

Emotion Analysis on Twitter: the Hidden Challenge

Luca Dini, André Bittar

Holmes Semantic Solutions

Grenoble, France

E-mail: dini@ho2s.com, bittar@ho2s.com

In this paper, we present an experiment to detect emotions in tweets. Unlike much previous research, we draw the important distinction between the tasks of emotion detection in a closed world assumption (i.e. every tweet is emotional) and the complicated task of identifying emotional versus non-emotional tweets. Given an apparent lack of appropriately annotated data, we created two corpora for these tasks. We describe two systems, one symbolic and one based on machine learning, which we evaluated on our datasets. Our evaluation shows that a machine learning classifier performs best on emotion detection, while a symbolic approach is better for identifying relevant (i.e. emotional) tweets.

Keywords: emotion analysis, Twitter, corpora

1. Introduction

In recent years so-called “emotional marketing” has become a key factor of success for many B2C companies, especially global ones. Put simply, emotional marketing has the basic goal of convincing customers that a brand or a product are not just brands or products, but that they are imbued with qualities that favor an emotional response or attachment in consumers. Any emotional marketing strategy starts with an in-depth understanding of customers’ emotional or motivational drivers, and this is achieved either via open question surveys or by analyzing emotional reactions on social media. This paper has the goal of validating the feasibility of automatic emotional analysis on social media. It should be noticed that standard sentiment analysis, by which we mean the detection of opinions as positive versus negative, or possibly ranked on a scale of intensity, is of little use for our purposes, both practically and conceptually. Practically, the fact that a customer expresses, for instance, a positive opinion has no impact on an emotional evaluation. For example, the sentences *I love this cup* and *This cup is very good* have an identical opinion value, but definitely a different emotional tonality. Conceptually, whereas opinions are usually expressed in a quite explicit way, the language of emotions is much more difficult to interpret as it relies on indirect signals.

We emphasize the difficulty of the task at hand by comparing a symbolic approach to emotion analysis and one based on machine learning. We show that while the task of classifying emotions in a closed world assumption (i.e. each text has an emotional value) is a relatively easy task, the real challenge lies in distinguishing emotional texts from non-emotional ones (information filtering). We also show that a statistical approach outperforms a symbolic one in emotion classification, while the contrary is true when considering information filtering.

Concretely, we focused on two different tasks. The first is the categorization of tweets according to the emotion they express – a single emotion per tweet, from a set of 6 basic emotions. As we explain in detail in **Erreur ! Source du renvoi introuvable.**, currently available corpora do not, to our knowledge, provide appropriate data for evaluating the automation of this task and a major objective of this work was to make up for this shortfall. The second task was to distinguish tweets that express an emotion from those that

do not. We also created a corpus as evaluation data for this task.

2. Background

Much research has been carried out on sentiment analysis in the past decade. (Wiebe et al, 2005) carried out an ambitious annotation effort, producing a 10,000-sentence corpus of world news texts marked up for a range of sentiment-related phenomena, although the typology of emotions focused on polarity rather than actual emotion types.

(Alm et al, 2005) created a corpus of children’s tales where each sentence (1,580 in total) is annotated with one of 6 emotions (Ekman, 1993), or a neutral value. They also describe a supervised machine learning system that detects emotional versus neutral sentences and detects emotion polarity within a sentence. Although this work captures actual emotion types, rather than just polarity, the type of texts is not compatible with the task we are tackling, namely analysis of tweets. (Pak & Paroubek, 2010) created a corpus of tweets for sentiment analysis, but again, only focus on emotion polarity.

(Vo & Collier, 2013) created a corpus of Japanese tweets annotated according to an ontology of 5 emotions (similar to the Ekman typology) and a system to automatically detect the prevalent emotion in a tweet, achieving a global F-score of 64.7. This work approaches our current work, although for a different language. Finally, the numerous SemEval campaigns that have included various tasks on emotion analysis have almost exclusively focused on the assignment of a polarity value to detected emotions. An exception is the 2007 campaign (Strapparava & Mihalcea, 2007), Task 14: Affective Text, in which participants were required to detect and classify emotions and/or determine their polarity in a corpus of manually annotated news headlines extracted from news websites. In the annotated corpus, each headline is annotated with a numeric interval (0-100) that indicates, for each of 6 emotions (ANGER, DISGUST, FEAR, JOY, SADNESS, SURPRISE), its strength in the headline. The 6 human annotators who prepared the corpus were instructed to “select the appropriate emotions for each headline based on the presence of words or phrases with emotional content, as well as the overall feeling invoked by the headline.” The nature of this task is rather subjective,

