A Hybrid Mood Classification Approach for Blog Text

Yuchul Jung, Hogun Park, and Sung Hyon Myaeng*

School of Engineering, Information and Communications University, South Korea
119, Munjiro, Yuseong-gu, Daejeon, 305-732, Korea
{enthusia77, gsgphg, myaeng}@icu.ac.kr

Abstract. As an effort to detect the mood of a blog, regardless of the length and writing style, we propose a hybrid approach to detecting blog text's mood, which incorporates commonsense knowledge obtained from the general public (ConceptNet) and the Affective Norms English Words (ANEW) list. Our approach picks up blog text's unique features and compute simple statistics such as term frequency, n-gram, and point-wise mutual information (PMI) for the SVM classification method. In addition, to catch mood transitions in a given blog text, we developed a paragraph-level segmentation based on a mood flow analysis using a revised version of the GuessMood operation of ConceptNet and an ANEW-based affective sensing module. For evaluation, a mood corpus comprised of real blog texts has been built semi-automatically. Our experiments using the corpus show meaningful results for 4 mood types: happy, sad, angry, and fear.

1 Introduction

A blog is a web site, where anybody can write about his or her own expriences and thoughts on a voluntary basis. As a result, it reflects user's personality and cultural biases, sometimes forming a unique society. Since blog texts often carry the emotions of the writers, they should lend themselves for automatic categorization based on moods. Compared to topicality-based classification of text, mood classification is challenging in many aspects.

A recent approach to mood classification of text uses Support Vector Machine (SVM) with 6 features: frequency counts, lengths, semantic orientations, Point-wise Mutual Information for Information Retrieval (PMI-IR), emphasized words, and special symbols [1]. While this approach of using surface level features can allow for reasonable accuracy, it seems to have a limit because its inability to deal with idiosyncratic aspects of moods and blogs. For example, although an author is under a certain mood when starting to write a blog document, the initial mood may not be maintained all the way to the end. Some blogs are so intertwined that even human readers would have difficulty in identifying the mood, not to mention the statistically motivated method using surface level features.

To detect the mood of blog text more accurately, we propose a hybrid approach to mood classification that incorporates commonsense knowledge obtained from the

* Corresponding author.

Q. Yang and G. Webb (Eds.): PRICAI 2006, LNAI 4099, pp. 1099 – 1103, 2006.

general public (ConceptNet) [2] and the Affective Norms English Words (ANEW) list [3]. ConceptNet is an easily usable, freely available commonsense knowledge base and natural language processing toolkit which support many practical textual-reasoning tasks over real-world documents, including topic-gisting, affect-sensing, analogy-making, and other context-oriented inferences. The knowledge base is a semantic network presently consisting of over 1.6 million assertions of commonsense knowledge covering the spatial, physical, social, temporal, and psychological aspects of everyday life.

The ANEW list, created from a psychological study, contains 1,034 unique terms with affective valence (unpleasant ~ pleasant), arousal (calm ~ excited), and dominance (submissive ~ dominance) scores. It can be used to identify different mood types based on lexical analysis by mapping terms in text to those in the list.

Our approach is hybrid in the sense that several tools are integrated: the SVM classification model [4] that has shown superior performance over other existing classification models in many application domains, the GuessMood function of ConceptNet [2], and an affective sensing model based on ANEW [3], and Open Mind Common Sense (OMCS)[1][5]. We observed that some features like term frequency, PMI-IR, emoticons, abbreviated words, and mood-specific terms contribute to detection of the mood of a given text. In addition, a paragraph level segmentation and a mood flow analysis were applied to handle various blog texts of different lengths and writing style.

For an evaluation of our hybrid mood classification approach, we have built a mood corpus based on a large number of blog documents extracted from Live Journal.com. More than 50GB text has been processed to semi-automatically classify the documents into four categories: *happy, sad, angry*, and *fear*.

2 Proposed System

Our system includes two steps as in Fig. 1. In the first step, when a blog document comes in, the system performs statistical analyses to obtain term frequency, n-gram, and PMI-IR[7] sequentially. Based on the statistical features, the system applies SVM[2] based mood classification to assign a mood category to the document.

