Evaluation of Sentiment Polarity Prediction using a Dimensional and a Categorical Approach

Ioana Mureșan[*], Adriana Stan, Mircea Giurgiu
Communications Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania

Rodica Potolea
Department of Computer Science
Technical University of Cluj-Napoca
Cluj-Napoca, Romania

***Abstract*—In this paper we evaluate two approaches for predicting the sentiment polarity of an utterance. The first method is based on a 3-dimensional model which takes into account text expressiveness in terms of valence, arousal and dominance. The second one determines the word's semantic orientation according to Chi-square and Relevance factor statistic metrics. We describe the general flow of the methods and their extracted features, as well as their predictability potential using different machine learning algorithms, Naïve Bayes, SVM and C4.5. The evaluation is performed on four emotional datasets: Semeval 2007 "Affective Text", ISEAR (International Survey on Emotional Antecedents and Reactions), children's fairy-tales and a movie review dataset. The results show a high correlation of the prediction performance with the database content, as well as to the average number of words within the classified text instances.**

Keywords- sentiment polarity; VAD model; statistic metrics.

I. Introduction

The recognition of emotions has gained popularity over time, mostly due to the efforts of psychology and behavioural sciences which focused their attention towards understanding human nature and its components. Human emotions tend to have a multi modal aspect, being conveyed by multiple means of expression that can include acoustic, textual and visual features. With the expansion of social networks and the increase in the amount of subjective online content, emotion recognition became an important domain for the Natural Language Processing (NLP) and speech processing community, also. The interest was intensified by using the collected emotional data in practical applications, such as marketing analyses and opinion mining, artificial intelligence and robotics, natural language interfaces for e-learning environments or edutainment games. The identification of emotions from text becomes a fundamental part for the new generation of human-computer devices that strive to offer a highly natural interaction.

Taking as a basis the concept that each word has a given emotional state that can vary upon context and the large amount of available textual data, two major text-based emotion representation models have emerged. The first model consists of a *dimensional* representation of each word, consistent with the psychological studies [1], and the second one is based on a *categorical* model, which is more suitable for computational linguistics, consisting in generating subsets of words around certain associated labels. Most notable categorization models are the following: **Ekman's six basic emotions** (anger, disgust, fear, joy, sadness and surprise), **subjectivity** (subjective vs. objective), **polarity** (positive vs. negative vs. neutral) and **stubbornness** (opinionated vs. non-opinionated). The categorical model benefits from the existing emotional thesaurus, like WordNet-Affect [2], ConceptNet [3] and SentiWordNet [4], but it can also generate important affective word lists using techniques like bag-of-words, keyword spotting and lexical affinity.

In this paper we evaluate the sentiment polarity prediction, bearing in mind the long-term goal of performing this task in an unsupervised, language and domain independent manner for text-to-speech applications. We therefore reduced the number of supervised steps involved in a common sentiment polarity prediction framework. The remaining processes can be easily replaced by unsupervised techniques. The evaluation is performed on several available emotional datasets using the dimensional and categorical models, and by employing three different types of classifiers: discriminative, generative and tree-based.

The paper is organised as follows. Section 2 will provide a brief overview of relevant research in the field of sentiment analysis. Section 3 describes the proposed method. The datasets and a set of experiments to measure the algorithm performance are presented in Section 4. We conclude in Section 5 with discussions and future work.

II. Related Work

A first attempt to classify emotions using textual data and a categorical model, was focused on determining the semantic orientation of words using two base categories: positive and negative. Hatzivassiloglou et al. [5] used a set of manually labeled seeds and explored the semantic orientation of adjectives by exploiting constraints on conjoined adjectives. They assumed that the conjunction "and" links adjective of the same polarity while the conjunction "but" links adjectives of opposite polarity.

