STAT 215B Final Project, Spring 2012

Christine Kuang, Siqi Wu, and Angie Zhu

April 22, 2012

1 Introduction

the largest microblogging website in China, Sina Weibo 新浪微博
A post can be text, an image, video, or other multimedia.

2 Methods

2.1 Sampling

2.2 Processing

UTF-8 encoding, GBK, Unicode, Big5

2.2.1 Characteristics of Chinese Language

No explicit delimiter between words in Chinese texts
OOV
ambiguity
[7]

2.2.2 Characteristics of Sina Weibo Posts

Our analysis takes not only the characteristics of Chinese language, but also the characteristics of Sina Weibo posts into consideration.

The writing of Weibo posts are generally informal. The users may not use standard punctuation marks for separation of sentences and parts of sentences. The most important features are described as follows:

Reposting A user may repost other post. Reposting does not automatically imply agreement or liking. This type of post usually consists of two parts: the reposting user's comment and the post being reposted. The user's comment may be empty or set as default text "Repost" or "转发微博" ("Repost Weibo"). The reposted post itself may include multiple reposting. The topic of keeping track of reposting and identifying agreement or disagreement can be a project itself. In our analysis, only the reposting user's comment is kept.

Spams There are a fair amount of spams on Weibo. Some spam posts are identical except for the URL. Hence, URLs are removed and then we check for duplication in the pre-tagging processing step.

Mentioning A user may mention other users whose usernames are preceded by the **@** symbol. The mentioned usernames may be an integrated part of the post. We define a set of topic-related usernames and substitute the mentioning of these usernames by the corresponding proper nouns. The other mentioned usernames are removed.

Emotion Symbols and Internet Slangs Sina Weibo provides the users a set of emotion symbols, which are corresponding words surrounded by square brackets in text. The users may use other emotion symbols, such as `:)` for smile and `T_T` for crying. Internet slangs are a large part of Out-of-Vocabulary (OOV). Some substitute the characters in a word with the characters which have similar pronunciation, such as “蜀黍” (Shu3 Shu2) for “叔叔” (Shu1 Shu1, means “uncle”). Some are Internet popular interjections, such as “喵了个咪” (喵: “meow,” 了: past tense marker, 个: universal measure word, 咪: “mew”) which means “dog my cats.”

Topic Topic words are surrounded by the pound signs `#` since Chinese language has no explicit delimiter between words. The topic word can be an integrated part of the post.

2.2.3 Pre-tagging Processing

The data set `Han.txt` contains 22,398 posts.

In order to obtain labeled messages for training and testing purpose, the authors manually provided sentiment tags to a data set containing 3000 messages. The three types of sentiment tags used here are positive, negative, and noninformative.

The data need to be cleaned before manual labeling. A typical post looks like:

```
1165303315 2012-04-16 09:55:40 《韩寒收到网友死亡威胁》(来自 @新浪娱乐) http://t.cn/z0prKap
```

1. The user identification number and time stamp are removed.
2. Only the reposting user’s comment is kept. The reposted part is removed from further analysis. If the resulting string is empty, it will be eliminated as well.
3. URLs are removed.
4. Duplicates are removed.

The output file `hanhanweibo.txt` consists of 13,070 unique posts. 3000 posts are chosen from this data set and will be manually tagged.

2.2.4 Pre-segmentation Processing

As discussed in Section 2.2.1, sentences in Chinese are normally strings of Chinese characters without spaces between words. Hence, word segmentation is crucial for our word-based analysis. According to the characteristics of Weibo posts described in Setion 2.2.2, the following processing is preformed:

1. A set of topic-related usernames are defined. Then the mentioning of these usernames are substituted by the corresponding proper nouns. The other mentioned usernames are removed.
2. A set of emotional symbols and Internet slangs are defined. Then they are substituted by the corresponding word surrounded by square brackets.

2.2.5 Segmentation

汉语词法分析系统 ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) is a well known Chinese word segmentation system developed by Institute of Computing Technology, Chinese Academy of Sciences [3]. It offers the functionality of Chinese word segmentation, lexical tagging, named entity recognition, unknown words detection, and the user-defined dictionary. The current version is ICTCLAS 2011, which supports GB2312, GBK, UTF8 and several encodings and has precision rate of 98.45%.