In the second step, the system initiates a mood flow analysis to identify a global mood for the given blog document. As in the details found in Fig. 2, our mood flow analyzer segments a blog document into paragraphs. After that, in paragraph analysis, the number of mood terms is counted to select a scheme between a revised ConceptNet's GuessMood [2] and a PAD [3] based affective sensing module. If the number of mood terms is bigger than the experimental threshold, the latter is chosen. If a mood is sustained without transitions throughout the whole blog document, the "final resolver" module only checks the consistency and assigns the mood as the final result. When some paragraphs have different mood types, heuristically measured weights are multiplied into the results of paragraph analysis. A global mood score is

[1] http://commonsense.media.mit.edu/cgi-bin/search.cgi
[2] http://svmlight.joachims.org/

calculated by averaging the weighted sum of every paragraph analysis in the final resolution phase. The heuristic weights were obtained through several hundreds of trials with our training corpus (Fig. 2).

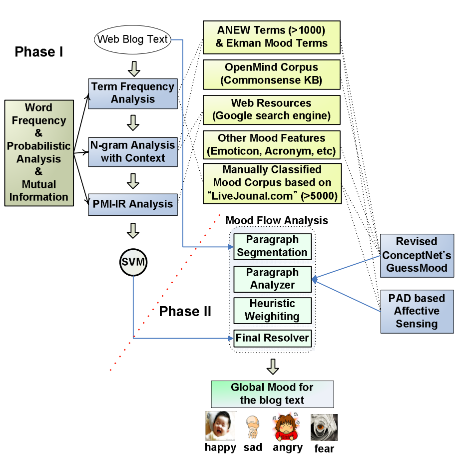

Fig. 1. Overall Flow

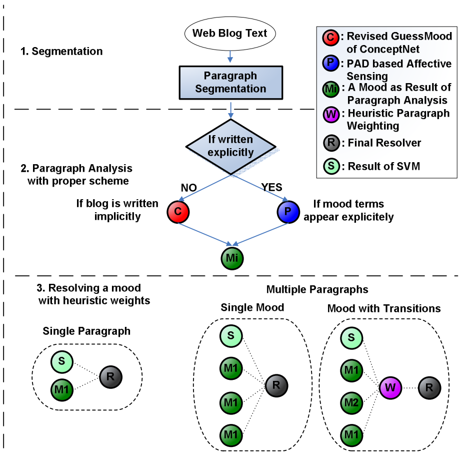

Fig. 2. Details of Mood Flow Analysis

3 A Blog Mood Corpus

We have collected over 50GB of blog text from "LiveJournal.com" semi-automatically, which are appropriate for our mood categories (happy, sad, angry, and fear) and the ANEW list to build a trustable mood corpus that can be used for training/testing purposes. In order to select blogs which fall into our mood categories, we adopted LiveJournal's mood hierarchy[3]. In this process, 108,892 blog documents were extracted from 50G blog text. Because the authors' mood "annotation" sometimes doesn't match the real content, we performed a specially designed refinement process for more trustable mood corpus. It uses a keyword spotting technique to remove non-affective parts of the text, which contain few or no mood key terms that were predefined. In addition, to avoid irregular, meaningless, and ambiguously long blog documents, it ignores blog texts whose length is less than 5 or more than 40 sentences. Finally, now we have 10,479 blog documents (about 10MB) as a highly refined mood corpus.

4 Experiments and Discussion

For improved accuracy of mood classification, diverse features of blog documents were considered. In addition, each of the two methods, one with GuessMood of ConceptNet and the other with the PAD based affective text sensing model. Internal

[3] *http://www.livejournal.com/moodlist.bml?moodtheme=140&mode=tree*

parameters were tuned based on experimental work. The main goal of the experiments was to evaluate the hybrid mood classification model. In case of the *happy* and *sad* mood categories, classification results approached to the range of 89~92%. Table 2 shows classification accuracy comparisons between the baseline SVM and our proposed approaches. A total of randomly chosen 20,000 blog documents (happy: 5,000, sad: 5,000, angry: 5,000, and fear: 5,000) and semi-automatically refined 5,000 blog documents (happy: 2000, sad: 1800, angry: 600, fear: 600) were used for training and testing, respectively. 5-fold cross validation was taken for SVM classifier's evaluation. In Table 2, "random" and "refined" mean the corpus with 20,000 documents and that with 5,000 documents, respectively.