This approach was extended by Turney et al. [6] who took into account other parts-of-speech considered to be responsible for the expressiveness, such as adverbs. Therefore, based on the concept that a positive semantic orientation denotes praise (e.g. "honest") and a negative semantic

[*]Corresponding author: ioana.muresan@com.utcluj.ro (I.Mureșan)

orientation indicates criticism (e.g. "disturbing"), their goal was to determine the intensity (mild or strong) a term brings to the class determined by valence (positive or negative). In order to achieve this, Point Mutual Information (PMI) and Latent Semantic Analysis (LSA) have been employed to infer direction and intensity of the semantic orientation in an automatic manner. This study raised criticism due to language variation, words potentially having multiple meanings [9].

Another study that felled short of the same type of criticism was made by Guohong Fu et al. [7], and it took into account the semantic similarity of two neighboring words. They started from a list of polar words, named "sentiment morphemes", trying to extend the analyses from the level of isolated words to the sentence level by using fuzzy sets and Chi-Square. "Sentiment morphemes" were categorized as static polar words and dynamic polar words, first referring to words that are not affected by changes of context ("love"), and words sensitive to context ("very").

Taking benefit from the intensive growth of subjective web content, new approaches emerged. In order to avoid the language dependent aspect and the limited information provided by lexical resources, Leonid Velikovich et al. [8] proposed a fully unsupervised method, which aimed at building large polarity lexicons semi-automatically from the web using a graph propagation algorithm. Another approach [9] took a large amount of web data, respectively a corpus of 200,000 online reviews, in order to try and determine the impact of high-order n-grams and the most suitable classifier for prediction of the reviews' polarity. High-order n-grams are required in polarity analysis due to the larger context they can capture. Both generative and discriminative classifiers were taken into account for categorization of text fragments into positive and negative, their prediction accuracy being influenced by the features given as input. For the generative approach, a Language Modeling based classifier (CMU – Cambridge Language Modeling Toolkit [10]) was considered, whose prediction is based on the probability of generating a word sequence, while the discriminative approach has been evaluated using the Passive Aggressive algorithm [11], which is a margin based online learning classifier. Winnow [12] was the third classifier under testing, this being an online linear classifier for sentiment analysis. Due to the significant differences between classifiers, some preliminary characteristics should be signaled. For example, the generative classifier may present data sparseness due to the limitation of training data, while the discriminative one is very adequate to large data because it uses an online learning pattern and it has a theoretical loss bound which makes its performance predictable.

Considering both models, dimensional and categorical, [13] had the purpose of recognizing the four affective states in text data: anger, fear, joy, sadness and neutral. The dimensional model takes the approach of numerically mapping emotions based on the three dimensions (valence, arousal and dominance) provided by the normative database ANEW (Affective Norms of English Words) [14]. The categorical model implies the use of emotional thesaurus WorldNet-Affect, Vector Space Models, and three dimensionality reduction techniques: Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA) and Non-negative Matrix Factorization (NMF). Experiments show that the categorical model using NMF and the dimensional model tend to perform best.

A similar method, [15] combines a chain of common tools in the natural language processing domain, with the dimensional model and statistical techniques, for feature reduction and weighting. A lexical analyzer was employed to filter out stop words, highlight valence shifters and negations, while a sentence splitter divided the text fragment into sentences. Following up, the presumed affective words were reduced only to nouns, adverbs and adjectives, through a part-of-speech tagger, and word context variation was resolved by a word-sense disambiguation. The last two chained components were a stemmer and a keyword spotter.

Lately, speech researchers developed an interest into emotion conveyed through text, such that semantic orientation of text might enrich the expressiveness of synthesized voices, by integrating lexical semantics as an input feature to text-to-speech systems. Zhiyong Wu et al. [16] represented text according to PAD (pleasure, arousal and dominance), and extracted expressive speech features (intonation, intensity, speaking rate and fluency). By analyzing emotional and non-emotional speech recordings, their strategy was to create a nonlinear perturbation model to transform neutral speech into expressive speech. This nonlinear perturbation model is composed by two perturbations: a local at the prosodic word level and a global at the utterance level. The first two dimensions, V (valence) and A (arousal), were used to describe the expressivity at prosodic level based on lexical semantics, while the third dimension, D (dominance), was used to draw the intonation at utterance level.

