Public Mood and the Collective Sentiment of Tweets

Yujie Lu Kotaro Sakamoto Hideyuki Shibuki Tatsunori Mori
Graduate School of Environment and Information Sciences, Yokohama National University
{luyujie, sakamoto, shib, mori}@forest.eis.ynu.ac.jp

1 Introduction

Many applications using user-generated contents from social media, such as Twitter, have been proposed [6, 4, 2]. These applications would benefit from higher performance of Twitter sentiment analysis. To support the development and evaluation of Twitter sentiment analysis systems, [3] constructed a multilingual corpus for deeper sentiment understanding in social media (the MDSU corpus, for short). This corpus is built to reveal the key principles behind the expression of feeling and to explore the linguistic clues to tweet-level sentiment classification.

However, instead of demanding the global polarity of each tweet, some systems only need the public mood of a group of people (i.e., collective sentiment). For systems developed to forecast election results, to predict stock price movement and to poll public opinion on social events, the collective sentiment takes priority over the tweet-level sentiment. As such, if we only need to grasp the collective sentiment of collections of tweets, is there any easy way to obtain it? Besides, will components, such as negation and rhetoric, affect the collective sentiment?

Compared with the discussion on the emotional mechanism of one tweet (i.e., tweet-level sentiment classification), collective sentiment is much less studied. Although we include the discussion of collective sentiment in [3], we did not shed too much light on it to avoid lose the focus of that paper (i.e., no findings about collective sentiment is included in the main conclusions). Therefore, in this paper, we separately present the main findings about collective sentiment based on the analysis of the MDSU corpus.

2 MDSU Corpus

The MDSU corpus involves 3 languages (i.e., English, Japanese and Chinese) and 4 international topics (i.e., iPhone 6, Windows 8, Vladimir Putin, and Scottish Independence[1]), which consists of 12 collections. Totally, the corpus has 5422 tweets, with each collection containing approximately 450 tweets that were carefully selected following our selection strategy. The English and Japanese tweets are collected from Twitter[2], and the Chinese tweets are collected from Weibo[3], a Chinese version of Twitter.

We proposed a novel sentiment annotation scheme that embodies the idea of separating emotional signals and rhetorical context, and required our native annotators to identify key components of expression of feeling including rhetoric devices, emotional signal, degree modifiers and subtopics, in addition to global polarity. Further, to improve the inter-annotator agreement, we asked the annotators to recheck their original answers by comparing their original answers with a pivot dataset. A gold-standard dataset is then obtained by merging the annotators' revised datasets, which is the MDSU corpus.

Here is an annotated example tweet from the corpus. In the first two sentences, there are three positive signals (i.e., wow, can, and like) and two intensifiers without a specific context (i.e., just and any). Next, the polarity of iPhone 6 is compared to a negative object in the third sentence. The sarcasm identified across the three sentences then finally determines the global polarity of the original tweet as being "negative."

Wow(positive), **with #iPhone6, you can**(positive) **send a message just**(intensifier) **by talking! In any**(intensifier) **voice you like**(positive)**. [So can my mom's old**(negative) **[rotary dial]***(Comparatively equal)*.**]***(Sarcastically negative)* ⊙ Global Polarity to iPhone 6: **Negative**

3 Collective Sentiment

The collective sentiment (denoted as the PN ratio) for an object is used to represent public opinion, measuring the degree of happiness of a group of people [4, 2]. The PN ratio of object X of a collection is defined as

$$\text{PN ratio(X)} = \frac{\text{\#positive tweets of X in the collection}}{\text{\#negative tweets of X in the collection}} \quad (1)$$

[1] I6, W8, PU and SI for short, respectively.

