A :) Is Worth a Thousand Words: How People Attach Sentiment to Emoticons and Words in Tweets

Marina Boia, Boi Faltings, Claudiu-Cristian Musat, Pearl Pu
École Politechnique Fédérale de Lausanne
Lausanne, Switzerland
firstname.lastname@epfl.ch

***Abstract*—Emoticons are widely used to express positive or negative sentiment on Twitter. We report on a study with live users to determine whether emoticons are used to merely emphasize the sentiment of tweets, or whether they are the main elements carrying the sentiment. We found that the sentiment of an emoticon is in substantial agreement with the sentiment of the entire tweet. Thus, emoticons are useful as predictors of tweet sentiment and should not be ignored in sentiment classification. However, the sentiment expressed by an emoticon agrees with the sentiment of the accompanying text only slightly better than random. Thus, using the text accompanying emoticons to train sentiment models is not likely to produce the best results, a fact that we show by comparing lexicons generated using emoticons with others generated using simple textual features.**

Keywords*—*Twitter, emoticons, sentiment classification

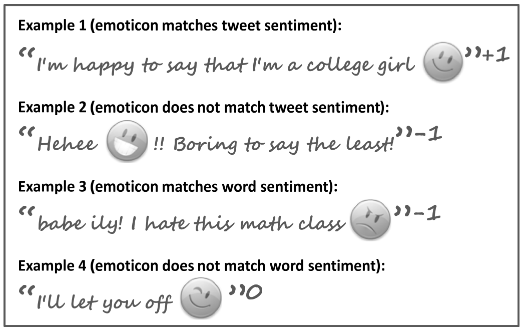
I. Introduction

Since its launch in 2006, Twitter[1] has become one of the most popular microblogging platforms, with over 500 million registered users [1]. A main Twitter communication pattern is the sharing of opinions on companies, products, or events [2]. This behavior is molding Twitter as a form of collective, aggregated wisdom. With proper exploitation, this resource could become invaluable for entities that base their strategic decisions on public opinions forming about relevant topics.

Corpus-based methods are a main direction in Twitter sentiment classification. They typically rely on machine learning algorithms to train classifiers on sentiment annotated datasets. To replace human labelers, automated annotation approaches have been explored that extend the sentiment of emoticons to sentiment labels of tweets. Such methods have a clear advantage, in that they produce substantial amounts of annotated data. However, they are dependent on the effectiveness of emoticons in highlighting representative training instances.

To the best of our knowledge, no previous work has quantified the consensus between the sentiment of an emoticon, the sentiment of the accompanying text, and the sentiment of the entire tweet. Thus, the efficiency of emoticons in Twitter sentiment classification is unclear.

We would like to know to what extent the sentiment of emoticons is in tune with the sentiment of tweets. For instance, in the tweet shown as the first example of Figure 1, the sentiment of the emoticon matches the sentiment of the tweet (tweet sentiment marked with the closing quotes). However, the second example of Figure 1 shows the opposite. Knowing the sentiment consensus between emoticons and tweets would reveal how useful emoticons are as sentiment features.



Example 1 (emoticon matches tweet sentiment):
"I'm happy to say that I'm a college girl "+1

Example 2 (emoticon does not match tweet sentiment):
"Hehee !! Boring to say the least!"-1

Example 3 (emoticon matches word sentiment):
"babe ily! I hate this math class "-1

Example 4 (emoticon does not match word sentiment):
"I'll let you off "0

Fig. 1: Examples of emoticon, word, and tweet sentiment

Moreover, emoticons offer a straightforward way of expressing positive and negative sentiments. We would thus like to know how much more sentiment there is in the words that emoticons accompany. For instance, the tweet presented as the third example of Figure 1 shows sentiment not only through the emoticon, but also through its words (word sentiment marked with the closing quotes). However, the fourth example of Figure 1 does not convey sentiment through its words. Knowing the sentiment consensus between emoticons and words would reveal how useful the sentiment features are that classifiers learn from emoticon-derived sentiment labels.

This work contains two parts. In the first part, we present a live user study on the use of emoticons on Twitter. We analyze the link between the sentiment of emoticons, the sentiment of accompanying words, and the sentiment of tweets. In the second part, we present two methods that generate Twitter subjectivity lexicons from sentiment seeds. We obtain lexicons from emoticons and emotion words and evaluate them to reinforce the findings of the user study.

