Precision Impact of Emoticons for Social Media Sentiment Analysis

(Research Report for the Summer Intern Project with Netbase Solutions Inc.)

Tanya Lee
Cupertino High School
Cupertino, CA 95014
tanyalee@berkeley.edu

June 22, 2014

***Abstract* —** Traditionally, businesses spend tons of money performing customer surveys on brands. In the big data age, such insights including both consumer compliments and complaints can be automatically collected from social media such as twitter. The enabling technology is sentiment analysis. It is observed that social media often uses emoticons such as :) and :(mixed with text to express sentiment. This research presents a novel study of how emoticons can help sentiment analysis precision. Data analysis shows that emoticons alone cannot determine sentiments towards a brand and they can only be used together with other evidence. Further study has discovered a use of emoticons as counter evidence to block glaring errors in sentiment analysis.

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

I. What Inspired Me for This Research

1.1. How it started

It all started with social media. Like many Facebook fans of my age, a significant part of my life was spent on social media. As we take knowledge from the infinite pool of cyberspace, cyberspace, in return, instilled appalling social habits, and my social interactions simply became competitions of who can glue eyes to their screen the most. Consequently, for me (and my 819 friends), my speech patterns rescinded to a level akin to "OMG LOL I have to get to class". I lived in social media, knowing it inside and out.

In sophomore year, I had an opportunity to put my social media expertise to some use as a paid summer intern at a Silicon Valley startup that automatically tracks public opinions and sentiments from social media. Their system uses natural language technology to do sentiment analysis of consumer opinions about a brand or topic. Once a query is given, the system looks through terabytes of social media data to mine public opinions and sentiment, displaying the insights through infographics such as word clouds, tables and charts. For example, searching "Miley Cyrus" on the day of her VMA performance would show her tumbling public appeal (net sentiment -19%) despite rising number of mentions.

On my first day as an intern, I was thrown into the business casual wild of Silicon Valley, where software engineers join hands with linguists to build software that reads human languages. My initial job was to incorporate social media jargon into the system, especially the emotional expressions from Urban Dictionary. I was also assigned to test entries from Facebook fan pages, sorting positive sentiment from negative. I soon immersed myself into my work routine but noticed that the system always disregarded smiley faces (emoticons) as these are things beyond words, extra-linguistic symbols. As visible representations of emotion, isn't that a missed opportunity to help gather sentiment? A happy face like :) usually denotes a positive tone of sentiment while a sad face :(a negative tone. Intuitively, it should help the system for the purpose of sentiment analysis.

Opening my Baskin Robbins Facebook fan page, I gathered all the emoticons on the page to study the underlying structures of emoticons. Emoticons are often made of different facial parts, such as smiling/crying eyes, nose and mouth. The combinations are many but there is a fairly clear pattern structure behind them. I went ahead to start coding the patterns for identifying emoticons of all possible combinations using *regex*-like proprietary formalism. Social media users, especially teenagers who use smiley faces heavily, can be creative in using and recognizing them. Pattern matching proved to be an adequate approach to automatically identifying emoticons.

My creation was initially met with confused stares, "so many strange symbols". To many of my colleagues, the idea of identifying these symbols was unplanned, and so I was able to bring insight to people I respect. My mentor looked intrigued as he scrolled through my work. From there, he guided my research direction to the study of the emoticon distribution and its impact on sentiment precision with novel findings on brand sentiments. It marked the start of my long journey of three-summer dedicated research on emoticons. It made the parser smarter at reading people's emotions from smiley faces as well as from words.

The research evolved in stages. The first stage was, of course, automatic emoticon identification. The implementation uses both an emoticon lexicon and the pattern matching device for rules on the proprietary NLP platform. It reached almost perfect quality of precision and recall (*recall* is technical metric for coverage). The second stage was to study emoticon's distribution and its impact on brand-oriented sentiment analysis. This involved statistical data analysis from large corpora. The statistical data analysis and benchmarks are based on randomly collected testing corpora, using scripting language and commands. This part of research involves data analysis that lead to new findings such as the use of emoticons as counter evidence to block errors in sentiments. These findings were further tested and integrated into the system by the team, with follow-up projects on emoticon recall research. The nature of this applied research determines its applicable benefits to enhance real life system's data quality. The benefits are already felt by the marketing department as well as clients.

