

What Emotions do News Articles Trigger in Their Readers?

Kevin Hsin-Yih Lin, Changhua Yang and Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan

{hylin, chyang}@nlg.csie.ntu.edu.tw; hhchen@csie.ntu.edu.tw

ABSTRACT

We study the classification of news articles into emotions they invoke in their readers. Our work differs from previous studies, which focused on the classification of documents into their authors' emotions instead of the readers'. We use various combinations of feature sets to find the best combination for identifying the emotional influences of news articles on readers.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *Clustering*

General Terms

Algorithms, Performance, Experimentation

Keywords

News Articles, Reader-Emotion Classification, Text Classification

1. INTRODUCTION

Past researches on emotions conveyed by documents focused on detecting the feelings that the authors of the documents were expressing [1]. Such studies are useful when we need a quick poll on how people feel about a particular event or item.

Instead of uncovering the emotional states of the authors of documents, this paper aims to find out what emotions documents trigger in their readers. Such research has novel applications. Suppose someone who likes pets is in a gloomy mood. He or she will definitely be cheered up by reading documents that are both about pets and heartwarming stories. Existing information retrieval (IR) systems can deal with the former. That is, IR systems are quite capable of finding documents which contain information that users want to know about. In our example, they are the documents about pets. But current IR systems do not deal with the latter. They cannot discern a shocking story about pet abuse from a pleasant story about the cute things that pets do. Users can add positive words like "heartwarming" to a query, and request IR systems to exclude any documents containing negative words like "abuse". But instead of requiring users to formulate a complex query, it is more convenient for them to simply click on a checkbox representing an emotion to tell IR systems to return documents that are both content and emotion relevant, e.g., about pets and heartwarming. A method capable of identifying the emotional effect of a document on its readers can serve as a filter to retain only the documents that cause desired emotions.

2. CORPUS

We used Yahoo!'s Chinese news articles as our corpus (<http://tw.news.yahoo.com>). Yahoo!'s Chinese news webpage has a special feature which allows a reader of a news article to express how he or she feels after reading the news article. The eight choices that a reader may select from to describe feeling are *happy*, *angry*, *sad*, *surprised*, *heartwarming*, *awesome*, *bored* and *useful*. It is strange for the word *useful* to appear as one of the choices, because *useful* is not an emotion. To deal with this anomaly, we did two separate sets of experiments: one including *useful* instances and one excluding *useful* instances from the corpus.

Our corpus contains news articles spanning a period of 81 days. Articles that were published during the first 54 days and those that were published in the last 27 days were used as training and testing data, respectively. The training data contained 12,079 articles and the testing data contained 5,664 articles. For each news article, we were able to obtain the number of votes of each emotion. We collected an article seven days after it was published to ensure that the votes have stabilized so that the rank of each emotion would no longer be subject to change.

3. CLASSIFICATION METHOD

Our goal was to classify the news articles into one of the 8 emotion classes provided by Yahoo! Chinese news. Several features were extracted from the news articles. The first set of features consists of all the Chinese character bigrams that appear in the articles. For the second set of features, we applied Stanford NLP group's Chinese segmentation tool on the title and content of each article. The words output by the segmentation tool were used. The third set is the metadata of the articles, which are the news reporter, news category, location of the news event, time (hour of publication) and news agency. News category refers to the newspaper sections such as business and political. The fourth set is the emotion categories of words. Yahoo! Kimo Blog service (<http://tw.blog.yahoo.com/>) provides users 40 emotion categories. Words can be assigned emotion categories and weights by machine learning using the Yahoo! Kimo Blog data sets [2]. LibSVM (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>) was used as the classifier. We used binary weights for the first three sets of features. The last set of features had weights ranging from 0 to 1, depending on the relative frequencies of each of the 40 emotion categories.

Table 1. Testing accuracy using different feature combinations

| Incl.Useful | Metric | Baseline | BI | WD | MT | BI+WD | BI+MT | WD+MT | BI+WD+MT | BI+WD+MT+EC |
|-------------|--------|----------|--------|--------|--------|--------|--------|--------|----------|-------------|
| Yes | Top-1 | 15.34% | 70.81% | 69.75% | 49.77% | 71.11% | 71.85% | 70.51% | 72.01% | 72.05% |
| No | Top-1 | 17.35% | 69.77% | 68.87% | 50.67% | 70.35% | 70.49% | 69.11% | 70.75% | 71.03% |
| Yes | Top-2 | 45.33% | 86.22% | 85.22% | 65.21% | 86.17% | 86.97% | 85.57% | 87.02% | 87.19% |
| No | Top-2 | 46.90% | 85.75% | 84.75% | 66.97% | 85.89% | 86.25% | 84.77% | 86.43% | 86.67% |

