Role of Emoticons in Sentence-level Sentiment Classification

Min Martin, Tanya Lee, and Ray Hsu

Netbase Solutions, Inc. Mountain View, USA
{cmin,tlee,rhsu}@netbase.com

Abstract — Automated sentiment extraction from social media is enabling technology to support gathering online customer insights. The basic sentiment extraction is semantic classification of a text unit as positive or negative using lexical and/or contextual clues in a natural language system. From the input side, it is observed that social media as a sub-language often uses emoticons mixed with text to show emotions. Most emoticons, e.g. :=), are not natural language words, but textual symbols using characters to present a smiley face. Intuitively, such symbols are innately associated with emotions, whether *happy, annoyed* or *don't care*, hence important clues for helping sentiment classification. Previous research has involved the limited use of emoticons as noisy labels in sentiment learning but detailed study on how noisy or useful they are has not been done. This paper presents a comprehensive data analysis study of the role of emoticons in sentence level sentiment classification. Various investigations are conducted on a fairly large annotated social media corpus, selected by our consumer insight analytics system. This corpus consists of 40,548 sentiment-rich sentences which business users are truly interested in mining. The study shows that the consistency between positive/negative emoticons with human judgment in this corpus is as high as 75.2%. Another larger randomly selected corpus consisting of 300,000 sentences from social media shows its consistency with human judgment to be 40.1%. A further study finds that emoticons' recall contribution to sentiment classification is moderate, nevertheless, the data containing emoticons and brands are guaranteed to be quality social media representing customers' voice instead of businesses' voice such as press news. In addition, emoticon is an additional factor to help extract sentiments where other linguistic clues are insufficient.

Keywords - emoticon; sentiment extraction; sentiment classification; customer insight; social media

I. Introduction

The rapid growth of social data from such online social networks as Twitter and Facebook has aroused enormous interest in the mining and discovery of customer in-

adfa, p. 1, 2011.

sights, recognized as significant for businesses. For instance, Twitter has over 500 million registered users as of 2012, generating over 340 million tweets daily, which is equivalent to 3,935 tweets per second. Sina Weibo, a Twitter like micro-blog service in China, has an active user base of over 40 million. Such huge amounts of unstructured text data involve enormous amount of customer voice that is invaluable to leading brands and businesses, especially in marketing and sales departments. They need to monitor, react, engage, and publish at the speed of social in real time in order to compete. The main motivation behind the adoption of social media intelligence tools like ours lies in the fact that people are increasingly sharing their opinions on products and services they buy or want to buy on social networks. Recent estimates indicate that on average one in every three blog posts and one in five tweets involve comments on products, services or brands (Hogenboom et al. 2013). Netters freely talk about whether they love or hate a brand, and lots of times they compare it with other brands in the same category. Apparently, such information would be really important for businesses to keep track of consumers' attitudes toward their brands and the management can make faster decisions based on the social intelligence when it is extracted from the huge social data pool and analyzed properly.

Sentiment analysis is an essential part of a commercial social media intelligence platform. The majority of sentiment analysis systems are machine learning based, taking a traditional text classification approach to train models like Naïve Bayes, Maximum Entropy or Support Vector Machine. The text unit for classification is usually a document or paragraph consisting of multiple sentences. Typical examples of such data are movie or product reviews. When trained on domain data, those classifiers generally achieve over 80% accuracy for coarse-grained sentiment classification, as positive, negative or neutral, as reported in the study (Pang and Lee 2008). While machine learning based sentiment classification works well in domain data at document or paragraph level, it faces challenges in handling object focused short messages (e.g. twits) in a commercial sentiment analysis system today when the dominant social media platform is mobile and the posts tend to be shorter from mobile users.

Social media as a sub-language is sometimes full of emoticons mixed with text to show emotions of the poster. Most emoticons, e.g. :=), are not natural language words, but textual symbols using characters. Among various definitions, Wikipedia defines emoticon as "a meta-communicative pictorial representation of a facial expression … draw a receiver's attention to the tenor or temper of a sender's nominal verbal communication, changing and improving its interpretation". It mainly uses a combination of punctuation marks to mimic a smiley face to express a person's feeling, mood or intention. Intuitively, people would assume that such symbols are innately associated with emotions, whether *love*, *hate* or *don't care*, hence important clues for helping sentiment classification. With the rapid growth of social media uses globally, it's generally recognized that emoticons have been playing an increasingly important role in online communications, especially for the younger generation. In light of this, a natural language system built for sentiment analysis is expected to take advantage of this special linguistic phenomenon, which is typically not in the scope of the grammar or vocabulary research of a language.