III. PROPOSED METHOD

In this section we detail the two proposed methods, *categorical* and *dimensional*, and the machine learning algorithms used for classifying text instances into polarity classes. Fig. 1 presents the processing steps for both methods. The first part of the work flow is common and includes:

- *Pre-processing*: punctuation sign stripping and transforming the words to lowercase;
- *POS tagging*: determines the part-of speech for each word within an utterance. This step provides two advantages: stop words are eliminated and emotional words, which are usually adjectives, adverbs, verbs and sometimes nouns can be highlighted; computation is reduced by discarding words that do not have an affective value;
- *Lemmatizing*: determines the lemma of each word. This helps to normalize the term variety, and thus facilitates the search through the normative database. It also improves the building of the bag-of-words by considering their lemma as a unique form.

The difference between the dimensional and the categorical models lies within the manner in which the sentiments are mapped, that is numerically, using a normative

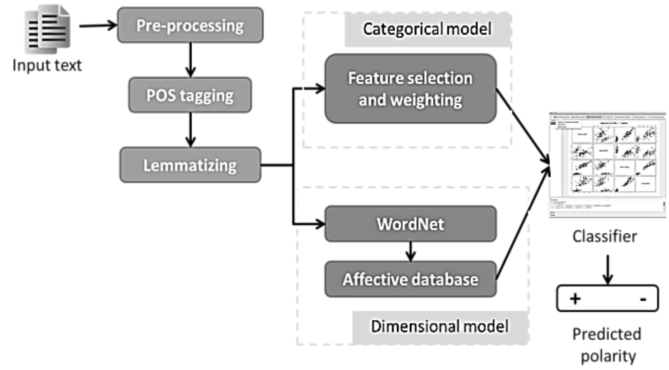

Fig. 1. Proposed methods block diagram. The separation between the categorical and the dimensional models is made after the lemmatizing block

affective database in the dimensional model and using the bag-of-words concept in the categorical model. For the dimensional model, due to the limited number of words within the normative database, we also included WordNet in order to extend the list of words by adding their synonyms. Next, we describe each model's particularities, and the features used for classification.

A. Dimensional Model

The dimensional approach is based on a psychological model in which word embedded emotions are mapped in a 3 dimensional space using the notions of: *valence*, *arousal* and *dominance* – VAD [17]. Valence represents the polarity aspect of an emotion (negative or positive), arousal measure its intensity, while the third dimension determines emotion's controllability nature. There are several available datasets which include the word-level VAD measures as obtained from the labelling performed by a group of participants. The selected dataset for this study is that of [17] which comprises 13,915 English lemmas. However, as the purpose of this work is to determine only the polarity, and not a continuous scale for sentiments, we split the VAD space into 3 subspaces: negative, positive and neutral, by using *k*-means to determine their centroids.

We also aim at going above the word level, and extend the dimensional approach to an utterance level, or even higher. For this purpose, there are several established features which are considered to be representative for a text fragment given its constituent words, such as: the VAD mean; the number of positive and negative words; the maximum arousal of the contained negative words and the maximum arousal of the contained positive words; all of which are included as features for the dimensional classifier presented in this paper. In order to compute them, we take each word from the utterance, and determine its VAD value. Stop words are eliminated by the pre-processing step, while for the words which are not found in the reference dataset, the VAD value is obtained through the use of their WordNet [18] synonyms.

B. Categorical Model

The categorical model classifies the text according to a set of predefined categories, employing statistical or probabilistic models. To perform the classification, the model determines a set of features which best represent their respective categories. The set of features might be so large that the classification would become tedious. Therefore, feature selection or reduction methods are employed. Feature selection consists of scoring each prospective feature by a particular selection metric, and then taking the best *k* features.[2] Chi-Square is one of these metrics, and also the one used in our experiments. To discriminate between the relevant and non-relevant terms, feature weighting metrics are used. In this study we considered the Relevance Factor, $RF_{t,c}$.

