Syntactic Genetic Algorithm for a Subjectivity Analysis of Sports Articles

Conference Paper · January 2012

CITATIONS
2

READS
308

2 authors:

Yara Rizk
American University of Beirut
18 PUBLICATIONS **32** CITATIONS
SEE PROFILE

Mariette Awad
American University of Beirut
111 PUBLICATIONS **420** CITATIONS
SEE PROFILE

Syntactic Genetic Algorithm for a Subjectivity Analysis of Sports Articles

Yara Rizk, *Student Member, IEEE*, Mariette Awad, PhD, *Member, IEEE*
Electrical & Computer Engineering
American University of Beirut
Beirut, Lebanon
{yar01, mariette.awad}@aub.edu.lb

***Abstract*—The abundance of information on the World Wide Web and the existing content authentication mechanisms render the ability to efficiently find factual information often challenging and time consuming. This situation calls to the user's judgment and knowledge about the sought topic. For sports articles, more specifically, where information is often used for casual and formal betting, an intelligent classifier which attempts to identify articles' subjectivity would be interesting to explore. Motivated to apply artificial intelligence to automate classification of sport articles, we propose to use genetic algorithm with syntactic features for subjective content analysis. The model was tested on a home-made corpus of three hundred sports articles of up to 1600 characters. The experimental results based on our proposed evaluation function and a 3-fold cross fold validation achieved an average of 96.2% on the training set and 94.5% on the test set which motivates follow-on research using different classifiers.**

Keywords—Natural Language Processing; Genetic Algorithm; Syntactic Features; Content Analysis

I. INTRODUCTION

While reading this report, thousands of sports articles are being written and published on the World Wide Web for readers to digest. The great progress in communication and storage technologies has rendered the average internet user overwhelmed with a myriad of information, with yet a limited authentication mechanism in place. Thus, whenever sports fans, for instance, aim to search for facts or unbiased information to make educated guesses or bets on some game winners, they end up examining many articles that mix between facts and opinions and are subjectively shaped by the writers' style and affiliations.

Subjectivity in articles is difficult to identify because it can embody few forms. It can be directly expressed or inferred. It can also be conveyed in a variety of sentence structures and through diverse words and phrases with varying intensities. It is sometimes difficult to identify if a certain part of an article represents the author's opinion or if it is simply a report on someone else's opinion. The latter would make this article segment an objective statement since it is not the author's personal opinion. Furthermore, people disagree on the matter of subjectivity, as some people might find certain articles subjective, others find them objective. In fact, making judgments on the subjectivity or objectivity of a text is yet an on-going outcome of advanced natural language processing (NLP).

Motivated to develop an automated text subjectivity analyzer that classifies sports articles based only on a syntactic analysis, which to the best of our knowledge has not been yet investigated, we propose in this paper a genetic algorithm (GA) framework to assess the subjectivity measure of such articles. Our focus on sport articles is partly supported by our genuine interest in sports and partly by the fact that popularity of sport betting is increasing. Some studies estimated that about 15% of Americans engage in sports betting, a figure that often increases to 25% during major sporting events to reach billions of dollars per year [1]. Also, a German firm had projected that the sport gambling market will reach 263 billion Euros in 2012 [2].

The organization of this paper is as follows. Section 2 summaries related work done mostly in the field of sentiment analysis. While Section 3 describes the proposed solution, Section 4 reports on the experimental results obtained. Finally, Section 5 concludes with follow-on research potentials.

II. LITERATURE REVIEW

Few researchers are focusing on sentiment analysis, a popular and emergent field of NLP aiming at identifying polarities and strengths of authors' opinions and views from text [3]. Proposed solutions varied: some studies addressed it at a word level, i.e. classifying words based on their sentiment or opinion content; others at a sentence level; and some based on the entire text. We will review, in that order, the surveyed literature that is most related to our research.

Reference [4] introduced the lexical library SENTIWORDNET, which computes for each word a positive subjectivity score and a negative subjectivity score using a group of three classifiers that assess if each term is objective, negative, positive or neutral. The classifiers were not trained on the same set of words but SENTIWORDNET was tested using the General Inquirer [5]. The best result obtained for objective-subjective classification was an accuracy of about 68% as reported in [6].

