Emotion Tokens: Bridging the Gap among Multilingual Twitter Sentiment Analysis[*]

Anqi Cui, Min Zhang, Yiqun Liu, and Shaoping Ma

State Key Laboratory of Intelligent Technology and Systems,
Tsinghua National Laboratory for Information Science and Technology,
Dept. of Computer Science and Technology, Tsinghua Univ., Beijing 100084, China
cuianqi@gmail.com, {z-m,yiqunliu,msp}@tsinghua.edu.cn

Abstract. Twitter is a microblogging service where worldwide users publish their feelings. However, sentiment analysis for Twitter messages (tweets) is regarded as a challenging problem because tweets are short and informal. In this paper, we focus on this problem by the analysis of emotion tokens, including emotion symbols (e.g. emoticons), irregular forms of words and combined punctuations. According to our observation on five million tweets, these emotion tokens are commonly used (0.47 emotion tokens per tweet). They directly express one's emotion regardless of his language; hence become a useful signal for sentiment analysis on multilingual tweets. Firstly, emotion tokens are extracted automatically from tweets. Secondly, a graph propagation algorithm is proposed to label the tokens' polarities. Finally, a multilingual sentiment analysis algorithm is introduced. Comparative evaluations are conducted among semantic lexicon based approach and some state-of-the-art Twitter sentiment analysis Web services, both on English and non-English tweets. Experimental results show effectiveness of the proposed algorithms.

Keywords: Multilingual sentiment analysis, Twitter sentiment analysis, Emotion token, Sentiment lexicon, Network informal language.

1 Introduction

Nowadays millions of users publish short messages on Twitter. It is widely spread all over the world and becomes a rich resource of texts in many different languages. Twitter's messages (*tweets*) are full of opinions and emotions, thus sentiment analysis for tweets is important for information spreading and marketing. However, this is more difficult than traditional text analysis.

Tweets are limited with no more than 140 characters and are usually composed on mobile devices, hence people often use irregular expressions both for convinience and to save room for more words. Emotion tokens (including *emotion symbols*, *irregular forms of words* and *combined punctuations*) are usually seen

[*] Supported by Natural Science Foundation (60736044, 60903107, 61073071) and Research Fund for the Doctoral Program of Higher Education of China (20090002120005). This work has been done at Tsinghua-NUS NExT Search Centre.

M.V.M. Salem et al. (Eds.): AIRS 2011, LNCS 7097, pp. 238–249, 2011.

in tweets, such as "*– Me tienes olvidada **:(** – (**-.-** otra vez esta...) Disculpa es que estaba dormido. – **AAAAH** ok, ¿qué harás ahora? – Dormir mucho más.*" (Spanish: "*– I have forgotten **:(** – (**-.-** again this...) Sorry is that he was asleep. – **AAAAH** ok, what will you do now? – Sleep more.*") Based on our observation (see section 3.2), there are about 0.47 emotion tokens per tweet; about one third tweets contain at least one emotion token. Emotion token is one of the most remarkable features of Internet text. It strongly expresses the feelings of the author and is often utilized across languages. Thus, emotion tokens are helpful for multilingual Twitter sentiment analysis, to determine if a tweet expresses a positive or a negative feeling, no matter what language the author uses.

Although the emotion tokens have been studied previously, they are usually considered as annotation of the texts and are chosen manually [6]. Different from these studies, we automatically extract different types of emotion tokens and use a propagation algorithm to label their polarities by few "seed" tokens. Therefore, many different tokens and their scores are discovered to build a sentiment lexicon which helps multilingual Twitter sentiment analysis.

The highlight of our work is: Different types of emotion tokens are extracted automatically, without considering the semantic information; their sentiment polarities are labeled with an unsupervised propagation algorithm. The sentiment lexicon built based on the tokens works as a bridge over the gap among different languages, while most state-of-the-art Twitter sentiment analysis approaches only deal with English tweets. In addition, the corpus for building the lexicon is independent of time which is practical and feasible for real-world applications.