Therefore, we compute the Chi-Square (x^2) and Relevance Factor (*RF*) for each word which has at least 100 occurrences within the data[3], using the following formulae:

$$x^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (1)$$

$$RF_{t,c} = log_2(1 + (A+B)) log_2 \frac{A}{\max(1, B)} \quad (2)$$

where *t* is the term, *c* the presumed category, *A* is the number of times *t* and *c* co-occur, B the number of times *t* occurs without *c*, C be the number of times *c* occurs without *t*, D be the number of times neither *t* nor *c* occurs, and N the number of text instances. To determine the category of a term we presumed that it belongs to the category A, this assumption turn out to be true when *AD-CB* is greater than zero, otherwise the term belonged to the other category B. Relevance factor is calculated just after establishing the appertaining category.

The classifier input features for this method are:

- Score per text instance, computed as:

$$\frac{\sum_{i=1}^{n} category(i) * x^2(i,c) \ * \ RF_{i,c}}{\sum_{i=1}^{n} x^2(i,c) \ * \ RF_{i,c}} \quad (3)$$

where *n* is the number of words, and *category(i) = sgn(AD-CB);*

- number of negative words: determined by category;
- number of positive words: determined by category;
- maximum weight of negative words: given by RF_i:
- maximum weight of positive words: given by RF_i .

C. Classifiers

To test the performance of each of the two proposed models, we used three standard classifiers with different classification techniques:

Naïve Bayes is a simple probabilistic classifier suitable for supervised training, which creates a language model based on the conditional independence assumption. Although this conditional independence does not hold for text data, the classification results are more often very good [19].

Support Vector Machines are a discriminative classifier based on the large-margin approach, which search for the optimal hyper plane that generates a maximum distance between classes. Given this hyper plane the support vector

[2] In this work we used the first 37,500 features.

[3] This number was obtained through empirical studies.

machine will consist of the features situated near the surface limit.

C4.5 is a statistical classifier which uses the information entropy concept to generate decision trees. Therefore its learning algorithm works by processing and deciding upon the attributes of the data to be classified.

IV. EVALUATION

A. Datasets

To test the methods, we selected 4 different datasets which contain emotional labels:

Semeval 2007 [20] is an emotional dataset with 1,250 news headlines extracted from major newspapers, and which are annotated according to six emotion classes (anger, disgust, fear, sadness and joy) and valence (scored with a number between -100 and 100).

ISEAR (International Survey on Emotional Antecedents and Reactions) [21] dataset comprises 7,666 sentences collected from 1096 participants with different cultural and intellectual backgrounds who reported experiences and reactions for various emotional states such as anger, disgust, fear, joy, sadness, shame, guilt.

Fairy-tales [22] is a collection of 176 children's fairy tales, from which only the paragraphs relevant for the sentiment analysis task were maintained. Text classification is performed according to the following emotional classes: angry-disgust, fearful, happy, sad and surprised.

The **Movies Reviews** dataset [23] is composed of 50,000 movie reviews, and it is designed for binary sentiment classification. A negative review is one that has a score less than 4, while a positive review has a score higher than 7, on a 10 point scale.

These datasets were created according to different emotional granularities, due to the variety of classifications that can be performed within the sentiment analysis task. The set of experiments comprised in this section have been performed by taking into account the polarity of a text fragment only, which means that our aim was to distinguish between negative and positive categories. Therefore, the datasets were divided into positive and negative as a function of emotional associated labels, meaning that negative emotions, such as anger, disgust, fear, sadness, shame or guilt were attributed to the negative set, while positive ones, such as joy, happy or surprised to the positive set. This separation was possible for the ISEAR, and Fairy-tales datasets. For Semeval 2007, which has two criteria for classification, i.e. emotion classes and valence, we estimated three centroids for the positive, negative and neutral classes, respectively. The resulting neutral being discarded. The Movie Review dataset already contains the negative/positive labels.