Table 2. Percentage of correctly classified testing instances for each emotion class including *useful* class using Top-1 metric

| Features | <i>Awesome</i> | <i>Heartwarming</i> | <i>Surprised</i> | <i>Sad</i> | <i>Useful</i> | <i>Happy</i> | <i>Bored</i> | <i>Angry</i> |
|-------------|----------------|---------------------|------------------|------------|---------------|--------------|--------------|--------------|
| BI+WD+MT+EC | 62.84% | 62.40% | 59.10% | 64.67% | 89.66% | 77.44% | 79.87% | 73.89% |

4. EXPERIMENTS AND DISCUSSION

For the experiments, we tried different combinations of features. The results are shown in Table 1. BI, WD, MT and EC denote bigram, word, metadata, and emotion category of word, respectively. For the baseline, we selected the emotion that occurred most frequently among the training instances as the predicted emotion. The “Incl. Useful” columns have the value “Yes” if *useful* was included as a class and the value “No” otherwise. Since news article readers rarely vote unanimously for a single emotion class, votes are usually distributed among several emotions in an article. Taking the multitude of emotional responses into consideration, we employed two kinds of evaluation metrics. For the Top-1 metric, a predicted class of an instance is correct if it agrees with the top-ranking emotion class of the instance. For Top-2 metric, a predicted class is correct if it agrees with one of the top two ranking emotion classes.

From Table 1, we see that the baseline method has the worst accuracies in all rows. This is not because the training and testing data have very different distributions of emotions. In both training and testing data, *happy* class has the most number of instances. Hence, the corpus does not have a strongly uneven distribution of emotions that can be exploited by the simplistic methods.

The figures in Table 1 are consistent in that BI performs better than WD, and BI+MT performs better than WD+MT. The p-values for their differences using paired t-test are 0.014 and 0.0016 respectively when *useful* class is included and the Top-1 metric is used. Holding all other conditions the same, Chinese character bigrams are better features than the segmented words. However, using BI and WD in combination with MT produces better accuracies than using BI and WD separately. Another observation from Table 1 is that the combination BI+WD+MT+EC performs slightly better than BI+WD+MT in all rows. It shows that the emotion category of words has certain influence on the classification accuracy.

Table 2 shows that different classes have very different performance, ranging from having 59.10% correctly-classified instances to 89.66% using features BI+WD+MT+EC. We were concerned that having highly distinguishing event words as features may be the cause of *happy* and *angry* having high percentage figures relative to other emotion classes. As event words may occur only for a short period of time and rarely be used again in future news stories, having event words as the primary distinguishing features is not going to be helpful in enhancing the general coverage of the classification system. To find out if this was really the case, we examined the most frequently occurring features for each class and computed the conditional probability $P(\text{instance } i \text{ 's true class is } c | \text{instance } i \text{ has}$

feature f) to give an indication of how distinguishing these features were. For *happy* class, we found that a feature shared by many instances is the news category sports. In particular, 48% of all *happy* instances belong to the news category sports. It is also observed that an instance with the news category sports has 67% chance of having the true class *happy*. So, the high accuracy of *happy* class can be a result of people’s general enthusiasm over sports rather than a result of a particular event.

A similar observation was obtained for *bored* class. Features that were shared by a great number of *bored* instances were mostly general terms related to politics instead of a particular time-dependent event. These features included the Chinese character bigrams that denote “legislator” and “minister”. One of the most shared features is the political news category metadata. Any training instance in the political category has 64% chance of belonging to *bored* class.

In Table 2, *useful* class has the highest percentage of correctly classified testing instances. As we examined the most frequently occurring features of *useful* class, we found a lot of features related to weather. An instance with the weather category has 84% chance of belonging to *useful* class in the training set and 95% chance of belonging to *useful* class in the testing set. The weather forecast news articles contribute to the performance of *useful* class.

5. CONCLUSION

The combination of bigrams, words, metadata and word emotion categories achieves the best accuracy in readers’ emotion classification, i.e., 72.05% and 87.19%, under Top-1 and Top-2 metrics, respectively. The accuracies of “useful”, “happy”, “bored” and “angry” classes are more than 73% on Top-1 metric. We will study additional cues to extend the methodology to other types of documents and ultimately integrate it into IR systems.

6. ACKNOWLEDGMENT

Research of this paper was partially supported by Excellent Research Projects of National Taiwan University, under the contract 95R0062-AE00-02. We would like to thank Ming-Feng Tsai for developing a tool for us to collect Yahoo! news data set.

7. REFERENCES

- [1] Yang, C.H. and Chen, H.H. A study of emotion classification using blog articles. In *Proceedings of 18th ROCLING Conference*, September 7th-8th, 2006, Taiwan, 253-269.
- [2] Yang, C.H., Lin, K.H.Y. and Chen, H.H. Building emotion lexicon from Weblog corpora. In *Proceedings of 45th Annual Meeting of Association for Computational Linguistics*, poster, June 23rd-30th, 2007, Prague, Czech Republic.