Other researchers proposed classifying sentences based on sentiment analysis. Reference [7] determined whether the phrase is neutral before computing its polarity based on context. Training and testing of the classifiers were done on the Multi-perspective Question Answering Opinion Corpus with an added contextual polarity to the corpus instances. The results for the first neutral-polar phase varied from 73.6 to 75.9%, while the second phase achieved an accuracy ranging from 61.7 to 65.7%.

Authors in [8] presented OpinionFinder, a software capable of classifying sentences as objective or subjective using a Naïve Bayes classifier. The classifier was trained on non-annotated corpus sentences however no concrete results were included in the paper.

In a recent work, [9] used information extraction techniques to which a lexicon of subjective indicators is bootstrapped to produce an unannotated classifier. The reported accuracy was 91% on a relatively small corpus.

Authors in [10] investigated the impact of adverbs of degree and adjective-adverb combinations on the objectivity and subjectivity classification of documents, in a syntactic and semantic approach to sentiment analysis. Based on 200 annotated news articles, experimental results recommended the usage of adverbs. Their work suggested that adverbs should be given a weight of 35% of the weight of adjectives. Unlike [10], we are assessing the impact of all the structural elements of the article, not only adjectives and adverbs, for sports articles using a GA and without taking any semantic information into consideration.

Reference [11] attempted to classify documents based on lexicons created using several methods such as PageRankSWN and SENTIWORDNET. A group of classifiers including Support Vector Machines (SVM) and Rocchio classifiers was used on a corpus of positive and negative movie reviews. Results ranged from 49.7 to 59.5% as an overall accuracy.

Pang et al. classified movie reviews as objective or subjective using Naïve Bayes, SVM and maximum entropy classification [12]. Different features such as unigrams, bigrams, adjectives, parts of speech, etc., generated varying results, for the different classifiers, in the accuracy range of 77 to 83% [12].

Authors in [13] focused on opinion mining in financial news documents. They analyzed lexical and syntactical features, punctuation and other part of speech. They compiled a corpus of financial news articles from different sources, of which only 30 articles were manually annotated and referred to as the Gold Standard. With an overall accuracy of 47%, their precision was about 66% for negative texts as opposed to only 38% for positive ones.

In [14], Heerschop et al. only studied the impact of taking into account negation in the classification of documents. They compiled a random set of articles that were manually classified. They deduced that accounting for negation improved accuracy by 1.17%, from 70.41 to 71.23% on their entire corpus.

Finally, [15] performed sentiment analysis on news articles and blogs using Wordnet. This was accomplished by breaking down the text into distinct entities and computing a score indicating how much positive or negative the opinion is in each entity. They also reported on how much positive or negative a person or issue is portrayed in news versus blogs. Testing their classifier on a large database of news articles and blogs that they compiled, authors noted a tradeoff between high precision and recall. For example, their precision decreased from 0.957 to 0.827 as the recall values decreased from 0.712 to 0.694.

III. Proposed Solution

A. Overview

As shown in Fig. 1, the input file is preprocessed first to eliminate unrecognizable characters or symbols. It is fed to the Stanford Part of Speech (POS) Tagger package [16] that is complemented based on [17] to identify POS of each word (i.e. nouns, adjectives, adverbs, verbs tenses etc.). Additional functions were added to the tagger to differentiate between first, second and third person pronouns and between imperative and base form verbs. Once the POS tagger extracts the article's syntactic features, a GA classifier labels the article either as subjective or objective based on a proposed classification function.

B. Feature Generation

Our classifier adopts a frequentist approach mainly motivated by human perception of speech. A human reader can usually evaluate the subjectivity of an article by inspecting, in a glance, the presence of specific syntactic features and their frequency of occurrences such as the excessive usage of adjectives. Upon focused discussion with some English linguistics at the American University of Beirut, we decided to choose the syntactic attributes listed in Table I as a measure of indication of objectivity and subjectivity in sport articles.