With guidance from my mentor, I was able to connect with my colleagues who integrated my research into the final product. This unique experience of seeing my work eventually enter a real life app used by Fortune 500 clients offers much of the satisfaction and motivation for further pursuing applied science in artificial intelligence as one of my major career goals although at this point I cannot pre-exclude other equally fascinating choices that might be available to me such as graphic design, marketing, or Pizza Hut Delivery (PhD).

1.2. Looking ahead

The internship allowed me to be introduced to the world of computational linguistics. I was working with a bunch of people generations above me. I was fascinated by my colleagues' conversations of "tokenization" or "parsing", and felt lucky to have senior members train me as their youngest trainee. As I became more involved with my research, my interest in artificial intelligence and computational linguistics grew. This is an interdisciplinary area which involves math (mainly statistics), logic, computer algorithm (automata) and linguistics.

The world is all material and information, and I became intrigued by the information side. One key problem with today's society is information overload. For example, Twitter by itself generates 340 million tweets daily, 3,935 tweets per second. It is impossible for humans to manually collect the information that they want from such an expansive sea, like finding a needle in the middle of hell. Google search can help, but only to a limited extent because they treat language as bags of unrelated key words, without understanding the underlying text.

Today, Artificial Intelligence has already escaped the realms of science fiction in linguistic products such as Siri and Google Translate. It is an enabling technology, which is bound to play an increasingly large role in this age of information overload. By combining linguistics with computer programs, we can build the language ability of a computer to model human intelligence and make the computer do all the reading for us. From that point, all kinds of applications can be built for various benefits, such as answering questions, summarizing information, or translating foreign languages (machine translation). Such progress will greatly simplify our lives and help us make more informed decisions.

1.3. Role of Math

A significant math class I took that influenced my study is AP statistics. The mainstream of natural language processing is statistical. Taking statistics opens a door that allows me to pursue the mainstream study in computational linguistics. Statistical study of the data is also necessary for uncovering the balancing impact of different pattern features on the quality of results. That is the key of my emoticon research. I was also trained in symbolic processing during the internship, which mainly used pattern matching and grammar engineering for sentiment analysis. The underlying mathematics of pattern matching mechanism is finite state automata (FSA) in Chomsky's formal language theory. The skills in writing *regex*-like (regular expression) patterns are most effective in identifying thousands of emoticon variants. Statistics allows us to see the whole forest, while symbolic processing allows us to dive into each tree.

II. RESEARCH INTRODUCTION

Traditionally, businesses use manual surveys to collect consumer opinions and sentiments on their brands in order to improve their products and sales. The rapidly growing amount of social data from online networks such as Twitter and Facebook has led businesses to seek methods of mining online consumer opinions and sentiments. For instance, Twitter has over 500 million registered users as of 2012, generating over 340 million tweets daily, equivalent to 3,935 tweets per second. Such big data contain enormous amounts of customer voice that is invaluable to the leading brands. People are increasingly sharing their opinions on products and services on social networks. Recent estimates indicate that on average in every three blog posts and one in every five tweets involve comments on products, services, or brands, often comparing it with other brands in the same category (Hogenboom et al. 2013). Such brand-oriented information allows businesses to keep track of consumer sentiments, which allows the management to make faster decisions in adjusting their product and marketing plans.

The majority of sentiment analysis systems are based on a statistical keyword model for classifying a social media post as thumbs-up or thumbs-down. Typical examples of social media data are movie or product reviews. When trained on domain data, sentiment classification generally achieves over 80% accuracy as reported in Pang and Lee's study (2008). However, today's dominant social media platform is mobile, the posts such as tweets tend to be shorter than review data. The statistical sentiment classification models are challenged for handling tweets as the keyword evidence is often insufficient in short messages.

The wide use of emoticons in social media, especially for the younger generation, leads to additional evidence beyond keywords. Social media as a sub-language involves significant amount of emoticons mixed with text to show the emotions of the poster. Most emoticons, e.g. :=), are not natural language words, but textual symbols using mainly a combination of punctuation marks. The main use of an emoticon is to express a person's feeling or mood, in categories such as *love* (positive), *hate* (negative), or *don't care* (neutral). In light of this, a natural language system built for sentiment analysis is expected to take advantage of this special extra-linguistic phenomenon, which is typically not in the scope of the grammar or vocabulary research of a language.