We conduct a two-stage sentiment annotation experiment. We analyze the implicit sentiment of an emoticon (as conveyed by its graphical representation), the perceived sentiment of accompanying words (as detected by readers), and the perceived sentiment of the entire tweet (as detected by readers through both emoticons and accompanying words). We explore three intuitions about emoticons. A first intuition is that *there is a high agreement between the implicit sentiment of an emoticon and the perceived sentiment of the entire tweet*. A second intuition is that *there is a considerable agreement between the implicit sentiment of an emoticon and the perceived sentiment*

[1] https://twitter.com

of the accompanying words. Finally, a third intuition is that *an emoticon has a strong impact on the perceived sentiment of the entire tweet*.

We show that the sentiment of an emoticon strongly coincides with the sentiment of the tweet. This means that emoticons are very effective sentiment features. We then show that, contrary to our beliefs, the sentiment of an emoticon coincides with the sentiment of the accompanying words only slightly better than random. This means that training sets obtained through emoticon labels might not reveal the most effective sentiment features. Lastly, we show that an emoticon has a substantial impact on the sentiment of the entire tweet. This means that emoticons are important in understanding tweet sentiment.

We explore two unsupervised Twitter subjectivity lexicon generation approaches. A first method derives word sentiment scores from their semantic association with a seed lexicon. A second method iteratively refines a seed lexicon through word frequencies in positive and negative tweets. To generate lexicons, we employ two seed sets. We compare emoticons, which form the first seed set, with emotion words *love* and *hate*, which form the second seed set. We prove that emoticons have a substantial sentiment prediction power, when compared to emotion words. However, we show that the sentiment words obtained through emoticons have a poor performance, when compared to the terms obtained through emotion seeds. We thus reinforce the first two findings in our emoticon study.

The remainder of this paper is structured as follows. Section II presents related work. Section III describes our analysis setup. Section IV illustrates our experiments and results, and Section V draws final remarks.

II. Related Work

A. Emoticons as Sentiment Labels

Emoticons have received a certain amount of attention when building Twitter sentiment models. [3], [4] used emoticons to label tweets as positive and negative, and then trained supervised sentiment classifiers. [5] derived a language model on manually annotated tweets, which they regularized with a second language model trained on emoticon labeled tweets. [6] proposed a graph-based approach, in which nodes were users, tweets, n-grams, hashtags, and emoticons, linked based on authorship and term inclusion. Starting labels were assigned to unigrams and emoticons, and propagated to remaining nodes.

These works made the common assumption that emoticons are a good ground truth for the sentiment of tweets. They also assumed that emoticons are strongly associated with other sentiment words and may thus help reveal them. To the best of our knowledge, there is no previous study that investigates the validity of these assumptions. We thus inquire whether the sentiment of emoticons indeed agrees with the sentiment of the tweets, and how much more sentiment there is in the accompanying words.

B. Subjectivity Lexicon Generation

There has been a substantial amount of work on subjectivity lexicon generation. A typical approach is to expand a small seed set to a full-fledged lexicon, using lexical and semantic resources. [7], [8], [9], [10] employed synonymy and antonymy relations in WordNet [11]. Similarly, [12] modeled a dictionary as a directed graph in which words were linked through synonymy and antonymy. Sentiment terms were added in a breadth-first fashion. [13] gradually refined a set of seeds with terms occurring in their dictionary definitions.

Other methods generated lexicons through word co-occurrence patterns. [14] verified that conjunctions (disjunctions) mostly link adjectives of the same (opposite) polarity. They then trained a model that predicted whether two adjectives had the same orientation. [15] developed a clustering method that started from a small amount of manual annotations and learned subjective adjectives. In [16], two bootstrapping algorithms were used to select subjective nouns. These algorithms started from a small collection of seeds and learned extraction patterns that covered new subjective nouns.

Several types of text have been exploited for lexicon generation. [17], [9] worked with news articles and blog posts, [18] exploited essay data, and [19] used word co-occurrence statistics over the entire web. Tweets, however, have received little attention.

To obtain Twitter lexicons, [20] and [21] provided good starting points. [20] inferred the polarity of words through their semantic association with the words *excellent* and *poor*, computed as pointwise mutual information. The approach used the intuition that positive (negative) words had stronger associations with positive (negative) seeds, and was thus unsupervised. [21] refined a seed lexicon through iterative classification of texts. Each iteration, the current lexicon was used to classify items as positive or negative. Through word frequencies in positive and negative texts, the lexicon was expanded with the best sentiment discriminant terms. The method was unsupervised, as initial labels were provided by the seed lexicon. We aimed to adapt these two approaches for unsupervised generation of Twitter subjectivity lexicons with emoticons and other textual features as seeds.