Our justification for this partitioning and selection is mainly because:

- *Quotations* rather insinuate objectivity since they indicate that the author is reporting something said by others; i.e. s/he is not expressing directly her/his opinion.
- *Question marks and exclamation marks* show evidence of subjectivity since the author could be inquiring for information, expressing her/his surprise or emphasizing some news.
- *Past tense verbs* point to objectivity since they involved events that occurred in the past and are being narrated by the author. However, it is important to also consider the pronoun used in conjunction with the verb since a past tense verb with the first or second person pronoun infer a sentence that is rather somewhat subjective.
- *Third person pronouns* refer to objectivity, while *first and second person pronouns* indicate subjectivity.

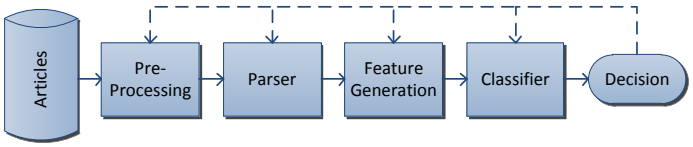

Fig. 1. Block diagram of the automated subjectivity analyzer.

TABLE I. Objective and Subjective Syntactic Features

Parameters		Objective	Subjective
Punctuation	Quotations	x	
	Question marks		x
	Exclamation marks		x
Pronouns	First & Second person		x
	Third person	x	
Verb Tenses	Past tense	x	
	Imperative tense		x
	Present tense	x	
Adjectives & Adverbs	Comparative & Superlative		x

- *Present tense verbs* with first person and second person pronouns show the opinion of the author and hence are subjective parameters. Present tense verbs with third person pronouns mostly indicate objectivity, with a few instances where they could indicate subjectivity.
- *Imperative verbs* reflect the author's opinion since s/he would be encouraging or promoting some action. Thus, this parameter indicates subjectivity.
- *Comparative and superlative adverbs and adjectives* in most cases indicate subjectivity since the author is comparing things based on subjective beliefs.
- *Numerical values and dates* in sports articles point out to some reported statistics which are an indication of objectivity.

C. Classifier

The classifier uses the features generated by the POS tagger to categorize articles according to a workflow that separates the inference step from the classification phase, for more leverage.

1) Classification Function: Each of the features discussed earlier has a different level of importance in revealing the degree of objectivity or subjectivity of the article. Since an article is best categorized with an overall structural analysis of all sentences, we selected a classification function which is a weighted sum of these features as shown in (1):

$$f(x) = \sum_i^{f_1} a_i x_i - \sum_j^{f_2} b_j x_j + c x_k \qquad (1)$$
$$a_i > 0, b_j > 0, c > 0$$

Note that in (1), x represents the syntactic features, f_1 and f_2 represent the number of objective and subjective features respectively, a_i the weights of the attributes indicating objectivity, b_j the weights of the features referring to subjectivity and c the weight of the *last sentence* class. Most authors conclude by restating the main idea of their article. Therefore, the *last sentence* can have an influence on the overall article classification. Hence, we propose adding the classification of this sentence as a feature in the classification function. $x_k = +1$ if the last sentence based on the syntactic features of Table I is objective and $x_k = -1$ otherwise. Hereafter, Model 1 refers to the design that uses the features in Table I as input to the classifier and Model 2 refers to the case that in addition to the features in Table I, the classification of the last sentence is added to the classification function.

2) Decision Stage: Obviously, if $f(\mathrm{x}) > 0$, the article is rather objective, and if $f(\mathrm{x}) \leq 0$, the article is mostly subjective. To insure confidence in the decision stage for values of $f(\mathrm{x}) \sim 0$, we propose the following inference steps. If:

- $f(\mathrm{x}) \geq 1$, then the article is classified as objective.
- $f(\mathrm{x}) \leq -1$, then the article is labeled as subjective.
- $-1 < f(\mathrm{x}) < 1$, then the classifier places the article in a fuzzy region for more in-depth analysis subject of our future work.