This paper is organized as follows. Related work on sentiment analysis and Twitter is introduced in Sec. 2. Emotion tokens and their characteristics are analyzed in Sec. 3. In Sec. 4 and 5, algorithms for sentiment lexicon construction and sentiment analysis are proposed respectively. In Sec. 6, the algorithms are evaluated. In the last section, conclusions and future work are addressed.

2 Related Work

Sentiment analysis plays an important role with the growing of user-generated-content services. In traditional studies, most researchers build statistical models for sentiment and affect analyses, where semantic information is highly considered as features [17,20]. These models require annotated corpus, which is often limited for online texts. Alternatively, manually built sentiment lexicons can be used as a useful resource [2,22]. Linguistic information (part-of-speech tags, syntactic information) is used for rule-based approaches [18]. Even though the feature words can be extracted with automatic algorithms [16], the semantic differences among languages limit multilingual analysis.

An intuitive idea for multilingual sentiment analysis is to translate languages into a well-studied language (e.g. English); hence traditional methods can be applied. Previous studies on news and blogs work on sentence level [5] or word level [11]. Cross-language dictionaries work as bridges between different languages [9]. Obviously, without language techniques, these methods do not work. Some machine learning models may be independent of languages, but training requires multilingual annotations [7,19] or is based on machine translation [3].

Emoticons and irregular words, however, are commonly seen in Internet texts of many languages. Emoticons are considered as annotations since they directly express the one's attitude [6]. Unfortunately, they have various forms; most studies manually choose some smileys (e.g. "*:-)*") as labels [19]. They fail to consider many other figures such as "*<3*" (heart, means *love*). In Twitter, we need to discover more possible emoticons with different forms, since they are independent of languages and are helpful for multilingual analysis. On the other hand, the irregular words are usually seen when people wish to save keystrokes, or the length of message is limited. We focus on the emphasized spelling, i.e. the repeating of consecutive letters in a word (e.g. "***nooooo*** *WTF everyone left me*").

Twitter is a popular research topic nowadays. Besides its network characteristics [15], social impacts reflected by Twitter sentiment are also of interests, such as word-of-mouth branding [13]. Other work follows the trends of sentiment [8]. As mentioned before, smileys are usually considered as annotations [6,19]. Most studies use linguistic rules or supervised learning to help sentiment analysis [14], which is difficult to be generalized into multiple languages.

There are websites that provide sentiment detection of Twitter. However, a study on the comparison of these results concludes that they contain much noise and lack precision [4]. The *twitrratr* site (http://twitrratr.com/) builds lists of positive and negative keywords and classifies the sentiment of tweets based on matching. The *twittersentiment* site (http://twittersentiment.appspot.com/) uses distant supervision to classify the sentiment of Twitter messages [12]. Although some smileys are used to collect training data as labels, emoticons are removed in their classification. We compare our results with these two websites.

To sum up, we notice that the traditional sentiment analysis methods are in shortage on Twitter and are limited to a specific language. To achieve a better multilingual Twitter sentiment analysis, we consider the emotion tokens as a bridge over the gap among languages.

3 Study on Emotion Tokens in Tweets

3.1 Types of Emotion Tokens

The three types of emotion tokens are listed in Table 1. They express emotions and cover most of the emotional informal words on the Internet.

Note that some of the repeating letters words may be relevant to language. However, based on our observation, many of them are onomatopoeic words (Table 2). Therefore, they can be simply considered as another type of *emoticons.*

3.2 Characteristics of Emotion Tokens in Twitter

In this paper, we use Stanford's SNAP data (http://snap.stanford.edu/data/) which contains more than 400 million tweets in over six months.