Additionally, from the Movie Review dataset we kept only the training set, which contains an equal number of positive and negative reviews. Because the dimensional model is quite computationally expensive, we also created 3 smaller subsets from the testing dataset, based on the maximum number of words from a review.

TABLE I. DATASET SPLIT INTO POSITIVE AND NEGATIVE INSTANCES AND THE AVERAGE NUMBER OF WORDS PER INSTANCE

Dataset	Negative	Positive	Average number of words
Semeval 2007	82	62	6
ISEAR	6578	1094	20
Fairy-tales	648	559	24
Movie Reviews	12500	12500	227
Movie Reviews 1 (max 50 wds/review)	321	352	40
Movie Reviews 2 (max 75 wds/review)	877	1071	54
Movie Reviews 3 (max 100 wds/review)	1548	1738	68

Table I presents the number of positive and negative instances for each of the evaluated datasets, as well as the average number of words within their instances.

B. Experimental Results

The following experiments evaluate the dimensional and categorical approaches on the four selected datasets: Semeval, ISEAR, Fairy-tales and Movie Reviews. The purpose is to classify the text instances or utterances into positive and negative polarity classes. For classification we employed Naïve Bayes, SVM (SMO) and C4.5 (J48) machine learning algorithms available in the Weka[4] (Waikato Environment for Knowledge Analysis) vs. 3.6.9 software. All results are expressed in terms of the classification F-measure using 10-fold cross validation. The part-of-speech tagging and lemmatizing were performed using the Natural Language Toolkit (NLTK).[5]

The classifier input features were calculated according to the results obtained using the affective database, in the case of the dimensional model, and using Chi-Square and Relevance Factor for the categorical one as presented in Section III.B.

As it can be observed from Table I, the ISEAR dataset is an unbalanced one, with more negative instances than positive ones. Therefore in order to obtain valid results, the negative set was divided into parts which are comparable in dimension to the positive set. The final F-measure for this dataset is the arithmetic mean of the individual subset ones.

We start by analysing the dimensional model. Table II shows the F-measure results for each dataset and classifier. It can be noticed that the Semeval 2007 dataset had the best performance, followed by ISEAR and the Fairy-tales. The worst performance was obtained by the Movie Reviews dataset. These variations are due to the genre of the text used in each dataset and the amount of emotional words contained within the instances, with a direct correspondence to the affective database. Semeval appertains to the news domain so that it contains several salient emotional words, being

[4] http://www.cs.waikato.ac.nz/ml/weka/
[5] http://nltk.org/

TABLE II. DIMENSIONAL MODEL RESULTS. F-MEASURE FOR EACH DATASET AND CLASSIFIER. THE BEST PERFORMING CLASSIFIER FOR EACH DATASET IS HIGHLIGHTED IN BOLD FACE.

	Semeval 2007	ISEAR	Fairy-tales	Movie Reviews 1	Movie Reviews 2	Movie Reviews 3
Naïve Bayes	**0.834**	0.742	0.742	**0.725**	0.698	0.690
SVM	0.812	0.725	0.726	0.722	**0.701**	0.687
C 4.5	0.779	**0.747**	**0.749**	0.708	0.689	**0.694**