The Java version of ICTCLAS 2011 preforms word segmentation and lexical tagging on a Linux 32-bit machine. A user-defined dictionary is provided. The entries in this dictionary contains proper nouns and some common Internet slangs. For instance, 微博 (Weibo, wei1 bo2) can be written as 围脖 (wei2 bo2, means "scarf"). Some users refer Han Han as 韩少 (韩: Han Han's surname, 少: abbreviation of 少爷, which means "young master of the house").

Even with the user-defined dictionary, some appearance of Han Han's name 韩寒 can not be segmented and tagged correctly. This is corrected directly using regular expression.

2.2.6 Conjunction Rules

Lee and Renganathan [4] presented that special consideration should be given to the sentences whose parts are linked by contrasting transitional expressions. In particular, if a sentence contains conjunctions such as "although" and "but," only the part being emphasized will be kept and used to infer the sentiment polarity of this sentence. There are three cases:

1. Although (part A), (part B).

2. (Part A), but (part B).

3. Although (part A), but (part B).

For each case, only part B will be kept.

The four words for "although" are 虽然, 虽说, 虽, and 尽管. The words for "but" are 但, 但是, 不过, 可是, 然而, 只是, 可, 只, 然, and 却.

2.2.7 Stop Words and Punctuation Elimination

Moreover, stop words, non-text strings, and punctuation marks are eliminated. The detailed process is as follows:

1. Remove prepositions, punctuation marks, English character strings, interjections, modal particles, onomatopoeia, and auxiliary words.

2. Remove pre-defined stop words and number strings.

Note that the pre-defined stop words do not contain the following six negation words: 不, 不是, 没有, 没, 无, and 别. These negation words will be used in sentiment score assignment.

2.2.8 Sentiment Score Assignment

Dictionary-based sentiment score can provide us some intuitive understanding of the sentiment polarity of the posts. The dictionaries are obtained from HowNet [2]. HowNet is an online extralinguistic common-sense knowledge system for the computation of meaning in human language technology.

Each post is examined and the numbers of positive and negative words are recorded. Positive word contributes +1 to the sentiment score, whereas negative word contributes −1. If there is a negation words among the three words before the positive/negative word, their combination will be treated as an entity and their updated contribution is −1 times the original contribution. The six negation words used are 不, 不是, 没有, 没, 无, and 别. The sentiment score of the post is the sum of all the contributions.

Also topic-related positive/negative words are added to the dictionaries. For instance, the users who refer Han Han as 韩少 (韩: Han Han's surname, 少: abbreviation of 少爷, which means "young master of the house") clearly have positive feelings about him.

Another interesting quantities are the numbers of positive and negative words in a neighborhood of a particular person. The neighborhood used in our analysis is three words before and after the person's name.

2.3 Sentiment Analysis

3 Results

4 Discussion

ROC curve precision and recall curve

4.1 Limitations

sampling [1] [5] [6]

reposting

other language, such as English

simplified Chinese and traditional Chinese: no simple one-to-one correspondence; word segmentation and then substitute words

5 Conclusion

References

[1] Boyd, S., and Xiao, L. Diaconis, P., and Xiao, L. Fastest mixing markov chain on a graph. *SIAM review* (2004), 667–689.

[2] Dong, Z., and Dong, Q. 知网 HowNet. `http://www.keenage.com//`.

[3] Institute of Computing Technology, Chinese Academy of Sciences. 汉语词法分析系统 ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System). `http://ictclas.org/`, 2011.

[4] Lee, H., and Renganathan, H. Chinese sentiment analysis using maximum entropy. *Sentiment Analysis where AI meets Psychology (SAAIP)* (2011), 89.

[5] Leskovec, J., and Faloutsos, C. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (2006), ACM, pp. 631–636.

[6] Wang, T., Chen, Y., Zhang, Z., Xu, T., Jin, L., Hui, P., Deng, B., and Li, X. Understanding graph sampling algorithms for social network analysis. In *Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on* (2011), IEEE, pp. 123–128.

[7] Wong, K., Li, W., Xu, R., and Zhang, Z. Introduction to chinese natural language processing. *Synthesis Lectures on Human Language Technologies 2*, 1 (2009), 1–148.