Table 1. Types of emotion tokens

Type	Definition	Example	Explanation
Emotion symbols	Every combination of symbols with alphabets and numbers. This type is an extended set of the traditional *emoticons*, since they are not limited to "faces".	***\o/*** *Domingo + Canjica quentinha = tudo de bom!* ***:D*** (In English: ***\o/*** *Sunday + warm Canjica = all the best!* ***:D***)	This Portuguese tweet contains "\o/" (man raising arms, cheering) and ":D" (laughing face), telling us the author is happy.
Repeating punctuations	The repeating (or combination) of the exclamation mark (!) and the question mark (?). They reflect if a tweet contains strong emotions.	*@sinceday1 http://twitpic.com/76hen - No****!!!!*** *Dnt eat it we got 2 eat healthy remember****?!?*** *Smh****!!*** *LOL*	The author doesn't want @sinceday1 to eat candies (the URL) for health reason, and he shake his head. The punctuations enhance his negative opinion of the candies.
Repeating letters	One letter or a group of letters[a]repeat within a real word, because of the author's excitement when typing the word.	*YAY. Presentation done. @Brockaldersley is the* ***besttt****. @BindinDTP thinks he's* ***soooo*** *cool.* ***Hummm hummm*** *hum.*	The author repeats letters in the words "best", "so" and "hum" to emphasize his praise on the user @Brockaldersley.

[a] Most real words contain less than three consecutive same le*tt*ers [10], so we set the minimal repeated times to be three. The repeating patterns are removed to recover the word's origin; hence "lololol" and "looooool" are all reduced to "lol".

Table 2. Word origins of the most frequent (more than 0.1%) repeating letters words

Rank	Origin	Frequency	POS[a]	Example	Explanation
1	ha	36,110	Excl	haha	Laughing
2	so	17,232	Adv	sooo	For emphasizing
3	ah	10,618	Excl	ahhh	Surprise, pleasure, etc.
4	hm	8,025	Onmt	hmm	Thinking or pondering
5	aw	7,572	Excl	awww	Entreaty, commiseration, etc.
6	oh	5,991	Excl	ohhh	Surprise, anger, etc.
7	m	5,370	Onmt	mmm	Similar as hmm

[a] Part-of-speech: Excl: Exclamation. Adv: Adverb. Onmt: Onomatopoeia.

For a simple classification of English and non-English tweets, we examine the Unicode of the characters in each tweet. If all the characters in one tweet are from the Basic Latin or symbols section, the tweet is called a *Basic Latin* tweet. We find that most of them are in English. If some characters are in the Latin extended section, it is called an *Extended Latin* tweet. These tweets are often in Portuguese, Spanish, German, etc. Tweets containing characters beyond these sections (such as Chinese) are not studied in this paper.

From the SNAP dataset of more than 400 million tweets, we uniformly sample five million tweets. Among them, 1,649,503 (33.0%) tweets contain emotion tokens. Their proportions in each character set are shown in Table 3. Shown in Table 4, each tweet with emotion tokens contains about 1.4 tokens. There are about 0.47 emotion tokens (either of the three types) per tweet.

Table 5 lists how many tweets contain such types of emotion tokens. Many types of emotion tokens co-occur in tweets. This is the basic idea of our propagation algorithm. For example,

@xClaire_Cullenx LOVE YOU TOO!!! *gives party hat* xxxxx <3
I'm in a slplendid mood right now =D

The misspelled word "splendid" may lead to a missing of the emotion in semantic-based methods, but the emotion tokens help us identify its sentiment.

Since the tweets are rich of co-occurred emotion tokens, a propagation algorithm based on the co-occurrence can be applied to label the polarities.

4 Multilingual Sentiment Lexicon Construction Based on Graph Propagation

As mentioned in Section 2, sentiment lexicons are commonly used for sentiment analysis. Previous studies take semantic links (WordNet relations, conjunctions, etc.) to build such lexicons [1]. The emotion tokens, however, do not have semantic links between each other. Considering the frequent emotion tokens in tweets, the co-occurred tokens are likely to have similar sentiment. Thus the co-occurrences between words are links for constructing a graph. Then a few initial seeds are used to propagate and discover new tokens.

