Using Sentiment Orientation Features for Mood Classification in Blogs

Fazel KESHTKAR
School of Information Technology and Engineering
University of Ottawa, Ottawa, ON, K1N6N5, Canada
Email: akeshtka@site.uOttawa.ca

Diana INKPEN
School of Information Technology and Engineering
University of Ottawa, Ottawa, ON, K1N6N5, Canada
Email: diana@site.uOttawa.ca

***Abstract*—In this paper we explore the task of mood classification for blog postings. We propose a novel approach that uses the hierarchy of possible moods to achieve better results than a standard machine learning approach. We also show that using sentiment orientation features improves the performance of classification. We used the Livejournal blog corpus as a dataset to train and evaluate our method.**

Keywords:

Sentiment Orientation, Classification, Hierarchy, Mood, Blog.

I. Introduction

Weblog services are free web sites that allow people to write about their opinions and thoughts regarding social, political, and other events. The users can also indicate their emotional state at the time of writing by choosing a mood label. One of the most popular weblog services which provides the above features is Livejournal.Livejournal provides 132 moods in a drop box list. Bloggers are able to indicate in what mood they were when they post a text or a blog. So, they are labeling their posts based on a set of possible moods. Our paper contributes to introduce modular, efficient, and an easy-to-implement hierarchical classification method in cooperation with Sentiment Orientation features and machine learning techniques.

Recently, a lot of research and progress have been done in opinion and sentiment analysis [1]. Research on emotion and mood detection is just starting. Holzman [2] and Rubin et al. [3] investigated emotion detection, but on very small data sets. Automatic classification of blog texts by mood is a challenging task. While authorship attribution and classification of texts by gender were shown to work well on long documents, mood classification needs to work on short documents.

A few researchers studied mood classification based in blogs. Mishne [4] collected a corpus of blog data annotated with mood labels, and implemented a Support Vector Machine (SVM) classifier. He used features such as frequency counts, lengths, sentiment orientations, emphasized words, and special symbols.

This was the first attempt at using surface-level features, and the classification accuracy was only a bit above the baseline. Therefore, we need to come up with methods to increase the accuracy of the classification. Mishne used only top 40 moods as classes, while we use all the 132 classes and their hierarchical organization provided by Livejournal service.

Another mood classification system was proposed by Jung [5], using some common-sense knowledge from ConceptNet [6], and a list of affective words [7], and treating only four moods: *happy, sad, angry, scared*.

In this paper, we introduce a hierarchical approach to mood classification. We address the variety of possible mood labels; our approach is flexible in the sense that we can change the set of moods (classes into which we classify the blogs). Hence, our effort is devoted to the development of a novel method for mood classification where the previous methods might not work. The main contributions of the current work is to devise an accurate and efficient hierarchical algorithm for mood classification using Sentiment Orientation features. This approach is different from the existing machine learning methods, i.e, Mishne [4] and Jung et al. [5] in that they use the direct flat classification.

Our results show that, although the blog corpus contains a huge amount of words in various domains, we can choose features that lead to accurate classification of the expressed mood. We start with all the words as features (most of the words in the training data), then we use feature selection techniques to reduce the feature space. We add sentiment orientation features and emoticons.

The rest of this paper is organized as follows. In Section II we take a look at the blog corpus that we consider for our research. In Section III we introduced our Hierarchy Classification approach. Section IV follows with details regarding the features that we used for the classification process, dividing them into sets of related features. Our experiments and results (we compare our work to existing work in affect analysis and related fields in VI) are reported in Section V. We conclude in Section VII.

II. Data Set

We used the blog data set that Mishne collected for his research [4]. The corpus contains 815,494 blog posts from Livejournal, a free weblog service used by millions of people to create weblogs. In Livejournal, users are able to optionally specify their ”current mood”. To select their mood users can

978-1-4244-4538-7/09/$25.00 2009 IEEE

Status	Counts
Number of Standard Moods	132
Number of User-defined Moods	54,487
Total Words	69,149,217
Average-words/Post	200
Unique Words	596,638
Individual pages	122,624
Total Weblogs	37,009
Total Posts	815,494

TABLE I
STATISTICS ABOUT WORDS AND POSTS IN THE DATA SET.