A GA was implemented to learn in a supervised way the values a_i *and* b_j on a corpus that was manually annotated by the authors. The structure of this GA is rather the generic one with the following specifics:

1) Encoding: Each coefficient is represented by a set of bits, genes in the chromosome, based on the permissible range of values for this coefficient as shown in Fig. 2.

2) Selection procedure: The GA uses a roulette wheel selection process to select the candidate chromosomes for crossover. The population contains 20 chromosomes, from which 10 chromosomes are selected for cross over in each generation.

3) Evolutionary operators: A two-point crossover scheme is used where the two cut points are randomly selected. After crossover, 5 children chromosomes are created. Mutation is performed on the generated children with a rate of 80%.

4) Fitness function: To evaluate how fit a chromosome is, the classification function discussed earlier in (1) is calculated for each article in the training phase. The number of correctly classified articles is used as the fitness of the chromosome. As the number of correctly classified articles increases, the fitness of the chromosome improves, indicating a fitter chromosome.

5) Replacement technique: A child chromosome with a fitness function greater than the least fit chromosome in the parent population replaces the latter.

6) Initialization and termination conditions: The initial population is randomly generated. The termination criterion is to correctly classify 100% of the total number of articles in the training set or to reach 1000 iterations.

IV. Experiments Results

A. Corpus Description

We restricted the scope of the problem at hand to English written sports articles for ease of implementation. Three hundred articles were manually annotated; half of these articles were subjective, and half were objective. The articles were obtained from fifty different websites such as ESPN, Foxsports, Eurosport UK, etc. and written by different authors. While some of these articles were written by sports fans on blogs, most articles were written by professional journalists. The length of these articles ranged from 70 to 1600 words.

B. Cross Fold Validation Results

To test the performance of our proposed solution, we designed few experiments with cross-fold and shuffling. The code was written in Java, NetBeans IDE 6.9 on an Intel Core i5 processor. The corpus was divided into three folds without any constraints on the demographic in each fold, i.e., the number of objective and subjective articles in each fold varied, often resulting in an unbalanced set.

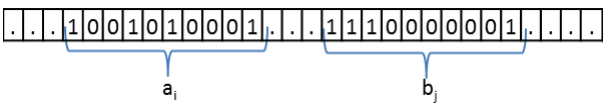

Fig. 2. Chromosome structure.

The GA was run for 1000 generations starting from a random initial population to obtain a set of coefficients for the classification function. Fig. 3 shows the accuracy of each set of coefficients versus the generation at which this set first entered the population. The percentage of data points in each region is reported in the rectangular boxes. Ideally, the aim is to reach fast convergence i.e. minimum generation number while maximizing the accuracy. Hence, we would have liked to obtain most points in Region 2. Based on the information in Fig. 3, Model 2 performed better than Model 1. First, Model 2 had a greater percentage of points in Region 2 compared to Model 1 (75 vs. 72%) in training. Model 2 had 88% of data points whose accuracy exceeded 95% compared to 85% of data for Model 1. Also, the minimum accuracy attained in Model 2 was 94% compared to 93% for Model 1. While testing, Model 1 produced better results than Model 2; 43% of the data points were in Region 2 compared to 45%, respectively. Furthermore, 52 and 46.4% of the data points attained accuracy greater than 95% for Model 1 and Model 2, respectively. Hence, Model 2 converged faster to higher accuracies.

The GA was trained on two out of three folds of the corpus. The bar graphs in Fig. 4 display the accuracy of the different runs of Model 1 and Model 2. The average of the results is indicated by the horizontal lines. Comparing the results for both models, Model 2 attained a higher accuracy than Model 1. Model 1 achieved 95.2% overall accuracy with 95.5% of the objective and 94.9% of the subjective articles correctly classified. Model 2 reached 97.2% overall accuracy with 97.8% of the objective and 96.7% of the subjective articles correctly classified. Overall, the percentage of correctly classified objective articles is greater than that of subjective articles. This is not surprising since objective articles have a similar structure whereas subjective articles are more diverse and hence difficult to identify.