4.1 Co-occurrence Graph Construction of Emotion Tokens

An undirected graph $G = (V, E)$ is constructed to represent the links of words. Each node $v \in V$ is a word, while each edge $(v_i, v_j) \in E$ represents a co-

Table 3. Proportion of tweets with emotion tokens in different character sets

	Basic Latin	Extended Latin
Total tweets	4,536,590 (90.7%)	266,831 (5.4%)
Tweets w/ emotion tokens	1,494,499 (29.9%)	115,025 (2.3%)

Table 4. Avg. number of emotion tokens per tweet (tweets with tokens)

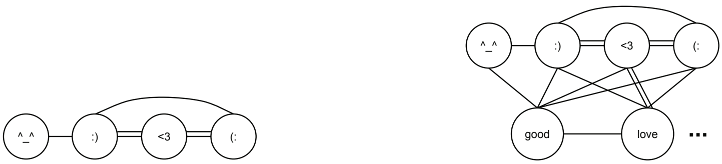
Token type	Basic Latin	Extended Latin
Emot.Symb.	0.79	1.00
Rept.Punc.	0.32	0.23
Rept.Ltr.	0.25	0.20

Table 5. Percentage of each type of emotion tokens in tweets (with tokens)

Token type	Basic Latin	Extended Latin
Emot.Symb.	65.9%	78.1%
Rept.Punc.	26.8%	19.4%
Rept.Ltr.	21.0%	16.9%

Tweets	Emotion tokens	Normal words
ˆ_ˆ yes we did ! RT @JetLife24_7 Me and @tinyy_tee had some good times last summer :)	ˆ_ˆ :)	good, . . .
@JASMINEVILLEGAS Pretty <3 , How are you ? . I wanna see you in the #MyWorldTour on S.America :) . Love you	<3 :)	love, . . .
That's A Good Boyfriend(: RT @CallMeYoshi : I Rather Stay In With My Girlfriend All Night Than Go Out To Party I Love Her To Much <3	(: <3	good, love, . . .
I can't wait for high-school (: gonna be back with my friends <3 on top of that my girls @_BRILove and @_THATSLEX are gonna be there too :)	(: <3 :)	. . .

(a) Example tweets for building the graphs



(b) Using emotion tokens only

(c) Using emotion tokens and normal words

Fig. 1. Example of co-occurrence graphs construction

occurrence between the two words v_i and v_j. The weight w_{ij} of edge (v_i, v_j) is the count of co-occurrence between v_i and v_j. This co-occurrence matrix W (i.e. the adjacent matrix of G) is symmetric. Each diagonal element w_{ii} of the matrix is the frequency of the corresponding v_i in the corpus.

A direct idea is to build such a graph with only the emotion tokens. However, the lexicon built with this graph does not contain any normal words; it could not deal with tweets without any emotion tokens. Therefore, we build the graph on both emotion tokens and normal words. Note that the semantic information of normal words is not considered. We show an example for the graph construction (with the four tweets) in Fig. 1. Only two normal words are shown in the figure for simplicity. Fig. 1(c) is the graph we propagate to build the sentiment lexicon.

To illustrate a clear scale of the graph, 100,000 Basic Latin and Extended Latin tweets from August, 2009 are sampled from the SNAP dataset. Both emotion tokens and normal words are extracted from them. The built graph contains 98,924 vertices (with only 10,390 emotion tokens) and 3,353,873 edges (22,515 among emotion tokens themselves, 314,186 among normal words and the rest are bridges). This undirected graph is extremely sparse (edges take only 0.08%).

4.2 The Propagation and Smoothing Algorithm

Similar to the SentiWordNet [1], we assign a positive score and a negative score to each word, which are calculated separately – the propagation starts with one seed for calculating the positive scores and one for negative scores, respectively.

A general algorithm for label propagation is used. Let x_k be the vector of scores of each word after the k-th iteration. The x_{k+1} is calculated by a co-occurrence matrix W and a bias vector $\mathbf{b}$, formally,

$$x_{k+1} = W \cdot x_k + \mathbf{b} \tag{1}$$

Normalizations in each iteration are applied after the $W \cdot x_k$ and $W \cdot x_k + \mathbf{b}$. The convergence of x_k has been proved [23]. This form of graph propagation is used in many algorithms such as Page-Rank and TrustRank. The bias vector $\mathbf{b}$ is set to the seed vector x_0 to keep the superiority of seeds. Also due to this reason, $W \cdot x_k$ is normalized before adding $\mathbf{b}$ to make them in a same scale. Since the initial x_0 is always added in each iteration, the seed token may have a much higher score than the other thousands of tokens. To smooth the scores into a reasonable scale, we add a logarithm transformation on each word followed by a normalization to the $[0, 1]$ interval. The positive and negative scores are normalized separately. This method maps the scores into a natural distribution.