III. Analysis Setup

A. Dataset

1) Collection: The Twitter Streaming API[2] allows near real-time access to a fraction of the global stream of tweets. Incoming instances may have location information attached, and the service allows the streaming of tweets with location within a specified set of bounding boxes. We collected 2.1 million tweets published in five English speaking cities (*London*, *Sydney*, *New York*, *Los Angeles*, and *San Francisco*), between *18 October 2011* and *7 November 2011.*

2) Preprocessing: We first converted the texts to lower case. We then substituted tweet syntax elements (links, hashtags, and usernames) with standard placeholders (the words *url*, *hashtag*, and *username*). We identified emoticons using the the Net Lingo dictionary[3], which we annotated with positive, negative, or neutral sentiment. We replaced positive and negative emoticons with the standard happy *:)* and sad *:(* examples. We replaced contractions with their full forms and slang terms with their definitions. We substituted sequences of

[2] https://dev.twitter.com/docs/streaming-api

[3] http://www.netlingo.com/smileys.php

repeated characters with two characters of the same kind only, and we replaced series of non-alphanumeric, non-emoticon characters with white spaces. Finally, we stemmed the resulting texts, to obtain the dataset T.

3) Structure: If w is a word, then let $T_w \subset T$ be the subset of items that contain w, and let $T_{\overline{w}}$ be its complement. Also, if W is a set of words, then let $T_W \subset T$:

$$T_W = \bigcup_{w \in W} T_w$$

be the subset of tweets that contain the words in W, and let $T_{\overline{W}}$ be its complement. Furthermore, if $T^{train} \subset T$ is a subset of training items, let T^{train}_W and $T^{train}_{\overline{W}}$ be defined similarly. To zoom in on the tweets with subjective content, we focused on the term sets *:)-:(* $= \{$*:)*, *:(*$\}$ and *lov-hat* $= \{$*love*, *hate*$\}$. Thus, to generate subjectivity lexicons for Twitter, we used $T^{train}_{\textit{:)-:(}} \subset T_{\textit{:)-:(}}$ (roughly 100000 items) and $T^{train}_{\textit{lov-hat}}$ (roughly 90000 items).

B. Emoticon User Study Hypotheses

We designed our emoticon user study so that it investigated:

1) The agreement between the sentiment of an emoticon and the sentiment of the tweet.
2) The agreement between the sentiment of an emoticon and the sentiment of the accompanying words.
3) The impact of an emoticon on the sentiment of the tweet.

A first intuition was that the sentiment of an emoticon is a good indicator for tweet sentiment. Hence, that *there is a high agreement between the implicit sentiment of an emoticon and the perceived sentiment of the entire tweet.* A second intuition was that an emoticon is in substantial sentiment consensus with the words it accompanies. Hence, that *there is a considerable agreement between the implicit sentiment of an emoticon and the perceived sentiment of the accompanying words.* A third intuition was that *an emoticon has a strong impact on the perceived sentiment of the entire tweet.* We summarized our intuitions with three hypotheses:

Hypothesis 1: The implicit sentiment of an emoticon highly agrees with the perceived sentiment of the entire tweet.

Hypothesis 2: The implicit sentiment of an emoticon substantially agrees with the perceived sentiment of the accompanying words.

Hypothesis 3: An emoticon strongly contributes to the perceived sentiment of the entire tweet.

C. Twitter Subjectivity Lexicons

1) Lexicon Generation through Semantic Association: The first lexicon generation method relied on semantic association of terms. For two words w_a and w_b, this can be estimated with *pointwise mutual information* $\text{PMI}(w_a, w_b)$:

$$\text{PMI}(w_a, w_b) = \log_2 \frac{p(w_a, w_b)}{p(w_a)p(w_b)} = \log_2 \frac{p(w_a|w_b)}{p(w_a)},$$

where $p(w_a)$, $p(w_b)$, and $p(w_a, w_b)$ are the probabilities for word occurrence and co-occurrence.