Mood	Occurrences	Mood	Occurrences
amused	24857 (4.0%)	contemplative	10724 (1.7%)
anxious	7052 (1.1%)	tired	20299 (3.2%)
awake	10121 (1.6%)	exhausted	6943 (1.1%)
happy	16471 (2.6%)	calm	10052 (1.6%)
crazy	6433 (1.0%)	cheerful	12979 (2.1%)
bouncy	10040 (1.6%)	depressed	6386 (1.0%)
bored	12757 (2.0%)	chipper	9538 (1.5%)
curious	6330 (1.0%)	accomplished	12200 (1.9%)
annoyed	8277 (1.3%)	drained	6260 (1.0%)
sleepy	11565 (1.8%)	confused	8160 (1.3%)
sad	6128 (1.0%)	content	11180 (1.8%)
busy	7956 (1.3%)	aggravated	5967 (1.0%)
excited	11099 (1.8%)	sick	7848 (1.3%)
ecstatic	5965 (1.0%)	working	2775 (0.55%)
angry	2832 (.60%)	confused	7724 (1.20%)
frustrated	4132 (.90%)	hyper	2678(0.52%)
blank	5441 (0.98%)	thoughtful	4295 (0.91%)
annoyed	7248 (1.15%)	loved	3883 (0.65%)
blah	10127 (1.78%)	hopeful	5059 (0.95%)
cranky	3945 (.85%)	comntemplative	10159 (1.76%)

TABLE II
THE MOST FREQUENT MOODS IN CORPUS [4].

choose from a list of 132 moods, or specify additional moods. We do not use these additional moods because very few posts are annotated with them. Some statistics of this corpus are shown in Table I. From the total posts, only 77% included an indication of the mood; we disregard the rest. There are 22 posts per blog on average.

Table II shows the most frequent moods in the corpus, based on their frequency counts.

III. OUR HIERARCHICAL APPROACH TO MOOD CLASSIFICATION

Researchers have been developing new techniques to classify large amounts of documents and texts such as directories or biomedical data, especially when the number of classes is large and the classes have hierarchical structure [8].

Hierarchical text categorization removes the assumption that categories are independent and tries to incorporate the inter-relationship among the different categories. In a regular or flat text categorization approach, it is generally assumed that categories are independent and non-overlapping. This means that classification of individual category is performed with the knowledge that all other categories are considered as unrelated. In practice, however, categories tend to overlap and not be independent of each other [9].

Wang et al. [9] introduced two types of hierarchical text categorization methods: *global* and *local*. Below we explain these techniques, then we show how we applied hierarchical categorization techniques in the mood classification task.

- ***Global Approach:*** In the global approach only one classifier discriminates all categories in a hierarchy. It differs from a flat categorization by considering the relations between the categories (using a similarity measure).
- ***Local Approach:*** In the hierarchical approach, a classifier is built for each internal node in the hierarchy. It proceeds in a top-down fashion, first picking the categories of the first level and then recursively making the choice among the lower-level categories (the children of the higher-level categories). There are two possibilities here. One is to train one classifier for the all the categories in the first level of the hierarchy (multi-class problem). Then one classifier is trained for each of the classes in the first level to differentiate among all its sub-classes. The process continues if there are more levels. The second possibility is to build a single-class classifier for each node in each level of the hierarchy (a binary classifier that differentiates the class of the node from any other sibling classes). We used the first approach, with SVM classifiers, because it is more efficient.

In the Livejournal weblog service, the moods are organized in a hierarchy, shown in Figure 1.

IV. FEATURE SET

We briefly explain the features that we used in our machine learning experiments. We have used most of features from Mishne [4], plus some additional sentiment orientation features, such as tagged words from the General Inquirer [10]. We assigned $+1$ to positive and -1 to negative words, and calculated the sum and the average score of each document for each of the three lists of words mentioned bellow for the third type of features.

A. Frequency Counts

Bag-of-Words (BoW) is the most common feature representation used in automatic text classification. We represent the words by their frequencies.

B. Length-related Features

Since, blog posts vary in length, we consider length features such as: the length of the document, the number of sentences, and the average number of words.

C. Sentiment Orientation

For mood classification, the sentiment orientation of some words can be a useful feature. Several sources are predictors for sentiment orientation. We calculate the total and the average orientation score for each document based on the words that are from the following resources:

- A list of 2,291 positive words and 1,915 negative words from the General Inquirer [10].