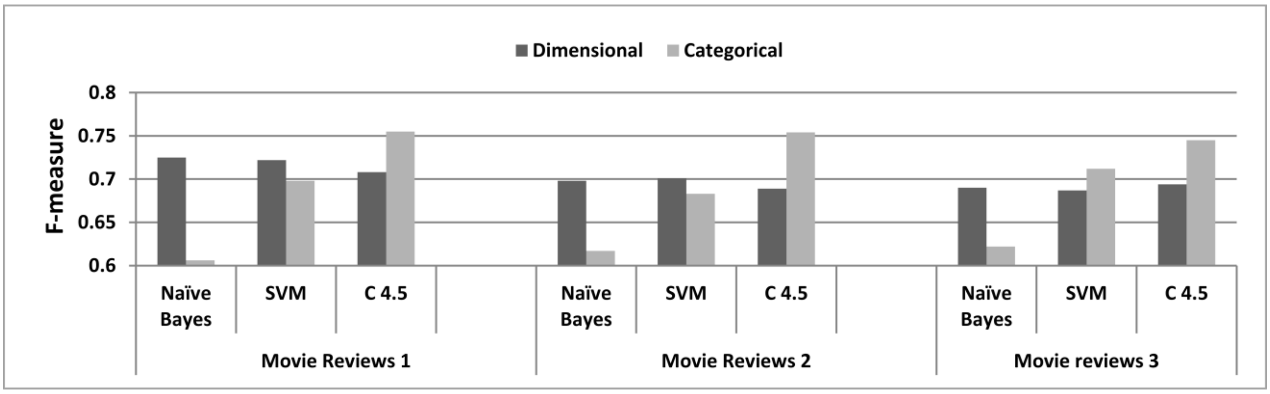

Fig. 2. Comparison between the dimensional and categorical approaches for the Movie Review dataset

constructed in a manner which appeals the reader. ISEAR comprises everyday human experiences, Fairy-tales is a compilation of children tales, while Movie Reviews has specific terms from movie critics, most of which cannot be found in the affective database. The type of classifier with best performances is highly dependent on the dataset. This leads us to believe that there are underlying dataset characteristics exploited differently by each classifier.

For the categorical model, due to the need for an extensive bag-of-words set, we could only test the approach on the Movie Review dataset. The features were extracted from all 25,000 training set movie reviews, and the results were estimated using the smaller datasets. As opposed to the dimensional model, here the training set domain is similar to the testing one, and the results should be above the previous ones (see Table III). However, this is true only for the C4.5 classifier (see Fig. 2). We therefore hypothesise that the conditional independence assumption of the Naïve Bayes classifier does not hold, and that the SVMs cannot determine a correct discriminative hyper plane using so many features. Another aspect worth noticing is that the length of the text instances does not have a major impact upon the classifier results, although a more broad set of review lengths should also be tested.

TABLE III. CATEGORICAL MODEL. F-MEASURE RESULTS FOR THE MOVIE REVIEW DATASETS AND EACH CLASSIFIER. THE BEST PERFORMING CLASSIFIER IS HIGHLIGHTED IN BOLD FACE.

	Movie Rev 1	Movie Rev 2	Movie Rev 3
Naïve Bayes	0.606	0.617	0.622
SVM	0.698	0.683	0.712
C4.5	**0.755**	**0.754**	**0.745**

V. CONCLUSIONS

In this paper, we evaluated two approaches for sentiment polarity classification, one based on the psychologist proposal which represents affective words in a three-dimensional space (valence, arousal and dominance) and a second one which make use of feature selection and weighting metrics to detect a word's polarity. Our evaluation showed that both models performed well, with an average 0.7 F-measure on all datasets, which is comparable to the results of [15]. However, we could not establish direct correlations between datasets and classifier performance, as well as the categorical and dimensional models. This means that there might be some underlying features within the datasets which cannot be captured by these simplified models.

As future work, we consider extending unigram features to higher order n-grams to capture larger contexts and thus performing a more precise classification. Due to the fact that sentiment polarity task requires large-scale annotated data sets, an interesting approach might be to create the features staring from a set of emotional seed-words and using web text data and statistic metrics to create an unsupervised normative dataset, similar to the one used in this work.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287678 (Simple^{4}All).

REFERENCES

[1] James A., Russel, “Core affect and the psychological construction of emotion”, Psychological review, Volume 110, No. 1, January 2003, pp. 145-172.