Unlike the main stream sentiment analysis based on statistical models using machine learning, we have built a rule-based high precision sentiment analysis system based on a parser for use in the product of mining consumer insights from social media. This system is designed to address two major challenges of machine learning systems: (i) brand focused sentence-level sentiment classification for short messages; (ii) extracting reasons behind sentiments to answer why questions. The research on emoticons reported in this paper is motivated by the need to enhance the system in handling (i) in the context of mining customer sentiments towards a brand. But the analysis and experiments we have done also serve the purpose of enlightening the researchers in both machine learning world and the grammar world with better understanding of the role of emoticons in sentiment analysis**.** In fact, it reveals a pitfall facing some earlier researchers (Read 2005; Zhao et al. 2012) who assume emoticons are handy and reliable sentiment indicators and therefore use them to collect training corpus for sentiment classification.

The major contribution of this study lies in the fairly comprehensive study of emoticon's distribution in social media and its role in a sentiment analysis system. We aim to accomplish two specific goals.

- Provide a statistical analysis of how emoticons are used in social media from various perspectives
- Provide an evaluation of emoticon's roles in our brand-focused sentence level sentiment classification system by calculating its precision and recall impact on system performance.

The remainder of the paper is organized as follows. Section II reviews related work in using emoticon as a clue for sentiment classification. Section III briefs our parsing-supported sentence-level deep sentiment system to provide a background for this study. In Section IV, emoticon-related experiments are set up and results are presented, focusing on emoticon's statistical distribution in social media, and its precision and recall impact on data quality. We present our findings and conclusions in Section VI.

II. Related work

The popularity of emoticons (or smileys) comes hand in hand with the growth of social network. They have been extremely popular in social media among the younger generation and the seasoned netters. Despite their use everywhere in the online text, linguistically, they are not a "legitimate" part of natural language vocabulary or morphology, hence belonging to so-called Unnatural Language Processing (UNLP, Ptaszynski et al. 2011). Some emoticons are fairly universal as symbols of emotion, and others are language dependent. Survey shows that they are the second most important vehicles for expressing emotions in online communication (Ptaszynski et al. 2011).

In the context of NLP, the use of emoticons has attracted machine learning researchers for the sentiment classification. Emotions seem to be a handy and reliable indicator of emotions and hence are used either to help automatically generate a training corpus for sentiment classification or to act as *seeds* or one type of evidence features to enhance sentiment classification (Davidov, Tsur and Rappoport 2010; Liu, Li and Guo 2012; Read 2005; Zhao et al. 2012; Yang, Lin and Chen 2007; Hogenboom1 et al. 2013).

Not much has been done in evaluating the contributions of emoticons in sizable real life social media corpora, in the context of brand-centered sentiment analysis. That is one major motivation and value for this study.

Ptaszynski et al. (2011) proposes that emoticon research consists of four lines of tasks: (1) detection; (2) extraction; (3) parsing; (4) semantic analysis; (5) generation; (6) evaluation. Our work involves (1), (2) and (6). The work involved in (3), (4) and (5) assumes the productive nature of emoticons, similar to the open morphology study in natural languages. For the following reason, at least for English, this is not a real issue.

There are thousands of varieties of emoticons due to the semantic compositionality of its components, in ways that are very close to flexible word formation in some natural language morphology, with a smiley face being made with various types of eyes, nose and mouth etc. (Strapparava and Mihalcea 2008). However, our study demonstrates that the frequency distribution of different emoticons is very different, and many theoretically possible combinations do not really add to the system due to the infrequency of their appearance in data. At least for English, the top n (n<1000) emoticons are easily listable in lexicon and can fulfill the identification of emoticons with very high precision and recall.