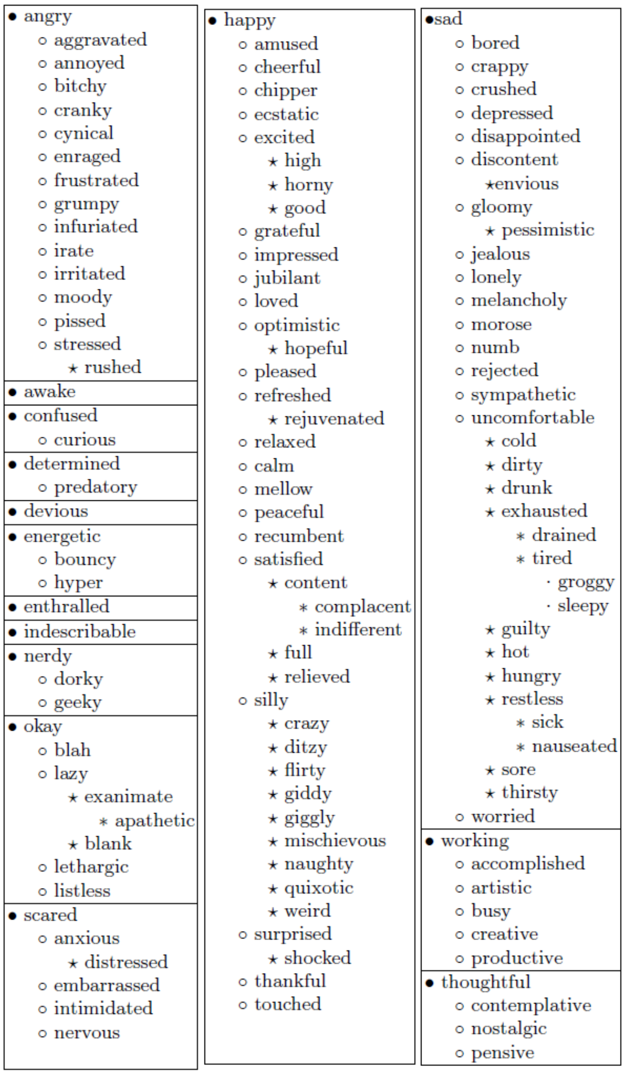

Fig. 1. The hierarchy of the 132 moods; •: level1, ∘: level2, ⋆: level3, ∗: level4, and · : level5 .

- A list of 21,885 verbs and nouns that were assigned a positive, negative, or neutral orientation score, Kim-Hovy list [11].
- A list of 1,718 adjectives with their scores of polarity values, constructed by using the method of Turney and Littman [12].

D. Special Symbols

We used special symbols called emoticons (emotional icons), that represent human emotions or attitudes. These symbols are textual representations of facial expressions, i.e. :) (smile) and ;) (wink) and so on. We used 9 most popular emoticons as features.

V. Experiments and Result

The purpose of this section is to evaluate the performance of the proposed method using a training and a testing dataset. We compare the performance of the simple flat classification into

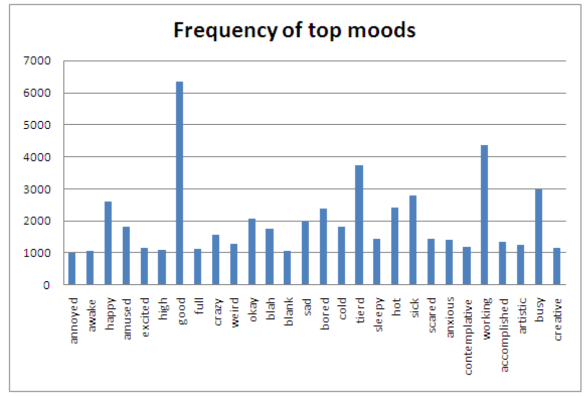

Fig. 2. Frequency of the top moods in the experimental data set.

132 moods to the performance of our Hierarchy-based Mood Classification. We explain the classification setting, then the experiments and the results will follow.

A. Classification Setting

Weka [13], a Data Mining Software was used for our experiments. We chose Support Vector Machine (SVM) as the classifier, because it was shown to perform very well in text classification tasks, including the previous studies on mood classification. Moreover, SVMs are able to deal with large amounts of features and instances [14]. We experimented with a few other classifiers, such as Nave Bayes and Decision Trees; the results for our task were lower.

