose of this phase is the transform the text that can be used for further text engineering activities. General Architecture for Text Engineering-GATE [10] is an open source tool, widely used by many research communities for text preprocessing. A Nearly-New Information Extraction System (ANNIE) is the information extraction application available in GATE. Pre-processing activities presented in this research are shown in Fig. 2.

C. The Feature Extraction Phase

1) The Senti word net database

SentiWordNet [11] is a database containing opinion scores for terms derived from the WordNet database version 2.0. Each set of terms sharing the same meaning, or *synsets*, is associated with three numerical scores ranging from 0 to 1, each indicating the synsets objectiveness, positive and negative bias. Such as:

$$PosScore + NegScore + ObjScore = 1$$

Data in SentiWordNet is categorized according to parts of speech because there are considerable differences in the level of subjectivity depending on its part of speech. After POS tagging each term has been associated with a POS tag to accurately apply the SentiWordNet scores.

Term Score Calculation: Every entry in the SentiWordNet takes the form term#sense. Obviously, different word senses can have different polarities in a single POS as shown in Table III:

TABLE III: SENTI WORD NET SCORE AGAINST SENSE

POS	ID	PosScore	NegScore	term#sense
R	00011093	0.375	0	well#1
R	00012531	0.5	0	well#3
R	00013092	0.75	0	well#6
R	00013626	0.125	0.25	well#12
R	00012129	0.667	0.333	well#13

An issue was raised in SentiWordNet while assigning the score to a term; different calculation formulae were discussed by [12].Average (AVE) and first sense (FS) were frequently used in previous studies and AVE is used for in this research experiments.

Subjective Terms Segregation: Usually, subjective terms are more meaningful for sentiment classification tasks. Therefore they were used as features. A term is said to be subjective, if sum of its PosScore and NegScore is more than its objective score else it is objective. No of subjective and objective terms in SentiWordNet estimated with both FS and AVE are shown in Table IV.

TABLE IV: NO OF SUBJECTIVE AND OBJECTIVE

	No of Subjective Terms	No of Objective Terms
FS	13052	142235
AVE	11678	143609

2) Content-specific features

Initially, a vocabulary of terms is constructed for each dataset using four parts of speech unigrams i.e. Adjectives (*a*), Adverb (*r*), Verb (*v*) & Noun (*n*) and syntactic phrases i.e. noun and verb phrases.

The unigrams features from these four part of speech groups are filtered to get most frequent unigrams. The new dimension features i.e. noun and verb phrases as a part of the content specific features are also filtered by removing the infrequent phrases.

3) Sentiment features

Sentiment features are the subjective terms available in the SentiWordNet and only consider those subjective terms as sentiment features which are present in the vocabulary. Sentiment features are considered as base line features in the proposed framework.

4) F-score for feature selection

F-score can be used to measure the discrimination between

[2] http://www.cs.cornell.edu/people/pabo/movie-review-data/

two sets of real numbers. F-score is the simple filter technique [13] to select the most discriminative features. Larger the F-score is, the more likely it is that the feature is discriminative.

Given training vectors x_k; k = 1, 2… m with $n+$ positive and $n-$ negative number of instances, then the F-score of the i^{th} feature is defined as:

$$F(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1} \sum_{k=1}^{n_+} (x_{k.i}^{(+)} - \bar{x}_i^{(+)})^2 - \frac{1}{n_- - 1} \sum_{k=1}^{n_-} (x_{k.i}^{(-)} - \bar{x}_i^{(-)})^2}$$

where $\bar{x}_i$, $\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$ are the averages of the i^{th} feature of the whole, positive, and negative data sets, respectively; $x_{k.i}^{(+)}$ is the i^{th} feature of the k^{th} positive instance, and $x_{k.i}^{(-)}$ is the i^{th} feature of the k^{th} negative instance. The discrimination between the positive and negative sets is indicated by the numerator.

To select more discriminative feature sets, F-Score of each feature is calculated with the help of LibSVM [14] feature selection tool (fselect.py) and then features with F-Score > 0.002 are selected in current experiments.

5) Features sets with/without feature selection

Incremental combination of features scheme was used to measure the effect and impact of new features addition in a base line set of features. In proposed framework effect and impact of content specific features has to be examined with sentiment features. Duplication at this stage is also removed by taking the union of different features during the process of combining different features.