[2] http://www.twitter.com

[3] http://weibo.com

Table 1: PN Ratio and Global Polarity Distribution of Each Collection, with Positive/Negative/Neutral Meaning the Number of Tweets of the Correspondent Polarity in a Collection

Collection		Positive #	Negative #	Neutral #	PN Ratio
I6	English	197	129	125	1.53
	Japanese	120	187	128	0.64
	Chinese	205	145	100	1.41
	avg.	**174**	**154**	**118**	**1.13**
W8	English	70	256	128	0.27
	Japanese	57	250	158	0.23
	Chinese	81	283	91	0.29
	avg.	**69**	**263**	**126**	**0.26**
PU	English	52	249	148	0.21
	Japanese	184	64	210	2.88
	Chinese	174	139	131	1.25
	avg.	**137**	**151**	**163**	**0.91**
SI	English	184	140	125	1.31
	Japanese	31	33	379	0.94
	Chinese	106	71	292	1.49
	avg.	**107**	**81**	**266**	**1.32**
Total #		1461	1946	2015	0.75

By definition (1), if the PN ratio is greater than one, people are happy with the object, while a value less than one indicates the opposite. When the size of the collection is too small or the polarity distribution is skewed, the numerator or denominator tends toward zero. In such instances, they are set to one in practice.

Table 1 shows the PN ratio and polarity distribution of each collection. Through the PN ratios, we can understand the public mood on each evaluation object for each culture. iPhone 6 was welcomed by English users (i.e., 1.53) and Chinese users (i.e., 1.41), whereas Japanese users (i.e., 0.64) showed an unfavorable attitude. As for Windows 8, all three cultures were complaining about their unpleased experience (i.e., 0.27, 0.23, and 0.29 for English, Japanese, and Chinese users, respectively). Individuals were evidently divided over Putin. Japanese users (i.e., 2.88) and English users (i.e., 0.21) markedly opposed one another, whereas Chinese users (i.e., 1.25) adopted a pro-center stance. Regarding Scottish Independence, both English users (i.e., 1.31) and Chinese users (i.e., 1.49) showed their support for independence. Japanese users (i.e., 0.94) were almost neutral on this issue. Based on the observation of PN ratios, we can see that public mood varies between cultures (its variance depends on the topic.). Note that the entire corpus is well-balanced, with 0.75 inclined to the negative side.

4 Similarities among WPN, SPN and GPN

Since the global polarity of each tweet is difficult to obtain, the word-level PN ratio is often used as a substitute for the tweet-level PN ratio [1, 5]. In this section, we verify whether this substitution is valid.

For ease of reference, we use WPN to denote the word-level sentiment ratio based on polarity lexicons[4]; SPN to denote the sentiment ratio based on hand-labeled emotional signals, which acts as the true value for WPN[5]; and GPN to denote the tweet-level PN ratio. By counting how many positive or negative words or signals occur in a collection, we can arrive at values for WPN and SPN. More specifically, the WPN and SPN of object X for a collection are defined as

$$\mathsf{WPN(X)} = \frac{\#\mathsf{positive\ words\ of\ X\ in\ the\ collection}}{\#\mathsf{negative\ words\ of\ X\ in\ the\ collection}} \tag{2}$$

$$\mathsf{SPN(X)} = \frac{\#\mathsf{positive\ signals\ of\ X\ in\ the\ collection}}{\#\mathsf{negative\ signals\ of\ X\ in\ the\ collection}} \tag{3}$$

Table 2 compares the three sentiment ratios. First, it shows that SPN has a stronger correlation and smaller gap (i.e., $r = 0.92$ on average, gap $= -0.19$ on average) with GPN than WPN does (i.e., $r = 0.76$ on average, gap $= -0.26$ on average) in all three languages; however, despite WPN being poorer than SPN, there is no statistically significant difference among GPN, SPN, and WPN (i.e., paired t-tests, all $p > 0.05$). In other words, SPN and WPN can both be possible substitutes for GPN, but SPN is more accurate. Therefore, it is acceptable to use WPN to represent public opinion in opinion-mining applications.

We also found that the correlation between WPN and SPN was relatively high and the gap between them was small (i.e., $r = 0.93$ on average, gap $= -0.07$ on average). We further computed the matching percentage of polarity words and emotional signals. Since emotional signals are allowed to be phrases (e.g., makes a...difference), we assume that if a polarity word hits any word of an emotional phrase, then it is a successful match. Further, the polarities of both sides should be identical.

The gap between WPN and SPN occurs primarily for two reasons. First, there was a failure in detecting emotional signals using polarity dictionaries. The average signal matching rates reached only 44.3%, 33.4%, and 33.2% for English, Japanese, and Chinese, respectively. These results have occurred

[4] The lexicon we used are Liu Bing's English opinion lexicon, Chinese emotion ontology lexicon, Japanese sentiment polarity lexicon and the SentiStrength emoticon lookup table.

