STAT 215B Final Project, Spring 2012

Christine Kuang, Siqi Wu, and Angie Zhu

May 1, 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 Tagging

In all, we tagged a total of 4000 Weibo posts. We made a search for the topic we were looking into and did this over time to get a total of 10,000 Weibo posts, from which we picked out 3000 posts as our training set. We each tagged 1000 posts. The final 1000 posts were for our test set and we tagged these together. We had a total of four different categories that we tagged the posts as - neutral, positive, negative, and irrelevant (spam).

As we each tagged the posts, we encountered and realized a few of the limitations involved. One limitation was in the fact that tagging these posts produced subjective responses. What one of us read as a negative response to the topic we chose, another may have read as a positive response. For example, the English phrase 'that wasn't too bad' could be taken as positive or negative. Positive - the experience was better than expected; negative - the experience wasn't great. Hence, it is difficult sometimes to truly know whether the original author of the post had in mind a negative or positive reaction to the topic he was posting about.

Another limitation was that some posts we were not sure how to tag. Many posts consisted of just a quote by the author we were looking into. These could have been seen as positive posts since the writer may have liked the quote and so posted it. But at the same time, the author could have been neutral and was just merely using the quote to apply to a specific circumstance in his life at the time. We were not entirely sure of how to tag each of the posts that fell into this category, and so again, the subjectiveness of tagging the posts by hand comes into play as a limitation in the forming of our model. Another uncertainty that occurred in trying to manually tag the posts was that some of the posts didn't talk about our chosen topic specifically, but a related topic. With our chosen topic, people who had negative responses towards 韓寒 (Hanhan), were usually on the side of the opposing author who was discrediting him. Some of the posts did not directly mention 韓寒, but would instead show support for the opposing side. With these posts, we typically labeled as a negative response. However, an argument could be made for just throwing out those posts since they do not directly say anything about 韓寒, and they could possibly just be supporting the opposing author in his own literary works and not necessarily in his stance against 韓寒. Another uncertainty in tagging is what to tag posts that have no subject. There were a few posts that

had nothing to do with the chosen topic at all, but there were also posts that had no subject but contained phrases such as "Keep it up!", "always a supporter!", etc. that could very well be taken as positive posts since our search is for our specific topic. But because there is no subject in the posts, this cannot be taken with 100% certainty. In these instances, where there is no subject in the posts, we marked them as irrelevant to be on the conservative side in our predictions.

These limitations in tagging will affect our model's accuracy in predicting whether a post contains a positive or negative response to the topic at hand. These limitations also indicate to us that in general, models for predicting whether or not a post is negative or positive towards a chosen topic is limited greatly by the subjectiveness of the sentence's meaning/interpretation which affects the training set used to form the model. This limitation could present a potential future question to look into on how we can improve tagging - whether we should just remove posts that have too subjective of an interpretation to tag or if by studying these types of ambiguous posts, we can create certain priors to help in the tagging process.

2.3 Processing

UTF-8 encoding, GBK, Unicode, Big5

2.3.1 Characteristics of Chinese Language

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

There has been quite a bit of research done into natural language processing in the English language, but not much on the Chinese language. This is due to the fact that the Chinese language contains unique characteristics that makes it difficult to do natural language processing well.

First, the Chinese language, unlike the English language, has no explicit delimiter between words. In the English language, spaces separate words and so to segment a sentence into its appropriate components is not too difficult since one could just separate based on spaces. The Chinese language however has no such delimiter between words.

An additional difficulty in segmenting a sentence into its appropriate components and words is that the Chinese languages made of separate characters where each character has a meaning of its own, but if you combine two or more characters together into a phrase, the meaning can be completely different. These types of ambiguities are typically easily resolved by humans reading the sentence and realizing what would make most sense in terms of the context of the sentence, but it is not so easy for a computer to preform those types of automatic word/phrase segmentations. For example, 他好吃 - this sentence could be separated in two ways. The first character means 'he'. The other two words can be taken as two different phrases: 'tastes delicious' or 'loves to eat'. It is obvious that the phrase should be taken as "he loves to eat" because "he tastes delicious" does not make sense but it is difficult to have a computer automatically recognize which sentence makes more sense. Some words also have more than one meaning. For example, 打 can be used in different ways with different meanings. It can be used in the contexts of playing a sport, hitting a person or object, or playing a game.