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

- Provide a detailed statistical analysis of how emoticons are used in social media
- Provide an evaluation of emoticon's roles in the brand-oriented post-level sentiment analysis by calculating its precision impact on system performance

Unlike the mainstream sentiment analysis based on statistical models using machine learning, Netbase has built a rule-based high precision sentiment analysis system based on a parser. This system is designed to address the major challenge of machine learning systems, namely, brand focused post-level sentiment classification for short messages. The research on emoticons reported in this study is motivated by the need to enhance the data quality of the Netbase system for consumer sentiment analysis of brands. The analysis and experiments also serve to enlighten other researchers with a 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 alone are reliable sentiment indicators and therefore use them to annotate a training corpus for sentiment analysis.

The remainder of the paper is organized as follows. Section III reviews related work in using emoticons as clues for sentiment analysis. Section IV presents the results of emoticon-related study and experiments, focusing on emoticon's statistical distribution in social media. Section V investigates the impact of emoticons on sentiment precision. The conclusions and future work are summarized in Section VI.

III. RELATED WORK

The popularity of emoticons (e.g. smileys) comes hand in hand with the growth of social networking. They have been extremely popular in social media among the younger generation and seasoned netizens. Despite their prevalence in 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 while others are language dependent. Surveys show that emoticons are the second most important vehicles for expressing emotions in online communication (Ptaszynski et al. 2011).

In the context of NLP (Natural Language Processing), the use of emoticons has attracted machine learning researchers for the sentiment classification. Emoticons 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 feature 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; Hogenbooml et al. 2013).

Ptaszynski et al. (2011) proposes that emoticon research consists of six lines of tasks: (1) detection; (2) extraction; (3) parsing; (4) semantic analysis; (5) generation; (6) evaluation. Our work on international emoticons as used in English involves all these

aspects except (5), with emphasis on (6). The work involved in (5) goes beyond the analysis phase, not within the scope of this study. The work in (1) through (4) is analysis work, similar to morphology analysis in linguistics.

Sentiments are not meaningful unless they are associated with an object, such as a brand or product. However, not much has been done in evaluating the contributions of emoticons in the context of brand-oriented sentiment analysis. This is the major motivation and value for this study.

IV. Experiments and Results

This section mainly answers two questions:

- How are emoticons generally used and statistically distributed in social media?
- What is the precision impact of emoticons on sentiment analysis?

These are questions that can help drive the design and development of an emoticon-inclusive sentiment system. It needs to be noted that in our approach to brand-focused, post-level sentiment analysis, the sentiment analysis must be targeted at a specific object (usually a brand).

4.1. Background of this study

This study is conducted in the context of the Netbase Consumer Insight system with the intent to help enhance the data quality of the brand-focused sentiment analysis. The post-level sentiment analysis is supported by natural language parsing. Each incoming social media post is analyzed by the Netbase parser, ending with a parse tree as results of automatic grammatical analysis, on top of which sentiment analysis is built. For instance, *I like the camera of iPhone.* From this sentence, the system is able to extract information as follows:

Sentiment: *positive*
Object: *iPhone*
Aspect: *camera*

The extracted sentiment must be associated with a topic or an object, such as Object: *iPhone* in this case. This is because identifying sentiments without an object is of little value in sentiment analysis of customers' insights. A third-party anonymous annotation service is used to benchmark the precision of sentiment results. The existing system reaches 87% precision in sentiment analysis.

4.2. Emoticon identification

Identification of emoticons and labeling them as positive or negative (and sometimes neutral if needed) are a precondition for investigating and making use of emoticons in sentiment analysis. An approach of combining emoticon lexicon and emoticon pattern rules is taken to do this job with almost perfect precision and recall.

Our emoticon lexicon houses over 1,000 emoticon symbols, which are manually collected and reviewed. Each entry is marked as positive, negative or neutral by hand. Fortunately, most of the western emoticon symbols in the lexicon are not ambiguous and the identification of them from text is a simple matter of lexical look-up.

A number of specific pattern rules are also implemented to disambiguate a handful of ambiguous trouble makers through an emoticon parser based on pattern matching using the Netbase NLP-oriented finite-state language. The precision and recall of emoticon identification through the combination of emoticon lexicon lookup plus emoticon parsing are almost perfect after several rounds of tune-up and debugging.

4.3. Emoticon statistics in social media

This sub-section investigates the general statistics of emoticons as used in brand-oriented social media data.