The results of the testing phase shown in Fig. 5 are consistent with the results obtained during training. The subjective articles have a higher error rate compared to objective articles and Model 2 has a lower error rate than Model 1. On average, Model 1 misclassified 6.5% of articles, of which 4.7% were objective, and 5.8% were subjective. Whereas for Model 2, an overall error rate of 4.5% was achieved, with only 2.6% of objective and 4.3% of subjective articles misclassified. If we compare the percentage of articles placed in the fuzzy area for the two models, we notice that Model 2 has a slightly lower percentage of articles in this region (1.1%) compared to Model 1 (1.3%).

C. *Leave One Out and Features Validation*

Since Leave-One-Out (LOO) validation portrays the worst case performance of the classifier, the GA was trained and tested using LOO cross validation. The average results over 5 runs are reported in Table II. The worst case accuracy obtained for Model 1 is 90.6% when using the features in Table I. To assess the importance of the comparative and superlative adjectives and adverbs as features, we removed them from the feature set. Results show an acceptable decrease in classification accuracies for both models.

TABLE II. LOO TRAINING AND TESTING RESULTS FOR DIFFERENT FEATURE SETS

Feature Sets		Training Overall (%)			Testing average (%)		
		Min	Max	Average	Correct	Wrong	Fuzzy
With Adjectives & Adverbs	Model 1	93.98	95.99	94.89	90.60	5.00	4.40
	Model 2	95.99	98.33	96.96	94.47	3.33	2.20
Without Adjectives & Adverbs	Model 1	93.31	94.98	94.06	88.33	7.00	4.67
	Model 2	95.99	98.66	96.96	94.33	2.34	3.33