Second, the Chinese language has many OOV words (out of vocabulary). These are new phrases that are not in dictionaries. These phrases typically come out of cultural references, current hot topics, acronyms, abbreviations, names/nicknames, or just plain slang words. Specifically to the

posts that we looked into, out of vocabulary words typically occurred in the case of cultural references, where there are nicknames created for some hot topic issue/person/event of the week just popular slang or informal words and phrases used to convey an emotion. Especially since social media posts are more popular with the young adult generation, many of the words used in the posts are popular informal phrases that would not normally be found in dictionaries. Knowledge of these types of OOV words comes out of knowing the current trends in Asian countries and being up to date with the typical language used by the younger generation.

Third, the Chinese language has two forms - traditional and simplified and it is not a 1-1 correspondence between the two languages which makes it difficult to convert between the two languages for translation or natural language processing.

2.3.2 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.

2.3.3 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/zOprKap`

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.

2.3.4 Pre-segmentation Processing

As discussed in Section **??**, 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 **??**, 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.

2.3.5 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 [**?**]. 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.

2.3.6 Conjunction Rules

Lee and Renganathan [**?**] 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.

2.3.7 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.

2.3.8 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 [**?**]. 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.

2.4 Sentiment Analysis

2.4.1 L_1 LASSO

Lasso regression is one of the most well known sparse regression methods. Lasso regression uses a L-1 norm penalty to make the regression coefficient vector beta be sparse. The larger lambda is, the more sparse the optimal beta will be. Our resulting models yield classification rules where we take the sign of the model to indicate whether it is in the category we are predicting (positive sign) or not (negative sign).

We used the lasso regression method to create three models model for predicting: positive response, negative response, and the unidentifiable response (we combined spam and irrelevant responses into this one category). We used the R package "lars" to form our models. This package efficiently fits an entire lasso sequence with the least squares loss function. To find the optimal lambda (the regularization parameter) for our L-1 penalty, we used a 10-fold cross validation on our training set of 3000 posts.

For each of the three models, we changed all responses that were not the category we were modeling to -1 and responses that fell into the category we were looking into as 1. Our observation matrix is the frequency matrix of dimension 756x3000. The columns represent each post and the rows represent 756 phrases. These 756 phrases have total occurance of at least 10 times over all 3000 posts and do not include stop words and negation words. Each element of the observation matrix represents the total number of times that particular word (indicated by row) occurrs in that post (indicated by column). Using 10 fold cross validation, we searched for the optimal regularization parameter. The function we used from the lars package in R does not directly return which lambda gives us the optimal model, but rather returns the fraction of the sum of all the absolute values of the betas for that particular model divided by the maximum of all the sums of absolute values of the betas for each model. We pick as our optimal model the model that minimizes our classification error.

To look at which words were most relevant to each category we modeled (positive, negative, and unidentifiable), we looked at the top 50 words based on the top 50 highest absolute beta values, the top 50 words based on the top 50 highest positive beta values, and the top 50 words based on the top 50 highest negative beta values. The top 50 words based on the top 50 highest absolute beta values tells us in general which words were most relevant to predicting that particular category. The top 50 words based on the top 50 highest positive beta values tells us which words typically were common in posts that fell into the category observed while the top 50 words based on the top 50 highest negative beta values tells us which words typically were not in posts that fell into the category observed but were rather more common in the categories outside of the one we were modeling.

Results: top 10 words & explanation & possibly interpretation.

3 Results

4 Discussion

ROC curve precision and recall curve

4.1 Limitations

sampling [?] [?] [?]

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., 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.