First, a representative corpus is collected to help analyze emoticon's various frequencies and distribution. This corpus is made up of randomly collected English social media data, altogether 440,000 social media posts, including tweets, Facebook posts and forum posts. For our purpose of brand-oriented sentiment analysis, the only requirement for the collection is that each post selected must contain at least one term in the represented brand set = {Pepsi, Walmart, Groupon, Taco Bell, Amazon}. For each brand, there are roughly 88,000 posts.

Table I lists the emoticon richness distribution. The emoticon richness, calculated as 3.27%, is defined as the ratio of the number of posts containing at least one emoticon to the total number of posts in the corpus. This metric informs us of the overall frequency of emoticons as used in social media and hence the suggestion of their maximum possible impact on a corpus.

TABLE I. EMOTICON DISTRIBUTION

BRAND	TOTAL POSTS	# OF POSTS WITH EMOTICON	PERCENT
PEPSI	84,483	5,006	5.93%
TACO BELL	85,790	3,727	4.34%
WALMART	94,339	3,067	3.25%
GROUPON	82,205	900	1.09%
AMAZON	92,223	1,653	1.79%
TOTAL	439,040	14,353	3.27%

Table II displays the distribution of Positive emoticons and Negative emoticons (excluding the 1,781 neutral emoticons as they are not involved in sentiment analysis). It is shown that the positive emoticons are used much more frequently (about 70%) than the negative emoticons (about 30%). This does not necessarily imply that social media involve more positive comments than complaints. One explanation could be that people tend to use the simple popular positive emoticons such as :) very heavily (Table III), not necessarily directed towards a specific object, but more to make the post light, friendly or humorous. The default of emoticon use is found to be a positive smiley face instead of a sad face or a neutral one.

TABLE II. RATIO BETWEEN POSITIVE EMOTICONS AND NEGATIVE EMOTICONS

BRAND	POSTS WITH EMO	POSITIVE EMO		NEGATIVE EMO		POSITIVE/NEGATIVE MIX	
PEPSI	4,559	3,058	67.2%	1,469	32.3%	22	0.5%
TACO BELL	3,112	2,141	68.8%	959	30.8%	12	0.4%
WALMART	2,661	1,764	66.3%	887	33.3%	10	0.4%
GROUPON	759	619	81.6%	138	18.2%	2	0.3%
AMAZON	1,491	1,210	81.2%	275	18.4%	6	0.4%
TOTAL	12,572	8,792	69.9%	3,728	29.7%	52	0.4%

Given the over 1,000 western emoticons in our emoticon lexicon, we are also 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 random corpus. The top 10 used emoticons are as follows in Table III.

How do people actually use these emoticons in online chatting and posts? One way to get an insight of this is to count the unique emoticons used by a particular user. In addition to the text body, other meta information of the text, including user id, is also available in the content store. Thanks to this, we are able to find out the fact regarding users' usage of unique emoticons from the corpus, shown in Table IV. Despite all the varieties of emotions in the web, majority of users (95+%) only use a couple of unique forms of emotions.

TABLE III. TOP 10 EMOTICONS

Emoticon	Frequency	Percent	Polarity
:)	4,833	33.67%	P: happy
<3	1,670	11.64%	P: love
:(	1,406	9.80%	N: sad
:D	1,270	8.85%	P: laugh
(:	1,174	8.18%	P: happy
;)	1,160	8.08%	P: wink
:-)	727	5.07%	P: happy
:P	719	5.01%	P: kidding
XD	554	3.86%	P: grin
=)	339	2.36%	P: happy
Total	9,019	62.84%	

TABLE IV. USER COUNTS OF UNIQUE EMOTICON

Unique Emoticons	# of Users	Percent
1	238,000	95.02%
2	6,325	2.53%
3	4,350	1.74%
4	1,265	0.55%
5	400	0.16%

V. EMOTICON'S IMPACT ON PRECISION

Precision and recall are key measures that are used to benchmark the quality of a sentiment analysis system. This section studies how the introduction of emoticons can help enhance the precision in brand-oriented sentiment analysis. Based on the community standards, precision used in this study is defined as: correct / (correct + wrong + spurious).

Human annotation is necessary to support in-depth sentiment analysis study. From the random corpus of 440,000 posts as used above, 2,000 posts are randomly selected for human annotation, based on the 14,353 posts that carry positive or negative emoticons only (i.e. excluding posts of neutral emoticons and the few positive-negative mixed posts). Each post is annotated by two annotators; and their agreement is used as the gold standard. Each post is annotated in two ways: (i) the general tone of the post as positive, negative or neutral, irrespective of an object; (ii) a sentiment choice of positive, negative, or neutral towards the corresponding brand associated with the sentiment. Of course, our focus is on the gold standard as defined in (ii), but it is insightful to perform a comparison of benchmarks between (i) and (ii), presented below in Table V and Table VI respectively.