Our feature space is very large. Due to efficiency reasons, we reduced the feature space for the *BoW* features only by using feature selection methods. We use the *Chi-Square* feature selection method to keep only the first 5000 features from the 43,109 BoW features, plus the six sentiment orientation scores (for each of the three resources, the average score of the words from the text that are also in the resource, and their total score) and the emoticons.

B. Experiments

Our training data set for the experiments includes 144,129 instances. They vary among all 132 different moods. From all the posts, we randomly selected 144,129 as training data and 90,000 as test data. Two sets of experimented are presented. We randomly selected the data set for testing among the 90,000 instances. Figure 2 shows the top 29 moods in our experimental data set. To achieve some preliminary estimation and to compare to our hierarchical approach, our first experiment was to use SVM for a flat classification into the 132 moods.

The second set of experiments is performed to evaluate the hierarchies classification method. For this purpose, we first trained a classifier to classify into the 15 categories from the first level of hierarchy: *happy, sad, angry, okay, working, scared, awake, thoughtful, nerdy, indescribable, enthralled, determined, confuse, devious*, and *energetic*.

In the next step, for each node from the first level of hierarchy we extracted the related instances and their mood labels. For instance, for the node *angry* we selected all the documents that have the label *angry, aggravated, annoyed, bitchy, cranky, cynical, enraged, frustrated, grumpy, infuriated, irate, irritated, moody, pissed*, and *stressed*. Finally, we ran the classifier for each node in the second level. We repeated this procedure for each of the 15 categories from the first level of the hierarchy. We continue similar steps for the third, fourth and fifth level of the hierarchy. For both sets of experiments mentioned above, we ran our classifiers using Bag of Words (BoW) features and using BoW plus Semantic Orientation (SO) features (including emoticons). We will see that adding the SO features greatly improves the classification results.

C. Results and Discussion

The first experiment was a flat classification into 132 moods. The results of this experiment was an accuracy of 24.73% for *BoW+SO* and 18.29% for *BoW*. This is an improvement compared to a baseline accuracy of 6.32% when always choosing the most frequent mood (*happy*). Just to clarify that baseline is the lower expectation in each level for a naive classifier that always chooses the most frequent class.

For the second experiment the hierarchical approach, for the classifier that classifies into one of the 15 moods from the first level, the accuracy was 63.5% for *BoW+SO* and almost 40% for *BoW*, compared to a baseline of 15%; the results for Level1 is illustrated in Table III.

	Baseline	BoW	BoW+SO
Level1	15%	40%	63.50%

TABLE III
ACCURACY FOR THE HIERARCHICAL CLASSIFICATION IN LEVEL 1 FOR BOTH BOW AND BOW+SO FEATURES.

In the next step, we have 15 classifiers in the second level, one for each node in the first level. In fact we have only 11 classifiers, because 4 moods did not have any children branches in the hierarchy, so for them the classification is already finished.the average accuracy was 92.33% for BoW+SO features, 83.30% for BoW features only, and 32.70% for a baseline of fhe most frequent class. The difference between the hierarchial approach with all the features and the baseline is 59.63%. There are several branches that have fewer children and show larger improvement; and there are several branches with many children that show lower performance improvement. For example the moods *happy*, *sad*, and *angry* have many children branches and the improvement is smaller. The gain in performance is bigger for moods such as *nerdy*, which has two branches. Two branches means three classes, in this case generic *nerdy* and more specific kinds of *nerdy*: *geeky* and *dorky*.

The results of the level 3 classifier are shown in Table V. The results of Level 4 are shown in Table VI. Level 5 has only one classifier, for *tired*, with an accuracy of 96.22% for *BoW+SO* features and 87.61% for *BoW*, with a baseline of 54.44%. Our experiments and results clearly show that the hierarchical classification leads to strong performance and it is well-suited for the task. The summary of results that shown in Table VII clearly support above arguments.

Level2	Baseline	BoW	BoW+SO
happy	8.64%	62.72%	86.97%
sad	10.38%	66.89%	86.88%
angry	11.67%	80.13%	91.90%
okay	24.55%	78.67%	82.25%
working	25.24%	87.74%	93.29%
scared	25.97%	89.21%	95.02%
thoughtful	35.99%	91.32%	94.84%
nerdy	41.40%	90.65%	97.68%
determined	65.52%	93.25%	95.83%
confused	56.32%	85.71%	94.33%
energetic	54.05%	90.05%	96.73%
Average	32.70%	83.30%	92.33%

TABLE IV
ACCURACY FOR CLASSIFICATION IN LEVEL 2 FOR BOTH BOW AND BOW+SO FEATURES.