If a word has a stronger semantic association with a highly positive term than with a negative one, then it is of positive sentiment. Similarly, if a word has a stronger semantic association with a highly negative term than with a positive one, then it is of negative sentiment. Following this reasoning, the sentiment score as^* of a word w^* can be estimated by comparing its semantic association with respect to a positive word w_+ and a negative word w_-. The highest *association score* dictates the word's polarity. Therefore, we can have $as^* = \text{as}(w^*, w_+, w_-)$:

$$\begin{aligned}\text{as}(w^*, w_+, w_-) =& \text{PMI}(w^*, w_+) - \text{PMI}(w^*, w_-)\\ =& \log_2 p(w^*|w_+) - \log_2 p(w^*|w_-).\end{aligned}$$

We can further extend this approach by comparing w^*'s association with respect to several positive and negative words. For a set of words W, let $L_W = \{(w, s)|w \in W, s \in \mathbb{R}\}$ be a lexicon on W. Moreover, let $W^+ = \{w^+|(w^+, s) \in L_W, s > 0\}$ and $W^- = \{w^-|(w^-, s) \in L_W, s < 0\}$ be the sets of positive and negative terms. We can define $\text{PMI}(w^*, W^+)$ as the average of w^*'s semantic association values with respect to each $w_+ \in W_+$:

$$\text{PMI}(w^*, W_+) = \frac{1}{|W_+|} \sum_{w_+ \in W_+} \text{PMI}(w^*, w_+).$$

We can introduce $\text{PMI}(w^*, W_-)$ similarly. Thus, we can obtain w^*'s sentiment score as $as^* = \text{as}(w^*, L_W)$:

$$\begin{aligned}\text{as}(w^*, L_W) =& \text{PMI}(w^*, W_+) - \text{PMI}(w^*, W_-)\\ =& \frac{1}{|W_+|} \sum_{w_+ \in W_+} \log_2 p(w^*|w_+)\\ &- \frac{1}{|W_-|} \sum_{w_- \in W_-} \log_2 p(w^*|w_-).\end{aligned}$$

Given a set of words W, a subset of training tweets $T^{train}_W \subset T^{train}$, a seed lexicon L_W, and the set of all words $W \subset W^*$ in T^{train}_W, we obtained a lexicon L^{assc}_W by attaching a sentiment score of $as^* = \text{as}(w^*, L_W)$ to every word $w^* \in W^*$. We estimated word probabilities using their frequency in T^{train}_W:

$$p(w_a|w_b) = \frac{\text{count}(w_a, w_b)}{\text{count}(w_b)},$$

where $\text{count}(w_a, w_b)$ is the frequency of w_a and w_b co-occurring, and $\text{count}(w_b)$ is the frequency of w_b occurring.

2) Lexicon Generation through Iterative Classification: The second method was iterative. Given a set of words W, a subset of training tweets $T^{train}_W \subset T^{train}$, a seed lexicon L_W, and the set of all words $W \subset W^*$ in T^{train}_W, we cycled through the following steps to obtain a lexicon for T^{train}_W.

a) Tweet Scores: Let L^i_W be the lexicon available at the start of the i^{th} iteration, with $L^1_W = L_W$. We used L^i_W to assign sentiment scores to tweets in T^{train}_W. For each tweet $t \in T^{train}_W$, its words were matched against L^i_W. Each word matched $w^* \in W^*$ was given an *effective score* $es^* = \text{es}(w^*, t, L^i_W)$:

$$\text{es}(w^*, t, L^i_W) = \frac{\text{length}(w^*)}{\text{length}(t)} \times s^* \times n^*, \tag{1}$$

where length(w^*) is w^*'s character length, length(t) is t's character length, s^* is w^*'score in L^i_W, and n^* is a negation flag indicating the presence of a negation near w^*. We summed the effective scores to obtain a *tweet score* $ts = \text{ts}(t, L^i_W)$:

$$\text{ts}(t, L^i_W) = \sum_{w^* \in t} \text{es}(w^*, t, L^i_W). \qquad (2)$$

b) Tweet Labels: We used the tweet scores to classify a subset of T^{train}_W. We separately sorted in decreasing order of absolute scores the tweets with positive and negative values. We disregarded the lowest ranked tweet scores from the larger set. We used the remaining scores to classify the corresponding tweets as positive or negative. For every item t (whose score we did not disregard), we assigned a *tweet label* $tl = \text{tl}(t)$:

$$\text{tl}(t, L^i_W) = \text{sign}(\text{ts}(t, L^i_W)). \qquad (3)$$

c) Lexicon Update: We used the tweet labels to update from L^i_W to L^{i+1}_W. For every $w^* \in W^*$ that appeared at least twice, we used its frequencies in positive and negative items $F^{w^*}_p$ and $F^{w^*}_n$ to assign w^* a *subjective discrimination score* $sds^* = \text{sds}(w^*)$:

$$\text{sds}(w^*) = \frac{|F^{w^*}_p - F^{w^*}_n|}{F^{w^*}_p + F^{w^*}_n}.$$

Given a threshold $\tau \in [0, 1]$, we kept only the words w^* for which $sds^* > \tau$. We added them to L^{i+1}_W with an *iteration score* $is^* = \text{is}(w^*)$:

$$\text{is}(w^*) = F^{w^*}_p - F^{w^*}_n.$$

d) Stop Condition: The cycle ended when there was no change from L^i_W to L^{i+1}_W, or when a maximum number of iteration was met. This gave the final lexicon $L^{\textit{iter-}\tau}_W$.

3) Sentiment Classification: Given a lexicon L, we classified a tweet t as follows. We matched its words against L. To every matched word w^*, we attached an effective score es^* as in Equation 1. We then combined the resulting effective scores to an overall tweet score ts, as in Equation 2. Finally, thresholding the resulting ts score at zero, as in Equation 3, gave a final tweet label tl for t. Items with null tweet scores remained unclassified.

IV. Experiments And Results

A. Emoticon User Study

1) Experimental Setup: To study the three emoticon hypotheses, we devised an experiment in which ten participants labeled tweets with their sentiment. Each participant p was assigned 120 tweets, randomly selected. One third of the assigned items contained positive emoticons ($T^p_{:)} \subset T_{:)}$), another third contained negative emoticons ($T^p_{:(} \subset T_{:(}$), whereas the rest did not contain emoticons ($T^p_{\overline{:)\text{-}:(}} \subset T_{\overline{:)\text{-}:(}}$). The set of 120 tweets $T^p = T^p_{:)} \cup T^p_{:(} \cup T^p_{\overline{:)\text{-}:(}}$ was revealed in shuffled order.

If a tweet t contained an emoticon e, it was initially presented with e hidden. The participant had to specify a first sentiment label l_{wrd} for the words in t, with possible options positive (+1), negative (-1), neutral, and unsure. Once l_{wrd} was confirmed, the emoticon e was revealed, and the participant had to indicate a second label l_{twt}, from the same set of options. Thus, l_{wrd} indicated the sentiment of the words in t, while l_{twt} indicated the sentiment of the entire tweet.

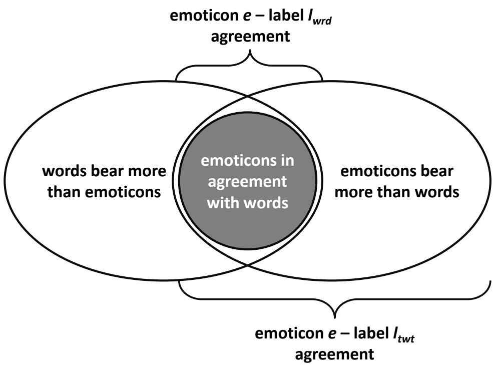

emoticon e – label l_{wrd} agreement
words bear more than emoticons
emoticons in agreement with words
emoticons bear more than words
emoticon e – label l_{twt} agreement

Fig. 2: Annotation agreement situations

We used the outcome of this experiment to obtain a test set T^{test}. We defined T^{test} as the set of 350 and 334 items that were positive and negative in the label l_{twt} respectively. We chose l_{twt} because we were later on interested in testing subjectivity lexicon efficiency in predicting tweet sentiment.

Furthermore, through this experiment we pursued our three emoticon hypotheses. We analyzed the first hypothesis through the agreement between the sentiment of an emoticon e and the label l_{twt}. Such an agreement happens in two cases. The first situation is that the emoticon agrees in sentiment with the words it accompanies. The second possibility is that the emoticon bears more sentiment than the accompanying words, and that it dictates the tweet sentiment (Figure 2).

We studied the second hypothesis through the agreement between the sentiment of an emoticon e and the label l_{wrd}. Such an agreement coincides with the first of the two situations mentioned as causes for sentiment agreement between an emoticon and the tweet (Figure 2).

Lastly, we studied the third hypothesis through the difference between the first and the second levels of agreement. This difference points out the cases where an emoticon bears more sentiment than the words it accompanies (Figure 2).

2) Results: Table I shows the agreement between an emoticon e and the label l_{twt}. We see that 71% of the tweets with positive emoticons are classified as positive, and 77% of the items with negative emoticons are classified as negative. Thus, the sentiment of an emoticon strongly coincides with the sentiment of the entire tweet. Hence, the level of agreement between the sentiment of an emoticon and the label l_{twt} supports our first hypothesis.