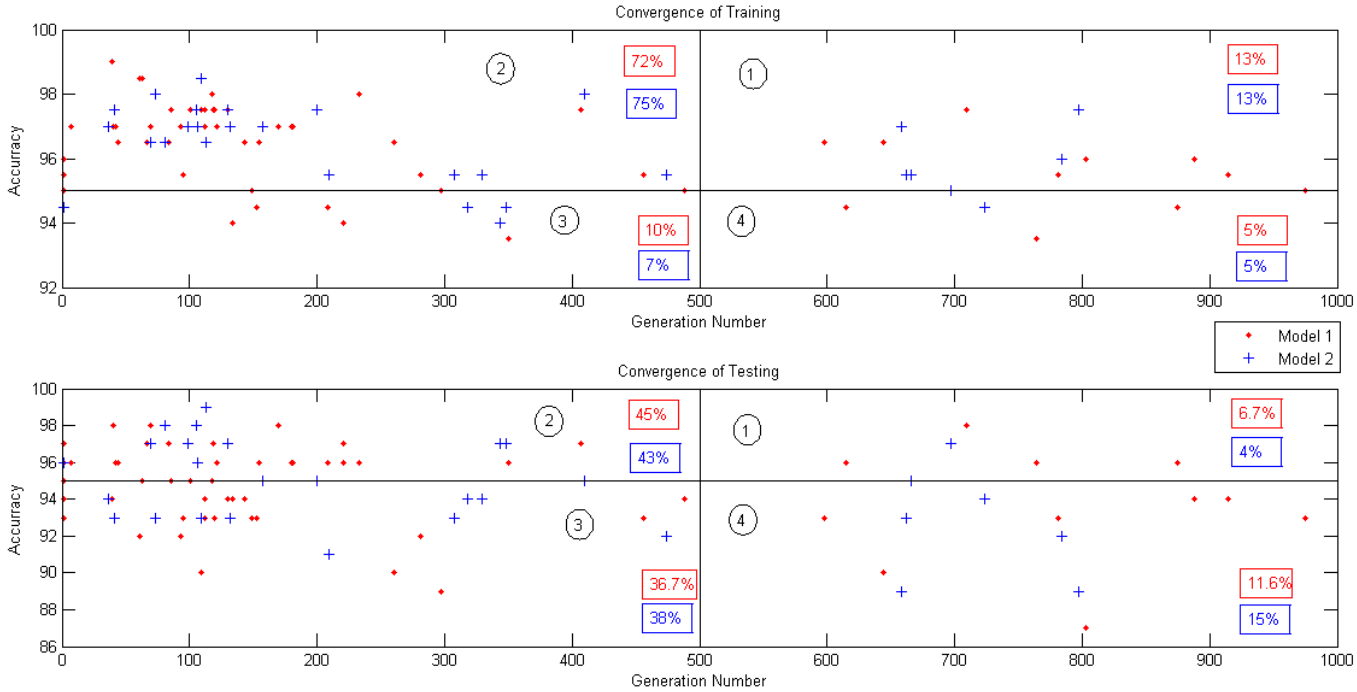

Fig. 3. Convergence of GA vs. accuracy in training and testing phases.

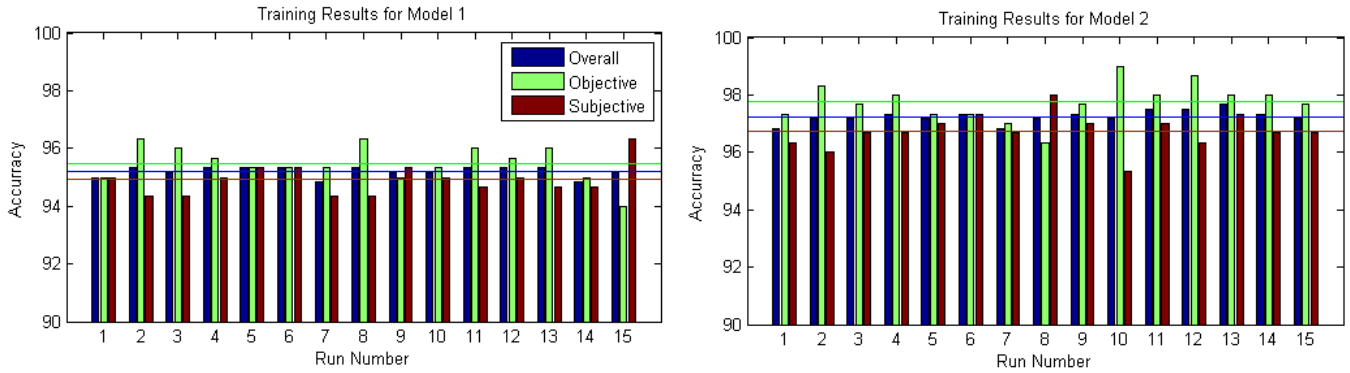

Fig. 4. Training results for model 1 and model 2.

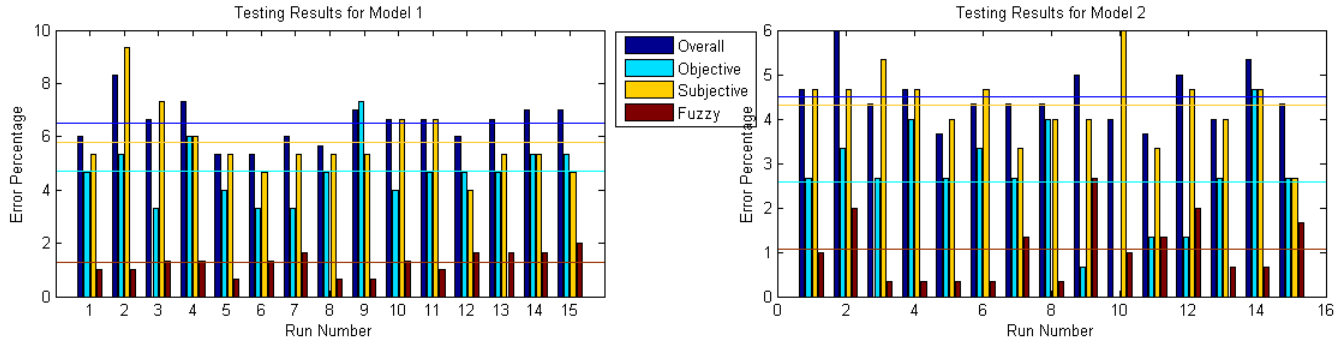

Fig. 5. Testing results for model 1 and model 2.