We choose only one seed to start the propagation (one for positive and one for negative, respectively). Since the graph contains both emotion tokens and normal words, two types of initial seeds are proposed: (1) smileys: ":)" for positive scores, ":(" for negative. (2) good/bad: "good" for positive, "bad" for negative. We build two *SentiLexicons* based on the two types of initial seeds.

With the graph built in Section 4.1, we find many of the scores (positive score minus negative score) of the emotion tokens are labeled correctly after the propagation. Many of the tokens do not have explicit emotions when judged by humans, but may contain hidden emotions brought by the context. We examine the scores by P^+@100 and P^-@100, i.e. the precision of the first 100 tokens with the largest absolute scores. Only 25% and 34% tokens have obvious emotions within the positive and negative ones, respectively. Among them, $P^+@100 = 0.92$ and $P^-@100 = 0.53$. Similarly, $P^+@200 = 0.88$, $P^-@200 = 0.56$, while $P^+@300 = 0.83$ and $P^-@300 = 0.59$. This demonstrates that the tokens are usually propagated with larger positive scores.

5 Sentiment Analysis with Emotion Tokens

The sentiment polarity of a tweet t is determined by both its positive score, $\text{score}^+(t)$ and its negative score, $\text{score}^-(t)$, shown in the equation below. Note the scores of the emotion tokens (v_e) and normal words (v_w) of t are looked up from the built *SentiLexicon*. The $\text{score}^-(t)$ is calculated similarly as $\text{score}^+(t)$.

$$\text{polarity}(t) = \begin{cases} \text{neutral} & \max\{\text{score}^+(t), \text{score}^-(t)\} \le \theta, \text{or score}^+(t) = \text{score}^-(t) \\ \text{positive} & \text{score}^+(t) > \max\{\text{score}^-(t), \theta\} \\ \text{negative} & \text{score}^-(t) > \max\{\text{score}^+(t), \theta\} \end{cases}$$

$$\text{score}^+(t) = \alpha \sum \text{score}^+(v_e) + (1-\alpha) \sum \text{score}^+(v_n) \tag{2}$$

This model is similar as a bag-of-words model. Though simple, it does not involve any linguistic (semantic) information of the sentence. Thus it can be used for multilingual sentiment analysis without much linguistic knowledge.

6 Experiments and Discussions

6.1 Dataset

Tweets for building the *SentiLexicon* and evaluating the algorithms are sampled from the Basic Latin and Extended Latin tweets in the SNAP dataset. Over 99.99% tweets are from June 11th, 2009 to December 31st, 2009, so only the tweets within this period are considered.

During this 204 days period, we pick eight tweets per day for evaluation (no overlap with the tweets building lexicons), including English, Portuguese, Spanish and German tweets (two of each language), which are among the most popular languages in Twitter [21]. Google's Translation API is used to automatically pick out the tweets of a certain language. Each tweet is then given one of the three labels: positive, negative or neutral with two annotators. The third annotator is introduced when there is no majority. If the label is still uncertain, the tweet is discarded (it is difficult even for human judgements). We finally have 449 positive, 211 negative and 553 neutral tweets (total 1,213).

6.2 Strategies of Comparative Evaluations

1. *SentiWordNet*: The baseline method

The SentiWordNet provides positive and negative scores for senses, part-of-speech tags of English words. We use this lexicon with the strategy as referred to its website (http://sentiwordnet.isti.cnr.it/), summing up the scores of each POS tags of a word. With this strategy, the SentiWordNet provides 17,778 positive words (whose positive score is greater than its negative score), 20,350 negative and 1,565 neutral words (with non-zero scores) among 39,693 words. The positive and negative scores of a tweet is the sum of each word. Then a threshold θ is used to classify its sentiment.

2. *twitrratr*

The *twitrratr* provides two lists of positive and negative keywords. We try to match them in a tweet and count their numbers. Similarly, we determine the sentiment of the tweet by the bigger count of positive and negative words. If two numbers are equal, we consider it as a neutral tweet.