III. Sentence-level Sentiment Classification

Our sentence-level sentiment system is supported by a natural language parser we developed. Each incoming sentence is analyzed by the full NLP parser, starting from tokenization, POS (Part-of-Speech) tagging, chunking and ending in decoding a dependency parse tree enriched with various syntactic and semantic features, on top of which deep sentiment extraction capability is built. For instance,

I like the camera of iPhone because the photo quality is higher.

From this sentence the system is able to extract information as follows:

Sentiment: *positive*
Object: *iPhone*
Aspect: *camera*
Reason: *photo quality (higher)*

The concept of deep, fine-grained sentiments is proposed in contrast to the dominant practice of shallow, course-grained sentiment analysis, thumbs-up and down (or plus neutral) classification, coupled with sentiment association based on proximity. This concept is inspired by the needs from the real world market analysts who were initially very happy with the precision of our sentiment insights and later told us they need more actionable insights and hope we can answer the why questions with regards to sentiments. Over time, the deep sentiments evolve in the process of engaging the users of customer insights and become fairly mature as a standard to drive the research and development supported by deep parsing. To shed some light in the process, a deep sentiment system should be able to extraction insights that can answer these questions in addition to the sentiment classification insight:

- Which brand is this sentiment about? (association insight)

- Can the system associate sentiments not only with a brand such as iPhone, but also with a feature of the brand, say, screen? (granularity insight)
- Who made the sentiment comment? (customer background insight)
- How intense is the sentiment? (passion intensity insight)
- Finally, most important of all, what is the reason of the sentiment? (why insight)

Systems that can answer such questions provide invaluable actionable insights to businesses. For instance, it is much more insightful to know that consumers love the online speed of iPhone 4s but are very annoyed by the lack of support to flash. This is an actionable insight, one that a company could use to redirect resources to address issues or drive a product's development. Extraction of such insights is enabled by our deep parsing.

Since our sentiment extraction must be object centered, meaning that the sentiment extracted must be toward a topic or an object mentioned in a sentence, which could be a brand, or person or location name. We typically select 10-20 brands to evaluate the system. For instance, in the most recent release the brands we used are:

> *iPhone, Walmart, Listerine, Costco, Olive Garden, Taco Bell, Tylenol, Camaro, Prius, Ikea, JetBlue, Skype, Yoplait, Playstation, fish oil, Pepsi*.

We use CrowdFlower's anonymous annotation service to annotate the data, and 75% or above inter-annotator agreement with at least four judges each time is used for benchmarking. The precision we have achieved is 87% on average across the brands.

Due to the fact that sentiment is generally sparse in randomly selected data, we have not really taken a standard approach to evaluate the recall, since that would require annotating a significant amount of data than precision evaluation. Also, our experiences have shown that to a certain point, increasing recall is an incremental process, especially when taking into consideration multiple domain factors. Thus, our evaluation has been focused on precision benchmark. For tracking relative recall, we simply measure the total number of extracted sentiment mentions and their percentage given a certain amount data processed by the system.

IV. Experiments and Results

In this section, we present three experiments aiming to answer three questions:

- How emoticons are generally used and statistically distributed in social media?
- What's the precision of sentiment classification when only using emoticon as evidence?
- What's recall contribution of emoticons to sentiment classification?

These are questions that can help drive the design and development in properly involving emoticons in a sentiment system. It needs to be noted that in our approach to brand-focused, sentence-level classification, the identification of sentiment from a sentence must be targeted at a specific object.

A. **Experiment Setups**

One of the major resources in the experiments is the emoticon lexicon with a total of 1258 entries, which we collected during the system development from social data. We manually marked 106 entries as positive indicator and 233 as negative indicator, leaving the rest as unspecified.

In addition to the emoticon lexicon, we use two corpora for the evaluation in the experiment. The first one is a human annotated corpus with 40,548 sentences, which is an accumulation of some of the data which our QA department prepared for system evaluation. Each sentence is annotated by four annotators, and a 75% agreement among annotators is the cutoff threshold we adopt for an agreement. To ensure the objective evaluation, our QA department uses a crowd sourcing service for the annotation so the human judges have no association whatsoever with the development team. Each sentence is annotated with a sentiment choice of positive, negative or neutral towards the corresponding object associated with the sentiment.