Finally, eight feature sets are formulated with and without feature selection from unigrams, noun and verb phrases and sentiment features, no of features for each domain against each features set are shown in Table V.

TABLE V: NO OF FEATURE FOR A FEATURE SETS

Features Set	Description	Books	Dvd	Electronics	Kitchen	Movies
Senti	Sentiment Features	935	855	312	302	1886
SentiUni	Joint Sentiment and Unigrams Features	4878	4354	2649	2502	14021
SentiPhr	Joint Sentiment and NP & VP Features	5278	5392	2889	2419	13441
SentiUniPhr	Joint Sentiment, Unigrams and NP & VP Features	7951	7679	4432	3952	21846
SelSenti	Selected Sentiment Features	101	124	68	61	273
SelSentiUni	Selected Joint Sentiment and Unigrams Features	599	645	488	477	2211
SelSentiPhr	Selected Joint Sentiment and NP & VP Features	613	704	564	457	1971
SelSentiUniPhr	Selected Joint Sentiment, Unigrams and NP & VP Features	935	1016	785	713	3213

D. Classification

Literature analysis revealed that mostly, for text classification Support Vector Machine-SVM is used as the front line algorithm. Reason for using SVM is that it consistently outperforms Maximum Entropy and Naïve Bayesian as presented in [1].

SVM as the discriminative model use $g(x) = W^T \emptyset(x_i) + b$ the discriminant function. It follows the maximized margin principle which prevents over-fitting on huge sets of features, such that it will be computationally better if only a small set of the data points are considered as support vectors. The SVM optimization is defined by [14] as:

$$\min_{w,b,\varepsilon} \frac{1}{2} W^T W + C \sum_{i=1}^{l} \varepsilon_i$$

$$\text{s.t.} \quad y_i(W^T \emptyset(x_i) + b) \geq 1 - \varepsilon_i, \quad \varepsilon_i \geq 0$$

where b is the bias, W is the weights vector, and $\emptyset(x_i)$ maps the input space non-linearly into high-dimensional feature space. $C > 0$ is the penalty parameter of error term.

Furthermore, if it is intended to map a very high-dimensional feature space input vectors for finding the maximum-margin separator; it will be computationally intractable. Kernel trick can be used to resolve this problem. It depends on considering a way to map the high-dimensional feature space which lets fast scalar products.

Linear kernel is used in text classification due to quite large dimension of feature space and also the text classification problem is always linearly separable.

E. Validation and Evaluation

When there is limited amount of data for training and testing, then swapping the roles of training data and testing data a technique known as cross-validation. When 10-Fold Cross Validation to be used i.e. divide the data into 10 equal partitions such that 10% of data for testing and remaining 90 % for training and repeat the process for 10 times, ensure that each partition is used for testing exactly once. Finally, the estimates are averaged for an overall estimate.

Accuracy usually taken as an evaluation measure for balance datasets for classification and can be defined in light of traditional confusion matrix as shown in Table VI.

TABLE VI: CONFUSION MATRIX

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	TP	FN
	Negative	FP	TN

IV. EXPERIMENT SETUP

The setting includes mainly the text representation and classification tool used for this study.

A. Text Representation

Before executing machine learning techniques on textual data, a structured document representation needs to be

devised.

Bag of Features: Bag of features used in this study for the document vector by taking the relational independence assumption between features. The number of columns in a word vector is a function of the number of distinct features in the document collection, or its dictionary. This number can grow quite quickly with larger and richer documents and in this study sparse data representation used for very high dimensional word vector spaces with several thousand attributes.

Bernoulli Document Model: In [1] both, term frequency and term presence are used for feature weighting but binary term presence shows better results. Term Presence for feature value in vector is used in this research.

B. LIBSVM

LIBSVM [14] is integrated software for support vector classification(C-SVC & nu-SVC), regression (epsilon-SVR, nu-SVR) and distribution estimation (one-class SVM).

In this study, C-SVC multi class classification is used with linear kernel available in the LIBSVM. Also, in current experiments accuracy is taken as the performance measure with 10 fold cross validation.