We performed a Welch test to establish whether $S^{twt}_{+} = \{I_+(l_{twt}) | (t, l_{wrd}, l_{twt}) \in T^{test}\}$ and $S^{twt}_{+,:)} = \{I_+(l_{twt}) | (t, l_{wrd}, l_{twt}) \in T^{test}_{:)}\}$ were statistically different, where:

$$I_+(l) = \begin{cases} 1, \text{if } l = +1 \\ 0, \text{if } l \neq +1 \end{cases}.$$

The empirical means of S^{twt}_{+} and $S^{twt}_{+,:)}$ were estimates for $p^{twt}(+)$ (the probability of the positive class as indicated by the label l_{twt}) and $p^{twt}(+|{:)})$ (the probability of the positive

TABLE I: Emoticons e - label l_{twt} agreement

e / l_{twt}	**positive**	**negative**	**neutral**	**unsure**
:)	**71.38%**	5.72%	15.06%	7.83%
:(	7.55%	**77.34%**	8.15%	6.94%

TABLE II: Emoticons e - label l_{wrd} agreement

e / l_{wrd}	**positive**	**negative**	**neutral**	**unsure**
:)	**56.32%**	12.65%	26.80%	4.21%
:(	19.33%	**58.61%**	18.12%	3.92%

class as indicated by the label l_{twt}, conditioned on a positive emoticon). We obtained a p-value of 3.74×10^{-33}. Similarly, we performed a Welch test on S^{twt}_{-} and $S^{twt}_{-,:(}$ and obtained a p-value of 6.54×10^{-50}. We thus concluded that the effect of positive (negative) emoticons on the positive (negative) class distribution as given by the label l_{twt} is strongly statistically relevant. On this basis, we established the relevance of the results in Table I, and we accepted the first hypothesis.

Table II shows the agreement between an emoticon e and the label l_{wrd}. Only 56% of the tweets with happy emoticons are classified as positive at the first attempt. Similarly, only 59% of the tweets with sad emoticons are classified as negative at the first trial. That is, the sentiment of emoticons coincides with the sentiment of accompanying words in a manner that is only slightly better than random. Hence, the agreement between emoticons and the label l_{wrd} does not support our second hypothesis. It seems that, in fact, when emoticons are used, they often replace words as the carriers of sentiment.

We performed a Welch test to establish if the samples $S^{wrd}_{+} = \{I_{+}(l_{wrd}) | (t, l_{wrd}, l_{twt}) \in T^{test}\}$ and $S^{wrd}_{+,:)} = \{I_{+}(l_{wrd}) | (t, l_{wrd}, l_{twt}) \in T^{test}_{:)}\}$ were statistically different. We arrived at a two-tailed p-value of 2.76×10^{-9}. We then performed a Welch test on the samples S^{wrd}_{-} and $S^{wrd}_{-,:(}$, defined similarly. We arrived at a p-value of 1.64×10^{-16}. We concluded that the effect of positive (negative) emoticons on the positive (negative) class distribution given by the label l_{wrd} is statistically relevant. We thus established the relevance of the results in Table II, and we rejected the second hypothesis.

The difference between the agreement scores presented in Tables I and II shows how the label l_{wrd} is changed to a positive (negative) label l_{twt} as a result of a positive (negative) emoticon. Happy emoticons prompt a switch to a positive l_{twt} 15% of the times. Similarly, sad emoticons produce a change to a negative l_{twt} 18% of the times. This means that the contribution of emoticons, seen as the amount of change from the sentiment of accompanying words to the sentiment of tweets, is important, thus supporting our third hypothesis. Given the proven relevance of the agreement results in Tables I and II, we accepted the third hypothesis.

B. Twitter Subjectivity Lexicon Evaluation

In the second experiment, we used the two generation methods to obtain Twitter lexicons. Through lexicon evaluation, we reinforced the findings of our emoticon study.