Another aspect of this corpus is that, sentences in this corpus are not randomly collected, but selected from our sentiment analysis system's output. The implication of this choice is that the data must be sentiment-rich, meaning that many sentences from this corpus would be either positive or negative due to the fact that it is a much smaller subset of data that has been filtered by our sentiment extraction system. One of the main practical reasons for this is that, the primary use of this corpus is for evaluating system's precision, not recall, and the standard precision metric is measured by the fraction of extracted sentiments that are correct against human annotations. For precision measurement, this selection of data is good enough and as a matter of fact, much better. This is because, if the data is randomly selected, the majority of the random data would not contain any sentiment and it requires a much bigger size of data to be annotated. That is not only costly but also meaningless for precision evaluation. Our empirical study has shown that about 15%-25% of data contains a sentiment, depending on a specific brand. Given the fact that this corpus is selected in a way that it is guaranteed to be sentiment rich, we name it SelectCorpus.

As opposed to the SelectCorpus, we also have a second corpus, RandomCorpus, made up of randomly collected data from our content store. The only requirement for this corpus is that each sentence selected must contain at least one brand term, e.g. Listerine. Unlike the sentiment-rich SelectCorpus, the sentiment is much sparser in RandomCorpus. As discussed earlier, on average only 15%-25% of sentences are expressing a sentiment, either explicit or implicit. From a system evaluation point of view, such a random corpus is more representative of social media content. Hence, we would like this corpus to be another set of evaluation data used in the experiments.

While SelectCorpus is human annotated with sentiment, RandomCorpus is not, and thus its size is much larger, with a total number of 300,000 sentences, as opposed to 40,548 sentences from SelectCorpus. Without annotation, how would we measure precision and recall based on this corpus? Here we are not seeking for measuring the absolute performance in the traditional sense. Instead, we estimate precision and recall through these two formulas:

Estimated Precision =
*count of agreed sentiment with emoticon * 0.87 / total count of extracted sentiment*

Estimated Recall =
*count of agreed sentiment with emoticon * 0.87 / total count of sentences * 0.2*

In the formulas, 0.87 is the system's average precision score across brands, which is obtained through our series of evaluation over the system development course. Likewise, 0.2 is the average sentiment richness score in the range of 0.15-0.25 obtained for different brands in our evaluation. Sentiment Richness is defined as the ratio of the number of sentences containing sentiment versus the total number of sentences.

B. Experiment Results

Experiments are conducted and results are reported from two perspectives in evaluating emoticons: (i) statistical distribution of emoticon uses in social media; (ii) emoticon's impact on precision and recall for sentiment extraction.

i) Statistical distribution of emoticon uses in social media

We primarily use RandomCorpus to analyze emoticon's frequencies and distribution, since it is much larger than SelectCorpus. Table I lists overall frequencies and distribution in the corpus.

TABLE I. EMOTICON SENTIMENTS IN RANDOMCORPUS

	counts	Percentile	Sent. count	Richness
Positive	6,752	65.4%		
Negative	1,314	12.9%		
Neutral	2,244	21.7%		
Total	10,310	100%	300,000	3.4%

Like the Sentiment Richness, the Emoticon Richness is defined as the ratio of the number of sentences containing an emoticon versus the total number of sentences in the corpus. This metric informs us of the overall frequency of emoticons used in social media and their maximum possible impact on a sentiment extraction system.

Given the over 1000 emoticons in our emoticon lexicon, we are curious about how each of these emoticons is actually used in social media. For this purpose, we count the frequency of each emoticon's use and its expressed emotion or mood based on the RandomCorpus. The top 10 used emoticons are listed in Table II.

TABLE II. TOP 10 EMOTICONS IN RANDOMCORPUS

Rank	Emoticon	Frequency	Percent	Emotion
1	:)	3,505	34.00%	Happy
2	:D	1,273	12.35%	Laugh
3	:(	927	0.89%	Sad
4	;)	773	7.50%	Wink
5	:-)	711	6.86%	Happy
6	:P	433	4.21%	Tongue out
7	=)	319	3.10%	Happy
8	(:	309	3.00%	Happy
9	;-)	226	2.19%	Wink
10	XD	175	1.70%	Grin
Total	N/A	8,651	83.90%	N/A

Table III demonstrates that although the number of emoticons in social media is big, only a handful of emoticons are used far more frequently than all others.