V. Results

The objective of this research was to identify the pitfalls in the pre-existential SentiWordNet based sentiment classification frameworks alongside the optimization of classification features as a resultant of this research effort.

Phrases as the new features dimension proposed and their effects along the sentiment features on sentiment classification are illustrated in the current research. Table VII shows the results on the five domains and the bold face represents the best result.

A. Comparison with Previous Work

Results of proposed approach are tabulated in Table VIII along comparisons with pre-existential approaches.

Initially, this study is compared with the most renounced and baseline work done by [2] in sentiment analysis. It was the first bench mark in sentiment analysis based on the phrases. The best achieved results were on automobiles, and they were 84% while worst results were exhibited by the movies domain as 65.83%. The results are far below when compared to this study results.

TABLE VII: Accuracy Results of Different Feature Set

Measure	Features Set	Books	Dvd	Electronics	Kitchen	Movies
AVERAGE ACCURACY	Senti	69.25	72.25	72.10	74.40	73.20
	SentiUni	74.95	75.15	77.80	78.45	84.40
	SentiPhr	73.35	75.75	77.05	77.80	82.50
	SentiUniPhr	76.45	76.65	78.70	79.35	85.25
	SelSenti	71.85	74.65	73.10	75.65	81.60
	SelSentiUni	83.50	82.95	83.30	85.15	86.95
	SelSentiPhr	82.00	84.15	85.55	84.75	86.90
	SelSentiUniPhr	**85.20**	**85.40**	**86.05**	**86.80**	**89.30**

TABLE VIII: Results Comparison

Approach	Books	Dvd	Electronics	Kitchen	Movies
Phrases and PMI [2]	_	_	_	_	65.83
Unigrams with presence [1]	_	_	_	_	82.90
SentiWordNet Subjective & Objective (with revised score) Words [5]	_	_	_	_	78.50
Unigrams and Bi-Tagged Phrases[6]	–	–	–	–	89.40 F-measure
Sentiment, Content-specific unigrams & bigrams and Content-free features [4]	78.85	80.75	83.75	84.15	_
Ensemble of unigrams, bigrams & dependencies [3]	78.35	81.00	83.35	86.75	87.25
This Study(sentiment, Content-specific unigrams, NP & VP)	**85.20**	**85.40**	**86.05**	**86.80**	**89.30**

Subsequently, this study is compared with [1] different types of features were experimented based on unigram, bigrams, adjectives, POS and Position. Their combination of unigrams and bigrams not reflected the best result due to huge features dimensions of bigrams. They best achieved accuracy 82.9% on Movie reviews using unigrams with feature presence as weight but less than this study results.

Objective words are used in [5] with revised score along with subjective word of SentiWordNet to reduce the limitation of SentiWordNet. According to their experiments, the average accuracy is 71.89% for the original SentiWordNet and 76.02% for the revised SentiWordNet. They achieved best accuracy 78.5% on movie reviews less to this study results i.e. 89.30%.

Another motivation of using phrasal features with unigrams to address the sentiment classification was taken from [6]. It was identified by them that individual bi-tagged phrasal features resulted adversely and even the bigrams however their combinations with unigrams performed better. Their results of F-Measure 89.4% are still comparable with current study accuracy 89.3%.

Ensemble of features (POS based & Word relation) and ensemble of classifiers was studied in [3]. Results that are comparable with this study approach are Joint features and

Ensemble of features of both POS and dependency word relation. They achieved better results with the ensembles word relational features than ensemble of POS-based features. Results of current study are much better than their ensembles word relational features. On the contrary ensemble of different features increases the complexity also increasing the computations manifolds.

The most comparable and motivational study to this research was carried by [4]. Sentiment features were new dimension to enhance the classification performance. But combination of content free and content specific features performed better than combination of content free and sentiment features. There was a need to re-think on content-specific features to get a good match. Current study proposed the new dimension of content specific features with noun and verb phrases and considered the sentiment features as base line features. On the multi domain product reviews dataset they achieved best accuracy 78.85~84.15 which is considerably less than current study i.e. 85.20~86.80% on the same dataset.