[5] As described by the example tweet in Section 1, emotional signals are the words/phrases that actually affect the global polarity of tweets.

Table 2: Comparison of WPN, SPN, and GPN, Including the Mean of WPN, SPN and GPN, Correlation Coefficient, and p-value of Paired t-tests Calculated Over all 12 Collections

Language	Ratio Type	Mean	Gap with GPN	Correlation with GPN (Correlation with SPN)	p-value of Paired t-test with GPN
English	GPN	0.83	—		
	SPN	0.99	−0.16	0.97	0.246
	WPN	1.07	−0.24	0.88 (0.97)	0.406
Japanese	GPN	1.17	—		
	SPN	1.49	−0.32	0.83	0.420
	WPN	1.52	−0.35	0.61 (0.95)	0.514
Chinese	GPN	1.11	—		
	SPN	1.21	−0.10	0.97	0.224
	WPN	1.31	−0.20	0.79 (0.86)	0.341

because many of the emotional phrases are composed of non-polarity words and some signals have not yet been registered in the polarity dictionaries. Second, many polarity words were mistaken as emotional signals. The average word mismatching rates were 53.6%, 83.0%, and 73.3% for English, Japanese, and Chinese, respectively, all of which are more than half. Here, the polarity words are not necessarily evaluating the objects, but rather can be narrative or off-topic, which accounts for the extremely high word mismatching rates for Scottish Independence in Chinese and Japanese, since both collections have a limited number of non-neutral tweets (Table 1).

Incorrectly registered non-opinionated words in polarity dictionaries can also further worsen the problem, since solutions to both problems above require high-quality polarity dictionaries. In our experience, WPN changes largely from dictionary to dictionary[6]. As for topic consistency, we regard it as an inherent gap between SPN and WPN, with WPN calculated only via simple counting, i.e., involving no topic-oriented technology. Finally, although there is plenty of room for improvement to use WPN as a proxy for GPN for all three languages, its adaptability in English is basically better than that in Japanese and Chinese.

5 Influence of Components on GPN−SPN

In this section, we further reveal the influence of components on collective sentiment.

Because both GPN and SPN were calculated from the manually-labeled annotations in the corpus, the gap between them can be regarded as originating from the context of the tweets[7]. Hence, we can use the gap between GPN and SPN (denoted GPN−SPN) to approximate context influence. If a particular type of context has no influence on global polarity, GPN−SPN will be similar regardless of whether it is present or not. We therefore conducted a one-way analysis of variance (ANOVA) to examine the influence of the presence or absence of certain types of components (i.e., independent variables), including degree modifiers and rhetoric devices, on GPN−SPN (i.e., a dependent variable).

Table 3: Difference of GPN−SPN and Results of ANOVA, with Presence/Absence Meaning the Mean of GPN−SPN at the Presence/Absence of Each Factor and p-value of ANOVA Calculated Over all 12 Collections

Factor	Presence	Absence	p-value
Modifiers	0.024	−0.293	0.087
Modifiers−Negation	−0.190	−0.193	0.986
Diminisher	−0.173	−0.189	0.942
Intensifier	−0.134	−0.206	0.663
Negation	−0.580	0.075	**0.004**
Rhetoric Devices	−0.336	−0.034	0.140
Rhetoric−Sarcasm	−0.124	−0.223	0.639
Comparison	−0.188	−0.160	0.911
Metaphor	−0.192	−0.068	0.717
Rhetorical Question	−0.139	−0.319	0.313
Sarcasm	−0.119	−0.775	**0.001**

Table 3 shows the GPN−SPN difference and the results of our ANOVA, showing that intensifiers and diminishers together (i.e., Modifiers−Negation) had little influence on the collective sentiment ratio (i.e., $p = 0.986 > 0.05$) and that their GPN−SPN difference was trivial (i.e., -0.003). Conversely, the influence of negation was significant (i.e., $p = 0.004 < 0.01$)[8]. Here, the GPN−SPN of the non-negation

[6] Low-quality dictionaries can generate rather meaningless results, so all three dictionaries we selected have been checked manually by their providers.

[7] We ignore quantization error (e.g., two positive tweets and one negative tweet (GPN = 2) may have five positive and two negative signals (SPN = 2.5) since our collection is quite large.