TABLE III. USER COUNTS OF UNIQUE EMOTICON

Unique users	Number of users	Percentile
1	238000	95.02%
2	6325	2.53%
3	4350	1.74%
4	1265	0.55%
5	400	0.16%

ii) Emoticon's impact on sentiment precision and recall

From Table I, we see that the positive, negative and neutral ratio of emoticons used is 65.4%, 12.9% and 21.9%. This gives us the impression that most emoticons used are conveying positive emotions, either explicit or implicit. However, to what extent is using emoticons as a single sentiment clue correct, especially in our context of brand focused sentiment extraction? For instance, in the message "*Someone asking a greeter at walmart to watch their child ----JOKE :) ha ha ha ha",* even though there is a positive emoticon indicating the poster is happy, there is no any indication of that sentiment is for the object "Walmart". Our major goal here is to find out how reliable it is to use emoticon alone as a sentiment indicator.

We run two experiments on both SelectCorpus and RamdomCorpus. The results are listed in Table IV and Table V, with the precision to be 75.2% from SelectCorpus and 40.1% from RandomCorpus.

TABLE IV. PRECISION USING EMOTICON AS SINGLE SENTIMENT CLUE BASED ON SELECTCORPUS

# of Pos agree	# of Neg agree	# of Pos as Neg	# of Neg as Pos	# of Neu as Pos	# of Neu as Neg	Precision
1052	88	211	84	56	25	75.2%

P = (1052+88) / (1052+88+211+84+56+25) = 75.2%

TABLE V. PRECISION USING EMOTICON AS SINGLE SENTIMENT CLUE BASED ON RANDOMCORPUS

# of Pos agree	# of Neg Agree	# of Pos as Neg	# of Neg as Pos	# of Neu as Pos	# of Neu as Neg	Precision
4068	641	779	421	3001	1400	40.1%

P = [(4068+641) /(4068+641+779+421+3001+1400)] * 0.87 = 40.1%

Each column in the tables represents:

- # of Pos(itive) Agree(ments): number of positive posts agreeing with positive emoticon mention
- # of Neg(ative) Agree(ments): number of negative posts agreeing with negative emoticon mention
- # of Pos(itive) with Neg(ative): number of positive posts with negative emoticon mention
- # of Neg(ative) with Pos(itive): number of negative posts with positive emoticon mention
- # of Neu(tral) with Pos(itive): number of neutral posts with positive emoticon mention
- # of Neu(tral) with Neg(ative): number of neutral posts with negative emoticon mention

The precision numbers differ significantly on the two corpora. It is not surprising however. Since the data in SelectCorpus is from our sentiment extraction system's output, meaning that it is already filtered by our sentiment system. As a result, the sentiment of this data set is much richer. However, the precision number obtained from SelectCorpus still offers insights into how emoticon's sentiment may overlap with the actual sentiment judged by human standards. The 75.2% precision seems to suggest that, when a sentence is guaranteed to be positive or negative, and when there is an emoticon in it, there is a fairly high chance of being the case that the emoticon can be trusted to be a fairly good sentiment indicator. In this case, the chance is 75.2%.

On the other hand, the 40.1% precision from RandomCorpus tells that in the real world, emoticon alone is not reliable enough to be taken as a sentiment dictator. As we can see from Table V, there are a large number of neutral sentences that have been incorrectly classified as either positive (3001) or negative (1400) using emoticon. Mistaking neutral ones as positive or negative has been a common problem for a sentiment analysis system, and our study shows that emoticon cannot be immune to this headache either. For instance,

*at the **marriott** hotel and i ran into vili, small world :D!*

The highlighted *marriott* is the brand we were evaluating. Although the emoticon generally expresses some type of emotion, it does not really indicate a sentiment for the brand we are interested in.

Finally, we would like to get a sense of the recall contribution if emoticon is used as a sentiment indicator. The 3.4% Emoticon Richness score listed in Table I informs us that the maximum contribution to the system recall would be 3.4%, assuming *i)* none of the sentences containing an emoticon have been correctly classified by the system without using emoticon; *ii)* emoticon's precision as an emoticon indicator is 100%. In reality, neither holds true though. As a result, the actual recall contribution would be lower. The result is presented in Table VI. The first column shows emoticons' agreement with the annotation, column 2 is for the emoticons' agreement with annotation missed by the system, column 3 is the recall improvement and column 4 shows the overall recall improvement.