TABLE V. EMOTICON PRECISION OF POST TONE IRRESPECTIVE OF BRAND

Human annotation (i)	Positive emoticon	Negative emoticon
Positive	1,230 (correct)	10 (wrong)
Neutral	384 (spurious)	38 (spurious)
Negative	16 (wrong)	322 (correct)

So the overall precision for sentiment tone irrespective of brand is: (1230+322)/2000 =77.6%

77.6% is decent for precision, showing that emoticons are indeed important clues for determining the sentiment tone of a post. But 77.6% is apparently not good enough to be used to collect data as gold standard for training a sentiment analysis system.

More importantly than just the general tone of the post, emoticon does not necessarily mean that there is a positive/negative emotion geared towards a brand mentioned in the post. For instance, in the post *"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 indication that the sentiment is towards the brand "Walmart". This shows that the sentiment as representing a general tone of a post may be different from the sentiment as associated with a specific brand. Only the latter is the brand-oriented sentiment analysis required by the businesses in understanding customer insights, as shown in Table VI.

TABLE VI. EMOTICON PRECISION OF BRAND-ORIENTED SENTIMENT

Human annotation (ii)	Positive emoticon	Negative emoticon
Positive	458 (correct)	76 (wrong)
Neutral	1,104 (spurious)	186 (spurious)
Negative	64 (wrong)	112 (correct)

So the overall precision for brand-oriented sentiment is: (458+112)/(458+112+1104+186+76+64) =28.5%

Table VI shows that there is a sharp drop, for almost 50 percentage points (77.6–28.5), in precision from the object-free sentiment analysis to the desired brand-oriented sentiment analysis. A precision of less than 30% is really too low for any uses. This means that for brand-oriented sentiments, emoticons alone are too weak to be useful. Therefore the future research should explore combining emoticon evidence with other evidence in enhancing the sentiment analysis.

Note that there is a large number (1104+186) of spurious cases where the human judges tag as neutral. As discussed before, positive emoticons are default ways of making the posts light, it does not have to carry strong sentiment towards a brand, especially for the simple common smiley faces like :) or :D. Hence, there are lots of gray area cases which are neutral but associated with positive emoticons, for example,

@tonymophoto It was American Balloons in Land O Lakes. I bought a Groupon for it. :).

Did you know there are Amazon links to all the books we discuss at http://t.co/rQZ8Uo0E? Buying thru them supports the show! :D.

As for spurious cases of negative emoticon, it seems that some simple negative emoticons could mean a weak "sorry" or "that's all I can do", while the post is still fairly neutral, e.g.

to be honest i can't really taste a difference between Coke and Pepsi ;-;.

I need to go to Walmart to buy my letters and my mirrors. :(.

The last example is a need statement which demonstrates some peculiar properties in terms of sentiment analysis. A need statement often goes hand-in-hand with negative emoticons (i.e. sad that there is a need not met) while the need statement is neutral from post sentiment perspective and it is often regarded as positive from brand perspective (because a needed product is a positive mention: e.g. *I badly need a new iPhone*).

As expected, wrong cases are very few (1%-3%) when the emoticon polarity mismatches a human tag (i.e. a positive emoticon is used with a negative post or vice versa). It can be considered as random noise, caused by mis-typing, and in some cases, it might involve a degree of sarcasm

Absolutely pissed that Taco Bell isn't open.:-).

RT @Jyoti_More: @rehannaaaaaa It's like Taco Bell but it seems so much better ;(.

RT @AfsahB: When drinking Pepsi/Coke out of a thela was so cool :'(. #PuranaPakistan (this looks like sarcasm)

Table VII benchmarks the precision of the Netbase sentiment analysis system in this emoticon-involved test corpus. Remember that the Crowd-source third-party benchmarks of Netbase system is 87% (Section 4.1) which is greater than 79.9% on this relatively small corpus. Eight percentage points are statistically meaningful, so a reasonable explanation is called for. As the testing corpora are all collected using brands as trigger words from the same social media sources, the only major difference is that posts in this study all involve emotion mentions, and the drop of precision seems to indicate that the Netbase system's data quality on emoticon-involved data is not as good as the quality on random data. This is fairly understandable as posts involving emoticons tend to be more casual, often involving more social media jargons and ungrammatical fragments (degraded text). Therefore, adequately making use of the emoticon evidence in sentiment analysis in such data is more important.