TABLE III: Lexicon evaluation

lexicon	**precision**	**recall**	**f-score**
$L^{assc}_{:)\text{-}:(}$	87.43%	100.00%	**93.29%**
$L^{iter\text{-}0.2}_{:)\text{-}:(}$	87.54%	98.54%	**92.71%**
$L^{assc}_{lov\text{-}hat}$	85.96%	100.00%	**92.45%**
$L^{iter\text{-}0.4}_{lov\text{-}hat}$	85.17%	96.64%	**90.54%**

1) Experimental Setup: We defined two seed lexicons. To capture the positive and negative words correlated with emoticons, we introduced a first seed set $L_{:)\text{-}:(} = \{(:), +1), (:(, -1)\}$. To derive the sentiment terms associated with emotion words *love* and *hate*, we defined a second seed set $L_{lov\text{-}hat} = \{(love, +1), (hate, -1)\}$.

We defined T^{train} as the set of tweets not in T^{test}. We applied $L_{:)\text{-}:(}$ and $L_{lov\text{-}hat}$ on $T^{train}_{:)\text{-}:(}$ and $T^{train}_{lov\text{-}hat}$ respectively. We employed the semantic association approach to obtain lexicons $L^{assc}_{:)\text{-}:(}$ and $L^{assc}_{lov\text{-}hat}$. We used the iterative method by varying the lexicon acceptance threshold τ from 0.1 to one, with a step size of 0.1. We obtained lexicons $L^{iter\text{-}\tau}_{:)\text{-}:(}$ and $L^{iter\text{-}\tau}_{lov\text{-}hat}$.

We evaluated the generated lexicons on T^{test}. To better investigate the effectiveness of the seeds, we also separately tested the two seed lexicons and the sentiment words they helped identify: $L^{assc}_{:)\text{-}:(*}$, $L^{iter\text{-}\tau}_{:)\text{-}:(*}$, $L^{assc}_{lov\text{-}hat*}$, and $L^{iter\text{-}\tau}_{lov\text{-}hat*}$ (where $L^{assc}_{W*} = L^{assc}_{W} - L_{W}$ and $L^{iter\text{-}\tau}_{W*} = L^{iter\text{-}\tau}_{W} - L_{W}$).

To quantify performance, we recorded precision, recall, and f-score. We interpreted precision as the percentage of items correctly classified, with respect to the subset of tweets for which sentiments were identified. We interpreted recall as the percentage of classified items, with respect to the entire document collection.

2) Results: Table III presents the lexicon evaluation results. It appears that both emoticon and emotion seeds produce lexicons of good quality, with f-scores above 90%. While recall is consistently close to 100%, precision seems to indicate emoticons as the slightly better choice of seeds.

Table IV presents the separate evaluation of the seed lexicons and of the newly added words. The seed emoticon lexicon gives a precision of 92%. Moreover, the sentiment words identified through emoticons give a precision of 75%. The seed emotion lexicon has a precision of 81%, while the terms gathered through emotion words give a precision of 86%.

The seed evaluation results show that emoticons have a very good sentiment classification power. In more than 90% of the tweets with emoticons, they indicate the correct sentiment orientation. The emotion words also perform appropriately, but, comparatively, they show a 11% relative decrease in precision. This considerable sentiment prediction power of emoticons reinforces the acceptance of the first hypothesis.

The added words evaluation shows that emoticons are not very efficient in gathering new sentiment words. The terms identified through emoticons have a poor prediction power. Comparatively, the words identified through emotion words bring a relative increase in performance of 11%. This low performance of the words added through emoticons reinforces the rejection of the second emoticon hypothesis.

TABLE IV: Seed lexicon and added words evaluation

lexicon	precision	recall	f-score
$\boldsymbol{L}_{\textit{:)-:(}}$	91.60%	78.36%	**84.47%**
$\boldsymbol{L}_{\textit{lov-hat}}$	81.08%	10.82%	**19.09%**
$\boldsymbol{L}^{\textit{assc}}_{\textit{:)-:(}*}$	73.54%	100.00%	**84.75%**
$\boldsymbol{L}^{\textit{assc}}_{\textit{lov-hat}*}$	85.82%	100.00%	**92.37%**
$\boldsymbol{L}^{\textit{iter-0.2}}_{\textit{:)-:(}*}$	76.31%	91.96%	**83.41%**
$\boldsymbol{L}^{\textit{iter-0.4}}_{\textit{lov-hat}*}$	86.00%	96.05%	**90.75%**

Lastly, going back to the evaluation results of Table III, we can remark that the good performance of the emoticon lexicons is a result of the structure of T^{test}. Because of the manual sentiment annotation experiment, numerous test items contain emoticons. This means that emoticons often have the chance to lead to correct classifications.