TABLE VI. EMOTICON'S IMPACT ON SYSTEM RECALL IN SELECTCORPUS

	Agree w. key	System miss	Recall up	Average
Pos	1052	20	1.94%	2.72%
Neg	88	3	3.52%	

So how to assess the 2.72% recall improvement on SelectCorpus? The number does not seem to be impressive and significant. However, from a system development point of view, this improvement is meaningful. In our fairly mature English system, we have a total of 400+ rules, which consist of thousands of linguistic patterns built upon a semantic parser. However, the top ten mostly fired rules contribute to nearly 50% of all sentiments extracted. The vast majority of all other rules account for the long tail of the remaining 50% sentiments, where many individual rules contribute to less than 1%. These are either domain specific or linguistically specific rules. Table VII lists the top ten fired rules in our system. Individual rules that contribute to little recall but can correct eye-catching errors cannot be ignored in a real life system. In this sense, the emoticon provides a low-hanging fruit for enhancing the data quality which should not be ignored either.

TABLE VII. TOP TEN RULES CONTRIBUTING TO SENTIMENT EXTRACTION SYSTEM

Firing Frequency	Positive	Negative
# 1	10.16%	9.89%
# 2	9.59%	7.12%
# 3	5.15%	6.11%
# 4	4.96%	5.22%
# 5	3.49%	3.93%
# 6	3.95%	3.42%
# 7	3.60%	3.11%
# 8	3.13%	2.73%
# 9	2.88%	2.50%
# 10	2.85%	2.25%
Total	49.76%	46.37%

However, it has to be noted that since data from SelectCorpus is not randomly sampled as discussed earlier, the actual number will be different from 2.72%, but the upper boundary would be 4.2% for RandomCorpus. More experiments will be needed in future to estimate the recall improvement using the formula "Estimated Recall" presented earlier, based on randomly selected data.

V. Conclusion and Future Work

We have performed a fairly comprehensive quantitative analysis of how emoticons are used in social media, how reliable it is to use emoticon alone as a sentiment trigger, and what could be the recall contribution of emoticon in a brand focused sentiment extraction system. We demonstrate that emoticon alone without considering other linguistic evidence is not sufficient to dictate a sentiment toward an object. Our study shows that its recall contribution in our context is not significant, but it is meaningful to help enhance linguistic and/or domain specific sentiment extraction. Ongoing and future work will be focused on what other linguistic factors could be used together with emoticon to improve precision and recall, such as sentence length, and other emotion-related lexical items including many weak emotion words. In addition, the study of language-dependent part of emoticons, especially for the Eastern vs. Western distinction, is also interesting and would be beneficial to our multilingual program.

VI. Acknowledgment

Thanks to our Quality Assurance team for preparing large testing corpora in which we are enabled to conduct the data analysis.

References

[1] Davidov, D., Tsur, O., A. Rappoport: Enhanced sentiment learning using twitter hashtags and smileys. Proceedings of COLING 2010: Poster Volume, 241–249, Beijing (2010)

[2] Hogenboom1, A., Bal1, D., Frasincar1, F., Bal1, M., Jong, F., Kaymak, U.: Exploiting Emoticons in Sentiment Analysis. Proceedings of the 28th Annual ACM Symposium on Applied Computing: 703-710 (2013)

[3] Liu, K., Li, W., Guo, M.: Emoticon Smoothed Language Models for Twitter Sentiment Analysis. Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (2012)

[4] Pang, B, Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2(1): 1-135 (2008)

[5] Ptaszynski, M., Rzepka, R., Araki, K., Momouchi, Y.: Research on Emoticons: Review of the Field and Proposal of Research Framework. 言語処理学会 第 17 回年次大会 発表論文集 (2011)

[6] Read, J.: Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification . Proceedings of the ACL Student Research Workshop: 43–48 (2005)

[7] Strapparava, C. , Mihalcea, R.: Learning to identify emotions in text. SAC (2008)

[8] Zhao, J., Dong, L., Wu, J., Xu, K.: MoodLens: an emoticon-based sentiment analysis system for chinese tweets. KDD 2012: 1528-1531 (2012)