STAT 215B Final Project, Spring 2012

Christine Kuang, Siqi Wu, and Angie Zhu

April 18, 2012

1 Introduction

2 Methods

2.1 Sampling

2.2 Processing

Sample format:

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

UTF-8 encoding, GBK, Unicode, Big5

2.2.1 Characteristics of Chinese Language

2.2.2 Characteristics of Weibo

2.2.3 Pre–segmentation Processing

* take out the user ID# and time stamp and save them into info.txt * keep only the first reposting. repostings can be identified by "//" (note: there are a lot of empty repostings on Weibo) * need to remove URL as well since ICTCLAS doesn't handle it well (need to do this before remove mentioning) * for topic-related Weibo usernames, change "@xyz" to the corresponding person name * change mentions to @, which will be removed eventually * substitute emotional symbols to "[words]" using pre-defined "emotionSymbols.txt" (Note: don't worry about topic "#...#" since we are going to remove punctuation marks later)

2.2.4 Segmentation

using the second level lexical tags (see ICTCLAS lexical tagging documentation) with user dictionary "userdict.txt" (still have issues to identity HanHan's name)

2.2.5 Post–segmentation Processing

* handle the incorrect tagging of HanHan's name * conjunction rules: see Lee and Renganathan 2011 * remove prepositions (labeled as "p"), punctuation marks (labeled as "w"), English character strings (labeled as "x"), interjection (labeled as "e"), Modal Particles (labeled as "y"), onomatopoeia (labeled as "o"), and auxiliary words （labeled as "u"). See ICTCLAS lexical taggingx documentation for details. * remove stopping words and number strings (note: the negation words are not in the stopping word list)

2.2.6 Conjunction Rules

2.2.7 Stop Words and Punctuation Elimination

2.2.8 Sentiment Score Assignment

2.2.9 Limitation

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

2.3 Sentiment Analysis

3 Results

4 Discussion

ROC curve precision and recall curve

5 Conclusion