V. Conclusion

We presented a live user study focusing on three aspects of the use of emoticons on Twitter. We showed that the sentiment of emoticons strongly coincides with the sentiment of tweets. This means that emoticons by themselves are very good sentiment predictors. We also showed that the sentiment of emoticons is only slightly agreeing with the sentiment of the accompanying words. This means that training sets obtained through emoticon labels might not lead to the most effective sentiment features. At the same time, we proved that emoticons have a substantial contribution in tweet sentiment. This means that emoticons are important in understanding this sentiment.

Moreover, we explored two unsupervised methods for Twitter subjectivity lexicon generation. Both approaches refined sentiment seeds into full-fledged lexicons. We generated lexicons from both emoticons and emotion seed words. By themselves, emoticons performed considerably better than the two emotion words. At the same time, the sentiment words gathered through emoticons produced a substantially poorer performance than the terms gathered through emotion words. These outcomes strengthened the first two user study findings.

Thus, emoticons offer a straightforward means of expressing sentiments, which words do not duplicate. They are a good ground truth for the sentiment of the entire tweet, but a bad ground truth for the sentiment of the accompanying words. On the one hand, this means that Twitter sentiment classification algorithms should give emoticons a central role as sentiment features. On the other hand, this means that emoticon-derived sentiment labels might not give the best results in Twitter sentiment classification.

References

[1] L. Dugan, “Twitter to surpass 500 million registered users on wednesday,” http://www.mediabistro.com/alltwitter/500-million-registered-users_b18842, 2012, [Online; accessed 23 November 2012].

[2] A. Java, X. Song, T. Finin, and B. Tseng, “Why we twitter: understanding microblogging usage and communities,” in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, 2007, pp. 56–65.

[3] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” Stanford, Tech. Rep., 2009.

[4] A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” in *Proceedings of the 7th International Conference on Language Resources and Evaluation*, 2010.

[5] K.-L. Liu, W.-J. Li, and M. Guo, “Emoticon smoothed language models for twitter sentiment analysis,” in *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 2012, pp. 678–684.

[6] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge, “Twitter polarity classification with label propagation over lexical links and the follower graph,” in *Proceedings of the 1st Workshop on Unsupervised Learning in Natural Language Processing*, 2011, pp. 53–63.

[7] S.-M. Kim and E. Hovy, “Determining the sentiment of opinions,” in *Proceedings of the 20th International Conference on Computational Linguistics*, 2004.

[8] S. Blair-Goldensohn, T. Neylon, K. Hannan, G. A. Reis, R. Mcdonald, and J. Reynar, “Building a sentiment summarizer for local service reviews,” in *Natural Language Processing in the Information Explosion Era*, 2008.

[9] N. Godbole, M. Srinivasaiah, and S. Skiena, “Large-scale sentiment analysis for news and blogs,” in *Proceedings of the International Conference on Weblogs and Social Media*, 2007.

[10] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 168–177.

[11] C. Felbaum, Ed., *Wordnet: An Electronic Lexical Database*. MIT Press, 1998.

[12] A. Paulo-Santos, C. Ramos, and N. C. Marques, “Determining the polarity of words through a common online dictionary,” in *Proceedings of the 15th Portugese Conference on Progress in Artificial Intelligence*, 2011, pp. 649–663.

[13] C. Banea, R. Mihalcea, and J. Wiebe, “A bootstrapping method for building subjectivity lexicons for languages with scarce resources,” in *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 2008.

[14] V. Hatzivassiloglou and K. R. McKeown, “Predicting the semantic orientation of adjectives,” in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, 1997, pp. 174–181.

[15] J. Wiebe, “Learning subjective adjectives from corpora,” in *Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, 2000, pp. 735–740.

[16] E. Riloff, J. Wiebe, and T. Wilson, “Learning subjective nouns using extraction pattern bootstrapping,” in *Proceedings of the 7th Conference on Natural Language Learning*, 2003, pp. 25–32.

[17] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 347–354.

[18] B. Beigman Klebanov, J. Burstein, N. Madnani, A. Faulkner, and J. Tetreault, “Building subjectivity lexicon(s) from scratch for essay data,” in *Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing*, 2012, pp. 591–602.

[19] L. Velikovich, S. Blair-Goldensohn, K. Hannan, and R. Mcdonald, “The viability of web-derived polarity lexicons,” in *Proceedings of The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

[20] P. Turney, “Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002, pp. 417–424.

[21] L. Qiu, W. Zhang, C. Hu, and K. Zhao, “Selc: A self-supervised model for sentiment classification,” in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 2009, pp. 929–936.