TABLE VII. SYSTEM PRECISION OF BRAND-ORIENTED SENTIMENT

Human annotation (ii)	Positive emoticon	Negative emoticon
Positive	580 (correct)	24 (wrong)
Neutral	86 (spurious)	26 (spurious)
Negative	52 (wrong)	168 (correct)

So the overall precision is: (580+168)/(580+168+86+26+52+24) =79.9%

There is another more significant finding in interpreting the relevant data and analyses. Although emoticons alone may not be reliable evidence for sentiment analysis and it requires more research on balancing them with other evidence, emoticons seem to be very good indicators as **counter-evidence** to block incorrect sentiment classification. This is especially meaningful as all incorrect classification in tagging positive posts as negative, or the other way around, involves embarrassing glaring errors. This is unlike the distinction between neutral and strong sentiments (whether positive or negative) where the mismatching between the system's tagging and the human judgment is not a big issue as there is a possible gray area involved. But the sentiment polarity error is a fatal mistake made in black and white area where human judges have no problems but a system often has trouble in sentiment identification (e.g. in tricky cases such as double negation). Extra-linguistic evidence such as emoticons can save these critical cases with high confidence. The implementation along this line is straightforward: no sentiment classification is allowed to be in conflict with the polarity of the emoticons. Based on the careful data analysis of 140 cases (i.e. all the wrong cases in Table VI), three restrictions apply as exceptions to the above rule: (i) the default positive emoticon *:)* should not be involved as a sentiment tagging blocker as it tends to be over-used; (ii) the negative emoticon should not be a blocker in a need statement, such as *I badly need a new iPhone :(* where the system tags *iPhone* as a positive mention correctly; (iii) an emoticon should not block sentiment tagging in a preference statement (e.g. *rather A than B*), a mixed case (e.g. *love A but hate B*), or a long post (thresholds can be set at 10-15 words in length). These restrictions are not difficult to enforce in the existing Netbase system because the required devices are already available in the system. As a result, some eye-catching, otherwise difficult-to-catch precision errors will be avoided when the above heuristic is in effect.

VI. CONCLUSION AND FUTURE WORK

A fairly comprehensive quantitative analysis is conducted on how emoticons are used in social media, how reliable it is to use emoticons as a sentiment indicator in a brand-oriented sentiment analysis system. It shows that emoticons alone without considering other linguistic evidence are not sufficient to dictate a sentiment towards an object. Ongoing and future work will be focused on what other linguistic factors could be used together with emoticons to improve precision, including sentence length and other emotion-related lexical items such as many weak emotion words (strong emotion words do not need further evidence from emoticons). The recall contribution of emoticons in the context of brand-oriented sentiment analysis is an interesting and equally important research topic that naturally follows this work. In addition, the study of the language-dependent part of emoticons, especially for the Eastern vs. Western distinction, is also interesting and would be beneficial to the multilingual program.

REFERENCES

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

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

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

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

[5] Ptaszynski, M., R. Rzepka, K. Araki and Y. Momouchi 2011. "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 年 3 月)

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

[7] Yang, C., K. H. Lin and H. Chen 2007. "Building Emoticon Lexicon from AWeblog Corpora", Proceedings of the ACL 2007. Demo and Poster Sessions: 133-136

[8] Yu, H. and V. Hatzivassiloglou 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. EMNLP

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

Glossary

Consumer Insight: consumers' comments or feedback on a brand or product/service, including both compliments and complaints.

Emoticon: graphic symbols (smileys) made of punctuations or other characters often used in social media for expressing the poster's emotion/sentiments.

Machine Learning: a school of research in Artificial Intelligence and Natural Language Processing based on statistical models.

Natural Language Processing: an area of research in using computers to process human languages for understanding its content.

Parsing: automated grammatical analysis of human language sentences, similar to sentence diagramming taught in schools.

Precision: a measure used to benchmark how accurate a system (often software) is in doing its job.

Recall: a measure used to benchmark the coverage of a system (often software) in doing its job.

Sentiment Analysis: classification of a chunk of text as positive, negative or neutral.