VI. Conclusion

It is clearly visible from current research results that sentiment classification performance has been enhanced with the inclusion of content specific features alongside sentiment features. Text exhibits limited sentiment performance with sentiment features only. In literature different individual features and joint features were used but the study in hand used feature sets which proved as best representative features for the sentiment classification task. A significant improvement in performance was observed after feature selection.

Although individual sentiment features are smaller in number and easy to use as compare to combination of sentiment and content specific features. However, content specific features add a valuable content representation in sentiment features. The performance of sentiment classification is more critical; the proposed features i.e. combination of sentiment , content specific phrases and unigrams, have much better classification accuracy i.e. 85.2~89.3%.

VII. Future Work

More pre-processing capabilities such as dependency parsing for word relation features could be used with sentiment features. Neural Networks and other latest classifiers may be adapted and ensemble of these classifiers can be experimented with this study features as baseline. This framework is general in nature so it can easily be adapted and ported to other domain datasets e.g. blogs, emails and tweets etc. as well as on other than product reviews datasets. We are also interested in future to validate this research framework in the multi lingual perspective based on the availability of respective lexicon resource.

References

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up: Sentiment classification using machine learning techniques," in *Proc. the ACL-02 Conference on Empirical Methods in Natural Language Processing*, July 2002, vol. 10. pp. 79-86.

[2] P. Turney, "Thumbs up or thumbs down: Semantic orientation applied to unsupervised classification of reviews" in *Proc. the ACL 40th Annual Meeting on Association for Computational Linguistics*, July 2002, pp. 417-424.

[3] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Information Sciences*, vol. 181, no. 6, pp. 1138-1152, 2011.

[4] Y. Dang, Y. Zhang, and H. Chen, "A lexicon-enhanced method for sentiment classification: An experiment on online product reviews," *IEEE Intelligent Systems*, vol. 25, no. 4, pp. 46-53, 2010.

[5] H. Chihli and H. Lin, "Using objective words in Senti word net to improve sentiment classification for word of mouth," *IEEE Intelligent Systems*, pp. 48-57, 2013.

[6] B. Agarwal, N. Mittal, and E. Cambria, "Enhancing sentiment classification performance using bi-tagged phrases," in *Proc. IEEE 13th International Conference Data Mining Workshops (ICDMW)*, Dec. 2013, pp. 892-895.

[7] N. Chomsky, *Syntactic Structures*, Publisher. Walter de Gruyter, ch. 3, pp. 26-33, 2002.

[8] L. Schwartz, C. C. Burch, W. Schuler, and S. Wu, "Incremental syntactic language models for phrase-based translation," in *Proc. the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, vol. 1, pp. 620-631.

[9] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Proc. ACL*, June 2007, vol. 7, pp. 440-447.

[10] H. Cunningham. "GATE, a general architecture for text engineering," *Computers and the Humanities*, vol. 36, no. 2, pp. 223-254, May 2002.

[11] S. Baccianella, A. Esuli, and F. Sebastiani, "Senti word net 3.0: An enhanced lexical resource for sentiment analysis & opinion mining," in *Proc. LREC*, May 2010, vol. 10, pp. 2200-2204.

[12] L. Gatti and M. Guerin, "Assessing sentiment strength in words prior polarities," arXiv preprint arXiv: 1212. 4315, 2012.

[13] Y. Chen and C. Lin, "Combining SVMs with various feature selection strategies," *Feature Extraction*, Springer Berlin Heidelberg, pp. 315-324, 2006.

[14] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 27, no. 3, 2011.

M. Latif was born in Wah Cantt on February 28,1981 and is currently a MS student (software engineering) in Department of Computer Engineering, College of Electrical & Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan. His research area is in text mining, opinion mining, named entity and relation extraction from biomedical texts etc.

U. Qamar got the Ph.D (information systems) in 2010 from University of Manchester, Manchester, England. He is an assistant professor in Department of Computer Engineering, College of Electrical & Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan.

He is a centre director in Data and Text Mining Centre (DaTCen). His research area is in data mining, outlier detection and feature selection.

A. Wahab got the Ph.D (scholar) degree in National University of Sciences and Technology, Pakistan. He is an officer (software) in Department of Computer Engineering, College of Electrical & Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan.

He is a member of Data and Text Mining Centre (DaTCen). His research area is in text mining, feature extraction and relation mining from biomedical text.