[8] Some opinions were toward opposites of Scotland in Scot-

Table 4: Results of ANOVA by Language, with p-value of ANOVA calculated over the 4 collections of each language.

Factor	English	Japanese	Chinese
Modifiers	0.498	0.141	0.162
Modifiers−Negation	0.425	0.907	0.387
Diminisher	0.232	0.305	0.665
Intensifier	0.487	0.844	0.904
Negation	0.739	**0.042**	**0.017**
Rhetoric Devices	**0.032**	0.664	**0.022**
Rhetoric−Sarcasm	0.330	0.551	**0.035**
Comparison	0.923	0.733	0.227
Metaphor	0.732	0.257	0.131
Rhetorical Question	0.335	0.726	0.027
Sarcasm	**0.002**	0.334	0.064

collection was small (i.e., 0.069), while it was large (i.e., −0.579) for the negation collection. For rhetorical phenomena, we found that sarcasm had the same influence as negation on collective sentiment (i.e., $p = 0.001 < 0.01$); although other rhetoric devices (i.e., Rhetoric−Sarcasm) were not statistically significant (i.e., $p = 0.639 > 0.05$), their overall GPN−SPN difference was not trivial (i.e., −0.302).

We performed similar ANOVA analyses for each language. Table 4 details the results here by language. The table indicates that Modifiers−Negation did not have a significant influence on collective sentiment for all three languages (i.e., $p > 0.05$), as expected. Surprisingly, the influence of Negation was significant for Japanese and Chinese (i.e., $p = 0.042$ and 0.017, respectively), but not for English (i.e., $p = 0.739$). This occurred perhaps because other contexts offset the influence of negation in English. For rhetoric devices, it appears that there was a significant difference for both Chinese and English (i.e., $p =$ 0.032 and 0.022, respectively), but not for Japanese (i.e., $p = 0.664$)[9].

In addition, we also conducted a two-way ANOVA to see how negation, rhetoric, and their interaction affect collective sentiment throughout the corpus. Results show that the interaction between negation and rhetoric had little influence on GPN−SPN (i.e., $p = 0.496 > 0.05$), while GPN−SPN was significantly different in terms of the presence of both negation (i.e., $p = 0.000 < 0.001$) and rhetoric (i.e., $p = 0.013 < 0.05$). From the above analyses, it indicates that we cannot deny that either of negation and rhetoric has influence on collective sentiment.

6 Conclusion and Future Work

From the above analyses, we mainly have two conclusions. First, we showed that WPN can be a relatively reliable substitute for GPN, which means that we can quickly get an approximate answer by simply counting polarity words in a collection of tweets using high-quality polarity lexicons for systems that only need to know public mood.

Second, negation and rhetoric most likely have strong influences on the collective sentiment for all three languages, indicating that WPN should be prudently used as a substitute for those collections with rich language phenomenon. Concerning that the size of our collections is limited, we hope that similar investigations can be conducted on larger datasets.

References

[1] Johan Bollen, Huina Mao, and Xiaojun Zeng. Annotating expressions of opinions and emotions in language. *Journal of Computational Science*, Vol. 2, No. 1, pp. 1–8, 2011.

[2] Yujie Lu, Jinlong Guo, Kotaro Sakamoto, Hideyuki Shibuki, and Tatsunori Mori. Predicting sector index movement with microblogging public mood time series on social issues. *Proceeding of 29th Pacific Asia Conference on Language, Information and Computation (PACLIC 2015)*, pp. 563–571, 2015.

[3] Yujie Lu, Kotaro Sakamoto, Hideyuki Shibuki, and Tatsunori Mori. Construction of a multilingual annotated corpus for deeper sentiment understanding in social media. *Journal of Natural Language Processing*, Vol. 24, No. 2, 2017.

[4] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM'10)*, pp. 122–129, 2010.

[5] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. Exploiting topic based twitter sentiment for stock prediction. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pp. 24–29, 2013.

[6] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Election forecasts with twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, Vol. 29, No. 4, pp. 402–418, 2011.

tish Independence (e.g., England); we temporarily regarded these opposites as negation here.

[9] Some rhetoric devices have low occurrences, causing the GPN−SPN values (Presence) somehow less reliable.