Table 2. Mood classification: SVM vs. Our Approach

Testing Type	Category	SVM	Our Approach
Type 1. Training: Random + Testing: Random	Happy	59.30	55.14 (-4.16)
	Sad	35.60	31.94 (-3.66)
	Angry	44.80	38.28 (-6.52)
	Fear	42.10	35.12 (-6.98)
	Average	45.45%	43.24%
Type 2. Training: Refined + Testing: Random	Happy	77.80	76.79 (+9.99)
	Sad	43.15	38.37 (-4.78)
	Angry	26.67	23.25 (-3.42)
	Fear	27.23	25.40 (-1.83)
	Average	43.71%	40.95%
Type 3. Training: Refined + Testing: Refined	Happy	90.13	92.85 (+2.72)
	Sad	89.58	89.27 (-0.31)
	Angry	80.77	83.12 (+2.35)
	Fear	57.68	61.97 (+4.29)
	Average	79.54%	81.80%

Although the semantic network of ConcpetNet consists of 1.6million assertions, it contains lots of needless commonsense knowledge that is not required for processing mood related concepts. Thus, we have re-organized the semantic network by filtering out unnecessary concepts, forming the "refined" corpus.

SVM: Randomly selected training data were not sufficient in constructing a classifier; with the limited coverage, its performance was only 45.45% on average. When the SVM classifier was tested with randomly selected testing data, which contains inconsistent lexical features, it almost failed in getting a reasonable level of accuracy except for the *happy* mood. In case of Type 2, although highly refined training data were used as the training corpus, classification result became worse due to heterogeneity between the training and testing corpora. However, highly refined training data helped in achieving average 79.54% of accuracy if well refined testing data were also used.

Our Hybrid Approach: In every testing, classification accuracy for the *happy* mood was enhanced when the refined corpus was used in training. However, in case of Type 1&2, when randomly chosen testing data were used, there were ups and downs. On the other hand, our hybrid approach obtained 81.80% of accuracy on average when well refined training and testing data were used. While the revised GuessMood and the PAD based affective text sensing modules caused noisy classification results because the semantic network is not comprehensive enough to cover all the mood-related terms that appear in the blog text, the experimental result with the refined data indicates the resulting is promising. Even the relatively small size of training corpus of refined data allowed for quite reasonable performance.

5 Conclusion and Future Work

This paper presents a hybrid model for mood classification of blog text, which uses statistical features, an informal commonsense reasoning with ConceptNet, and a PAD based affective text sensing method. In addition, a semi-automatically refined mood corpus has been built and used to evaluate our proposed model.

Mood classification of blog documents is a very difficult task because of diverse situations and expressions of authors. Although we can hardly catch author's internal, emotional status correctly, at least we could perceive a global mood for a given blog text at the surface level if statistical features and commonsense knowledge are incorporated. With the experiments using the blog text, we have developed a firm belief that sophisticated preprocessing and a specially designed hybrid mood classifier are quite feasible for mood classification of blog text.

References

1. Mishne, G.: Experiments with Mood Classification in Blog Posts. *Style2005 – the 1st Workhops on Stylistic Analysis of Text for Information Access, at ACM SIGIR 2005*, (2005)
2. Liu, H. and Singh, P.: ConceptNet – A practical commonsense reasoning tool-kit. BT Technology Journal, (2004), 211-226.
3. Bradley, M.M., and Lang, P.J.: Affective norms for English words (ANEW). Gainesville, FL. The NIMH Center for the Study of Emotion and Attention, University of Florida, (1999).
4. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proc. of ECML-98, 10th European Conf. on Machine Learning*, Springer-Verlag, (1998), 137-142
5. Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., and Li Zhu, W.: Open Mind Common Sense: Knowledge acquisition from the general public. *Proc. of the First Int. Conf. on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems*, (2002).
6. Liu, H., Lieberman, H., and Selker.: A Model of Textual Affect Sensing using Real-World Knowledge. *Proc. of the 2003 Int. Conf. on Intelligent User Interfaces, IUI 2003*, (2003).
7. Read, J.: Recognising affect in text using pointwise-mutual information. Master's thesis, University of Sussex, (2004).
8. Mehrabian, A.: Framework for a comprehensive description and measurement of emotional states. Genetic, Soccial, and General Psychology Monographs, 121(3), (1995), 339-361.