To directly compare the results of the flat categorization to results of the hierarchical classifiers, we can cumulate the errors from all the levels. This will give a global accuracy of 55.24% for all 132 classes (for Bow+SO), significantly higher than 19.28% for the flat categorization.As illustrated in Table VII, the improvement in performance between the flat and the hierarchy classification is significant, especially when adding the sentiment orientation features.

Level3	Baseline	BoW	BoW+SO
uncomfortable	17.97%	71.03%	90.72%
surprised	56.18%	96.19%	97.68%
stressed	67.62%	94.58%	98.72%
silly	14.17%	74.53%	90.32%
satisfied	31.97%	80.15%	96.44%
refreshed	52.92%	95.55%	97.06%
optimistic	66.41%	90.35%	98.80%
lazy	31.37%	87.40%	96.88%
gloomy	66.21%	93.68%	99.88%
excited	35.87%	86.33%	91.75%
discontent	84.27%	92.08%	100%
anxious	58.88%	89.91%	96.85%
Average	48.65%	87.64%	95.84%

TABLE V
ACCURACY FOR CLASSIFICATION IN LEVEL 3 FOR BOTH BOW AND BOW+SO FEATURES.

Level4	Baseline	BoW	BoW+SO
content	54.23%	89.73%	97.92%
restless	48.50%	90.27%	97.70%
exhausted	40.20%	89.17%	96.09%
exanimate	68.73%	96.46%	100%
Average	52.16%	91.40%	97.93%

TABLE VI
ACCURACY FOR THE HIERARCHICAL CLASSIFICATION IN LEVEL 4 FOR BOTH BOW AND BOW+SO FEATURES.

VI. COMPARISON TO RELATED WORK

Standard research on text classification tries to detect the topic of documents. Less research has been done to detect

Summary of the Results	
Hierarchical Classification BoW+SO	55.24%
Hierarchical Classification BoW	23.65%
Flat Classification BoW+SO	24.73%
Flat Classification BoW	18.29%
Baseline	7.00%

TABLE VII
THE RESULTS OF THE HIERARCHICAL CLASSIFICATION WHEN THE CLASSIFIERS FROM ALL THE LEVELS ARE APPLIED SUCCESSIVELY (THE ERRORS FROM ALL THE LEVELS ARE MULTIPLIED), COMPARED TO THE RESULTS OF THE FLAT CLASSIFICATION, FOR BOTH BOW AND BOW+SO FEATURES.

specific features in the text [15], such as gender detection of the writer [16] or authorship attribution [17].

In this paper, we focus on mood classification; the task is difficult because of the distinctive aspects of mood in blogs. For instance, a blogger can start in a certain mood and the document can ends with other content or mood. Some blogs are so intertwined that even human readers would have difficulty in identifying the mood, and finding the relation between the mood of the writer and the documents. As we mentioned in the introduction, we found two studies that started similar investigations.

Mishne [4] introduced an approach that used Support Vector Machine (SVM) to detect mood classification of blog data.

Mishne tried to answer some questions such as:

- How different is mood classification in blogs and other classification domain? Which types of features used in other classification tasks can be re-used in mood classification?
- How much data is required to have reasonable results and how many features are need to improve the results?

Mishne [4] used only the top 40 moods, shown in Table II. In his experiment he classified each of top 40 moods on different training sets. He showed the performance of each mood, based on different size of the training set. To compare with Mishne's results, we classified into the 40 most-frequent moods which showed in II and we obtained 84.89% accuracy, while Mishne obtained the best accuracy of 67%. He used a test set, randomly chosen and with a balanced distribution of classes; therefore we are not able to use exactly the same test set to compare our results to his result directly, but our data set is randomly chosen from the same data. Moreover, in our work, we consider all the 132 moods which makes the task more difficult. For such a large number of classes, we were able to obtained good accuracy only when using the hierarchical classification model.

In summary, the differences between our work and Mishne's work consist in the fact that we used all the 132 moods, not only the 40 most frequent moods, and in the fact that we enhanced the feature set with more sentiment orientation features. Moreover, we use the hierarchical classification in order to improve the results.