3. *twittersentiment*

We retrieve *twittersentiment*'s results of our data from its API. The authors test their method on their own dataset[12]. However, the API they provided makes it comparable for both their and our methods on our dataset.

6.3 Results and Discussions

We examine the lexicon from several aspects. Besides the evaluation on English and non-English tweets, we build the lexicons with different sizes of tweets to see if the size is "the bigger, the better". Moreover, the lexicons are built from different months' tweets in SNAP data, to examine if the lexicon built from a specific month is stable for the analysis on tweets in other months. Our results are all compared with the SentiWordNet baseline and the two websites.

Table 6. Comparative evaluations of algorithms in English tweets

Algorithm	Accu-racy	Positive (166)	Negative (62)	Neutral (111)	$\bar{F_1}$
		F_1 / P / R	F_1 / P / R	F_1 / P / R	
1. *SWN*, $\theta = 0.5$	51.3%	0.591 / 59.9% / 58.4%	0.429 / 38.5% / 48.4%	0.448 / 47.5% / 42.3%	0.489
2. *twitrratr*	49.3%	0.557 / 87.2% / 41.0%	0.276 / 27.9% / 27.4%	0.527 / 41.0% / 73.9%	0.454
3. *twittersentiment*	59.0%	0.648 / 78.0% / 55.4%	0.549 / 70.0% / 45.2%	0.548 / 44.2% / 72.1%	0.582
SentiLexicon	57.8%	0.642 / 65.2% / 63.3%	0.149 / 100.0% / 8.1%	0.606 / 49.7% / 77.5%	0.466

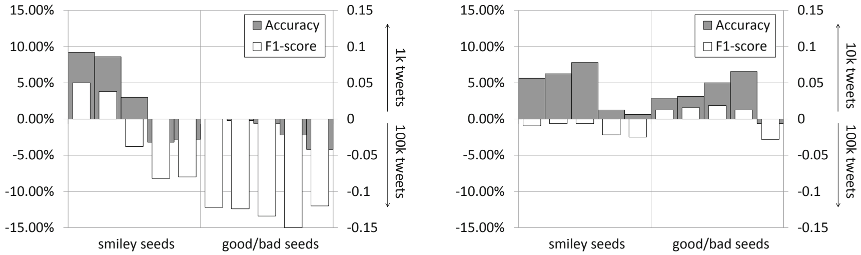
Table 7. Comparative evaluations of algorithms in non-English tweets

Algorithm	Accu-racy	Positive (283)	Negative (149)	Neutral (442)	$\bar{F_1}$
		F_1 / P / R	F_1 / P / R	F_1 / P / R	
1. *SWN*, $\theta = 0.5$	–	–	–	–	–
2. *twitrratr*	52.2%	0.311 / 72.7% / 19.8%	0.130 / 20.9% / 9.4%	0.659 / 52.9% / 87.3%	0.366
3. *twittersentiment*	51.0%	0.218 / 44.1% / 14.5%	0.129 / 52.4% / 7.4%	0.656 / 51.8% / 89.1%	0.334
SentiLexicon	**57.4%**	0.500 / 51.3% / 48.8%	0.123 / 76.9% / 6.7%	0.685 / 59.8% / 80.1%	**0.436**

Parameters and Seeds. The parameter α determines how much the emotion tokens influence the sentiment score, while θ determines the proportion of neutral tweets. We conduct experiments with several combinations of them (both in the $[0, 1]$ interval), based on the *SentiLexicon* built with both smileys and good/bad as seeds from 10,000 tweets in August, 2009 without loss of generality. In general, $\alpha = 1$ is better, i.e. the scores of emotion tokens are weighted with 1 while normal words are weighted with 0. This implies that it is the emotion tokens that affect the sentiment of the tweet. The θ is somehow stable among different α's. Similarly, there are no significant differences between the two types of seeds. For simplicity, we fix $\theta = 0.7$ and smileys as seeds in the following experiments.