[2] Ro Valitutti, "WordNet-Affect: an Affective Extension of WordNet", Proceedings of the 4th International Conference on Language Resources and Evaluation, 2004, pp. 1083-1086.

[3] Robert Speer and Catherine Havasi, "Representing General Relational Knowledge in ConceptNet 5", Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), May 2012, pp. 23-25.

[4] Stefano Baccianella, Andrea Esuli and Fabrizio Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), May 2010, pp. 19-21.

[5] V. Hatzivassiloglou, K.R. Mckeown, “Predicting the Semantic Orientation of Adjectives”, Proceedings of the 35^{th} Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, pp. 174-181.

[6] P.D. Turney, M.L. Littman, “Measuring Praise and Criticism: Inference of Semantic Orientation from Association”, ACM Transactions and Information Systems, Volume 21, Issue 4, October 2003, pp. 315-346.

[7] G. Fu, X. Wang, “Chinese Sentence-Level Sentiment Classification based on Fuzzy Sets”, COLING 2010, pp. 312-319.

[8] L. Velikovich, S. Blair-Goldensohn, H. Kerry, Hannan and R.T. Mcdonald, “The Viability of Web-Driven Polarity Lexicons”, The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 777-785.

[9] Hang Cui “Comparative Experiments on Sentiment Classification for Online Product Reviews”, AAAI-2006, pp. 1265-1270.

[10] Philip Clarkson and Ronald Rosenfeld, “Statistical Language Modeling Using The Cmu-Cambridge Toolkit”, 1997, pp. 2707-2710.

[11] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz and Yoram Singer, ”Online Passive-Aggressive Algorithms”, The Journal of Machine Learning Research, Volume 7, 12/1/2006, pp. 551-585.

[12] Kamal Nigam, Matthew Hurst,“Towards a Robust Metric of Opinion”, In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, 2004.

[13] S. Mac Kim, A. Valitutti, R. Calvo, “Evaluation of Unsupervised Emotion Models to Textual Affect Recognition”, Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 62-70.

[14] M. M. Bradley, P. J. Lang, “Affective norms for English words (ANEW): Stimuli, instruction manual, and affective ratings”, (1999).

[15] A. Trilla and F. Alias, “Sentence-based Sentiment Analysis for Expressive Text-To-Speech”, IEEE Transaction on Audio, Speech and Language Processing 2013 February, pp. 223-233.

[16] Z. Wu, H. Meng, Helen, H. Yang, and L. Cai, “Modeling the Expressivity of Input Text Semantics for Chinese Text-to-Speech Synthesis in a Spoken Dialog System”, IEEE Transactions on Audio, Speech and Language Processing, Volume 17 Issue 8, November 2009, pp. 1567-1576.

[17] A. BethWarriner , V. Kuperman ,M. Brysbaert, Marc, “Norms of valence, arousal, and dominance of 13,915 English lemmas”, Journal of Behavior research methods, February 2013, pp. 1-17.

[18] George A. Miller , “WordNet: A Lexical Database for English”, Communications of the ACM, Volume 38, 1995, pp. 39-41

[19] P.D. Turney, “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews”, Proceeding ACL ’02 Proceedings of the 40^{th} Annual Meeting on Association for Computational Linguistics, pp. 417-424.

[20] Carlo Strapparava, Rada Mihalcea, "Learning to identify emotions in text", Proceeding SAC '08 Proceedings of the 2008 ACM symposium on Applied computing, pp. 1556-1560.

[21] H. G. Wallbott, and K. R. Scherer, "How universal and specific is emotional experience?", Social Science Information, Volume 24, 1986, pp.763-795.

[22] Cecilia Ovesdotter Alm, "Characteristics of high agreement affect annotation in text", LAW IV '10 Proceedings of the Fourth Linguistic Annotation Workshop, pp. 118-122.

[23] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A. Ng and C. Potts, “Learning Words Vectors for Sentiment Analysis”, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, June, 2011, pp. 142-150.