Another research that investigate mood classification was done by Jung et al. [5]. They proposed a hybrid approach to identify mood in blog texts. For their research they used ConceptNet [6] and the Affective Norms English Words (ANEW) [7] list. In order to do classification they chose some unique features from blogs and calculate statistical features such as term frequency, n-grams, and point-wise mutual information (PMI) for the SVM classification method. They also used a paragraph-level segmentation based on mood flow analysis. They extracted their own blog data from LiveJournal. They show good classification results only when restricting the task to four mood types: *happy, sad, angry,* and *scared*. Also, they considered only documents with the length between 4 and 40.

Liu et al. [18] introduced another approach that works on a larger corpus and it is based on "common-sense" classification. The above common-sense classification used a corpus of 400,000 facts about the everyday life in the real world. They used the combination of four linguistic models in order to recognize the affect. The system analyzed affective qualities of text, sentence by sentence. This has a practical value when people want to evaluate the text they are writing.

VII. CONCLUSIONS AND FUTURE WORKS

This paper presents experiments for a hierarchy-based mood classification of blog corpus. In addition, a hierarchical approached was considered to classify the data using the SVM algorithm. Our corpus was collected from Livejournal, an online service that allows users to post their personal thoughts. Our results showed that the hierarchical approach leads to a substantial performance improvement over a flat classification. Classifying mood in blog text is a difficult task due to the variety of users, but our hierarchy approach shows that if we classify the blogs using the mood hierarchy, we can achieve very good performance. We also showed that using SO features on top of Bow features greatly improves the classification results.

In future work we plan to experiment with more features sets, types of features and features selection methods. We also plan to test other hierarchical classification approaches, such as binary classifiers for each node (one class versus all the other classes from the same level), in order to compare the performance and the efficiency in term of running time.

ACKNOWLEDGMENTS

The authors wish to thank Soo-Min Kim and Ed Hovy for providing their polarity-tagged lists.

REFERENCES

[1] B. Pang and L. Lee, "Opinion minding and sentiment analysis," *Foundation and Trend in Information Retrieval*, vol. 2:1-2, 2008.

[2] L. Holzman and W. Pottenger, "Classification of emotions in internet chat: An application of machine learning using speech phonemes," *Technical Report LU-CSE-03-002, Lehigh University*, 2003.

[3] V. Rubin, J. Stanton, and E. Liddy, "Discerning emotions in texts," *AAAI-EAAT*, 2004.

[4] G. Mishne, "Experiments with mood classification in blog posts," *ACM SIGIR*, 2005.

[5] Y. Jung, H. Park, and S. Myaeng, "A hybrid mood classification approach for blog text," *LNCS*, pp. 1099–1103, 2006.

[6] H. Liu and P. Singh, "Conceptnet a practical commonsense reasoning tool-kit," *BT Technology Journal*, pp. 211–226, 2004.
[7] M. Bradley and P. Lang, "Affective norms for english words (anew)," *University of Florida*, 1999.
[8] O. Deker, J. Keshet, and Y. Singer, "Large margin hierarchical classification," in *Proceedings of the 21st International Conference on Machine Learning*, 2004, p. 27.
[9] K. Wang, S. Zhou, and S. Liew, "Building hierarchical classifiers using class proximities," *Proceedings of VLDB*, pp. 363–374, 1999.
[10] P. J. Stone, D. Dexter, C. Smith, S. Marshall, and D. M. Ogilvie, "The general inquirer: A computer approach to content analysis," *MIT Press*, 2007.
[11] S.-M. Kim and E. Hovy., "Determining the sentiment of opinions," *Proceedings of the 20th COLING*, pp. 1367–1373, 2004.
[12] P. Turney and M. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM (TOIS)*, vol. 21(4), 2003.
[13] I. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
[14] T. Joachims, "Text categorization with suport vector machines: Learning with many relevant features," *ECML, UK*, pp. 137–142, 1998.
[15] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34(1), p. 147, 2002.
[16] M. Koppel, A. Argamon, and A. Shimoni, "Automatically categorizing written texts by author gender," *Literary and Linguistic Computing*, vol. 17(4), pp. 401–412, 2002.
[17] M. Koppel and J. Schler, "Authorship verification as a one-class classification problem," *ICML, Twenty-first international conference on Machine learning*, 2004.
[18] H. Liu, H. Lieberman, and H. Selker, "A model of textual affect sensing using real-world knowledge," *IUI, Proceedings of the 8th international conference on Intelligent user interfaces*, pp. 125–132, 2003.