Comparative Evaluations on Different Languages. To show our method's efficiency on multilingual tweets, we compare *SentiLexicon* (smileys as seeds, $\alpha = 1$, $\theta = 0.7$) with the three algorithms mentioned above. The lexicons are built with tweets from June to December, respectively. Since the results are similar, we only show the results built with the August tweets. The performances are compared on English tweets and non-English tweets respectively, shown in Table 6 and Table 7. The F_1, P and R under each class stand for F_1-score, Precision and Recall, respectively. The last column $\bar{F_1}$ is the average of three F_1-scores in the three classes. For the performance of *SentiWordNet*, we only list the best one with $\theta = 0.5$. Since this lexicon is for English, it should not be used for non-English sentiment analysis.

These two tables suggest that our method is efficient on multilingual tweets. Most of the F_1-scores, precisions and recalls of the *SentiLexicon* are higher than the current state-of-the-art methods. In English tweets, the recall rate on negative tweets of our method is rather low, which pull down the overall accuracy. We examine that these negative tweets do not contain many strong emotion tokens; hence we classify most of them as neural ones. Another reason is that the tokens are usually have larger positive scores. Therefore, many of the negative tweets are classified as positive ones. We find that in Twitter, there are usually more positive tweets than negative ones (e.g. 449 positive vs. 211 negative ones with our annotation). As a result, the construction of the co-occurrence



(a) 1k vs. 100k

(b) 10k vs. 100k

Fig. 2. Comparison of different datasizes for building lexicons

graph links many tokens with positive words implicitly; the propagation process assigns larger positive scores for many of the tokens. This indicates that the negative tweets may be processed differently from the positive tweets. However, in non-English tweets, our method outperforms the other methods.

Sizes of Datasets for Building Lexicons. With the tweets from August, 2009, three sizes of datasets are extracted: 1,000 (1k), 10,000 (10k) and 100,000 tweets (100k). We compare their performances on the evaluation tweets with the *SentiLexicon* built from them, and draw the differences between accuracies and average F_1-scores on the same scale to compare 1k vs. 100k (Fig. 2(a)) and 10k vs. 100k (Fig. 2(b)). We see that 10k vs. 100k is less different than 1k vs. 100k. Hence we conclude that 1,000 tweets is not sufficient to build a good lexicon, since many tokens may not even appear in such small amount of tweets. On the other hand, the lexicon built with 100,000 tweets does not perform much better than just 10,000 tweets. This finding is helpful for practical use – we do not have to build a very big lexicon. The tokens covered in 10,000 tweets are enough to build a helpful lexicon for sentiment analysis.

Stability of the Lexicons over Time. The lexicons are built with tweets from only one month, hence we propose to examine whether or not the lexicon from one month can work on future months. One strategy is to build the lexicon with tweets in the first month in the dataset (June 2009), and evaluate it on tweets in the succeeding months in the evaluation set. The other strategy is to build with each month (except the last one) and evaluate it on tweets in just the next month (e.g. use June to evaluate July tweets). We also build a lexicon with the first week (June 11th to June 17th, 2009) and evaluate it on July to December tweets. The performances of each strategy are shown in Fig. 3.

The results show the accuracies are all around 50% to 60% in each month's evaluation tweets, and the average F_1-scores are also within 0.4 to 0.5. The performances of the lexicons do not rely on tweets in a specific month or week.

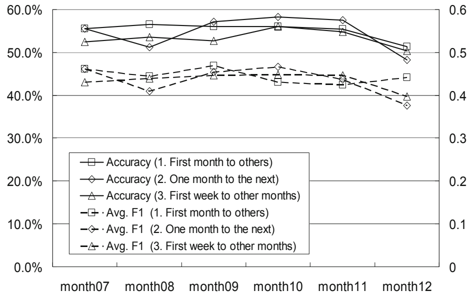


Fig. 3. Stability of *SentiLexicon* with smiley seeds, $\alpha = 1$ and $\theta = 0.7$

This infers that the lexicon is stable along with time. Therefore, we can build a lexicon with the current tweets and use it for future sentiment analysis.