V. CONCLUSION

This work proposed an automated text subjectivity analyzer capable based on a frequentist approach of classifying articles as objective or subjective by only examining the grammatical structure of the article. The classification function proposed a weighted sum of specific syntactic features. A GA was trained and tested to find the best set of weights. On average, the classifier reached an overall training accuracy of 96.2% using cross fold validation. While during testing, 94.5% of articles were correctly classified on average, 4.3% were wrongly classified, and 1.2% of the articles were placed in the fuzzy region. Although these results are good, future work may explore classifiers other than GA as well as different feature sets.

ACKNOWLEDGMENT

This work is supported by the University Research Board at the American University of Beirut. We would also like to acknowledge the help of Dr. Lina Choueiri from the Department of English at the American University of Beirut.

REFERENCES

[1] Jerry Martin (2012). *History of Sports Betting* [Online]. Available: http://www.ultimatecapper.com/history-of-sports-betting.html

[2] Bill Wilson (2010). *Sport betting industry looks to protect itself* [Online]. Available: http://www.bbc.co.uk/news/business-11309620

[3] *Handbook of Natural Language Processing*, 2nd ed., CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010, pp. 627-666.

[4] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining," in *Proc. of Language Resources and Evaluation*, 2006, pp. 417-422.

[5] P. J. Stone and E. B. Hunt, "A Computer Approach to Content Analysis: studies using the general inquirer system," in *Proc. Spring Joint Computer Conf.*, 1963, pp. 241-256.

[6] A. Esuli and F. Sebastiani, "Determining term subjectivity and term orientation for opinion mining," in *Proc. 11th Conf. of the European Chapter of the Association for Computational Linguistics the European,* Trento, Italy, 2006, pp. 193-200.

[7] T. Wilson, J. Wiebe and P. Hoffmann , "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proc. of the Human Language Technology Conf. and the Conf.on Empirical Methods in Natural Language Processing*, 2005, pp. 347–354.

[8] T. Wilson *et al.,* "OpinionFinder: A system for subjectivity analysis," in *Proc. of HLT/EMNLP on Interactive Demonstrations,* 2005, pp. 34-35.

[9] J. Weibe and E. Riloff, "Finding Mutual Benefit between Subjectivity Analysis and Information Extraction," *IEEE Trans. Affective Computing*, Dec. 2011 pp.1.

[10] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato and VS Subrahmanian, "Sentiment analysis: Adjectives and adverbs are better than adjectives alone," in *Proc. of the Int. Conf. on Weblogs and Social Media*, 2007.

[11] B. Heerschop, A. Hogenboom and F. Frasincar, "Sentiment Lexicon Creation from Lexical Resources," in *14th Int. Conf. on Business Information Systems,* Vol.87, Springer, 2011, pp. 185–196.

[12] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs Up? Sentiment Classifcation using Machine Learning Techniques," in *Proc. Empirical Methods in Natural Language Processing,* Philadelphia, 2002, pp. 79-86.

[13] A. Devitt and K. Ahmad, "Sentiment analysis in financial news: A cohesion-based approach," in *Proc. of the Association for Computational Linguistics*, 2007, pp. 984–991.

[14] B. Heerschop, P. Van Iterson, A. Hogenboom, F. Frasincar and U. Kaymak, "Analyzing Sentiment in a Large Set of Web Data while Accounting for Negation," in *Advances in Intelligent Web Mastering–3*, 2011 pp. 195-205.

[15] N. Godbole, M. Srinivasaiah and S. Skiena, "Large-scale sentiment analysis for news and blogs," in *Proc. of the Int. Conf. on Weblogs and Social Media*, 2007, pp. 219-222.

[16] K. Toutanova, D. Klein, C. D. Manning and Y. Singer, “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network,” in *NAACL '03: Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Edmonton, Canada, 2003, pp. 173–180.

[17] E. Atwell. *The University of Pennsylvania (Penn) Treebank Tag-set* [Online]. http://www.comp.leeds.ac.uk/amalgam/tagsets/upenn.html