7 Conclusion and Future Work

In this paper, we propose the emotion tokens to help sentiment analysis on multilingual Twitter messages. A graph propagation algorithm with a smoothing method is applied; hence the polarities of the tokens are labeled automatically based on their popular co-occurrences. With this lexicon, we perform a multilingual sentiment analysis for tweets, and achieve a better performance than traditional semantic based approach as well as several Twitter sentiment analysis websites. The comparative evaluations indicate that the emotion tokens are helpful for both English and non-English Twitter sentiment analysis, and are independent with the tweets in different time periods to build the lexicon.

There are also several technical issues we would like to address as future work, such as improving the bag-or-words model for sentiment analysis for higher accuracies, tracking Twitter's sentiment within a longer period and to discover if some tokens have opposite or weak emotions.

Acknowledgments. We would like to thank the anonymous reviewers for their valuable comments. We are deeply grateful to our annotators: Yongfeng Zhang, Qianli Xing, Yin Wang, Kuan Liu, Junwei Miao, Tong Zhu, Qian Wang, Fei Chen, Xiaoguang Wang, Bin Liang and many other participants. The NExT Search Centre is supported by the Singapore National Research Foundation & Interactive Digital Media R&D Program Office, MDA under research grant (WBS:R-252-300-001-490).

References

1. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: LREC, pp. 2200–2204 (2010)

2. Banea, C., Mihalcea, R., Wiebe, J.: A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In: Proc. LREC 2008 (2008)
3. Banea, C., Mihalcea, R., Wiebe, J.: Multilingual subjectivity: are more languages better? In: Proc. 23rd COLING Conference, pp. 28–36 (2010)
4. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Coling 2010: Posters, Beijing, China, pp. 36–44 (2010)
5. Bautin, M., Vijayarenu, L., Skiena, S.: International sentiment analysis for news and blogs. In: Proc. International Conference on Weblogs and Social Media (2008)
6. Bifet, A., Frank, E.: Sentiment Knowledge Discovery in Twitter Streaming Data. In: Pfahringer, B., Holmes, G., Hoffmann, A. (eds.) DS 2010. LNCS, vol. 6332, pp. 1–15. Springer, Heidelberg (2010)
7. Boiy, E., Moens, M.F.: A machine learning approach to sentiment analysis in multilingual web texts. Information Retrieval 12, 526–558 (2009)
8. Bollen, J., Pepe, A., Mao, H.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. arXiv:0911.1583 (2009)
9. Boyd-Graber, J., Resnik, P.: Holistic sentiment analysis across languages: multilingual supervised latent Dirichlet allocation. In: EMNLP 2010, pp. 45–55 (2010)
10. Brody, S., Diakopoulos, N.: Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In: EMNLP 2011, pp. 562–570 (2011)
11. Denecke, K.: Using SentiWordNet for multilingual sentiment analysis. In: IEEE 24th International Conference on Data Engineering Workshop, pp. 507–512 (2008)
12. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. Tech. rep., Stanford CS224N Project (2009)
13. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Micro-blogging as online word of mouth branding. In: CHI 2009, pp. 3859–3864 (2009)
14. Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: Target-dependent twitter sentiment classification. In: Proc. 49th ACL: HLT, vol. 1, pp. 151–160 (2011)
15. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about twitter. In: Proceedings of the First Workshop on Online Social Networks, pp. 19–24 (2008)
16. Li, Z., Zhang, M., Ma, S., Zhou, B., Sun, Y.: Automatic Extraction for Product Feature Words from Comments on the Web. In: Lee, G.G., Song, D., Lin, C.-Y., Aizawa, A., Kuriyama, K., Yoshioka, M., Sakai, T. (eds.) AIRS 2009. LNCS, vol. 5839, pp. 112–123. Springer, Heidelberg (2009)
17. Liu, B.: Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing, 2nd edn. CRC Press, Taylor and Francis Group (2010)
18. Neviarouskaya, A., Prendinger, H., Ishizuka, M.: Sentiful: A lexicon for sentiment analysis. IEEE Transactions on Affective Computing 2(1), 22–36 (2011)
19. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREC 2010 (2010)
20. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)
21. Semiocast: Half of messages on twitter are not in english. Tech. rep. (2010)
22. Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In: Proceedings of the 2008 ACM Symposium on Applied Computing, pp. 1556–1560 (2008)
23. Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation. Tech. rep., CMU-CALD-02-107 (2002)