relying on each annotator’s interpretation of a given headline and their emotional response. Furthermore, requesting annotation over such a fine-grained interval for each emotion allows even further room for disagreement. This degree of subjectivity is reflected in the reported inter-annotator agreement (Pearson correlation coefficient) for the task, which “is not particularly high” (Strapparava & Mihalcea, 2007, p74) (an average agreement of 53.67¹ across the 6 emotions). The resulting “gold standard” corpus for this task is, therefore, not a reliable yardstick against which to evaluate system performance. Indeed, evaluation results for the participating systems were relatively quite low (F-scores ranging from 16.00 to 42.43). In terms of corpus preparation and experimental setup, our work has much in common with that of (Suttles & Ide, 2013) which uses hashtags (among other things) to build an emotion corpus (which is however not publically available). Whenever relevant in the course of this paper we will provide comparisons between their approach and our own. In light of previous work and as, to our knowledge, no suitable corpus for the task at hand exists, it was necessary for us to create a new corpus of tweets annotated for emotions for the evaluation of our system.

3. The Corpora

3.1 A silver standard corpus for emotion classification in tweets

The first corpus, which we will call the “Emotion Tweet Corpus for Classification” (henceforth, ETCC) is a “silver standard” in which each tweet is classified with a single emotion. In creating this corpus, the basic premise we relied on was that some Twitter users, when expressing emotions, also tag their message with emotional hashtags (Saif, 2012). Based on this assumption we tried to construct a corpus where tweets are classified according to the 6 emotional classes used in SemEval 2007 (ANGER, DISGUST, FEAR, JOY, SADNESS and SURPRISE) (Strapparava & Mihalcea, 2007). The choice of hashtag to be associated with emotions (one hashtag per emotion) was very important, as they needed to be common enough to allow the retrieval of a significant number of tweets, and be unambiguous: for instance for the emotion *surprise* we could not use the hashtag *#surprise* as it is semantically highly ambiguous (as an interjection – “Surprise!” – as a noun meaning something that surprises, an act of surprising, etc.). Instead, we opted for the unambiguous hashtag *#astonished*. The 6 emotion hashtags we used were *#angry*, *#astonished*, *#disgusted*, *#happy*, *#sadness* and *#scared*. The corpus collection phase was made much easier by the fact that since November 2014 (Zhuang, 2014), Twitter proposes a search interface not emphasizing recentness and allowing the retrieval of tweets since 2006: emotional hashtags could then be used as search keywords and the necessary number of tweets for each emotion was collected (20,000 per emotion) using the approach described in (Dickinson, 2015). We then performed some filtering to remove inappropriate tweets (we call this “formal filtering”). We eliminated non-English tweets by specifying the language in the search query. We also removed tweets which were not composed of text (for example, by filtering out tweets that had a

higher proportion of hashtags than other tokens). Tweets containing links to multimedia content were also filtered out, as in general the emotional hashtag in such tweets relates to the indicated media rather than the textual content of the tweet. After manual inspection we still noticed that the number of tweets containing an emotional tag but no emotional text was still high. We therefore applied a further filter (we call this “affect filtering”) based on WordNet-Affect (Valitutti, 2004) – we indexed all contents with Lucene and then ran a fuzzy search selecting only tweets containing an emotional word from the WordNet-Affect lexicon. Figures for the resulting corpus are given in Table 1.

SemEval 2007 Emotion	Hashtag	After formal filtering	After affect filtering
Anger	#angry	8,738	5,105
Surprise	#astonished	16,970	8,635
Disgust	#disgusted	14,508	9,084
Joy	#happy	3,574	2,009
Sadness	#sadness	3,364	1,724
Fear	#scared	10,525	5,750

Table 1: figures for the corpus of emotion-classified tweets.

Finally, all hashtags appearing at the end of a tweet were removed and hashtags that occurred before the end of a tweet had their hash sign removed, as in such cases they are often used in the place of regular words. For example, after this step, the tweet in a becomes the tweet in b:

- a. #MoodSwings are a #symptom of being #Bipolar. If you're #scared, #sad, #paranoid, or #suicidal, there's help here: <http://ow.ly/SByb6>
- b. MoodSwings are a symptom of being Bipolar. If you're scared, sad, paranoid, or suicidal, there's help here: <http://ow.ly/SByb6>

Although the corpus was not created via a full manual annotation (hence its “silver” status), the criteria used for retrieving and selecting the texts were anchored in the actual textual forms of the tweets, as opposed to relying on highly subjective annotator judgements, as was the case in SemEval 2007.

The corpus collection method is similar to the one described in (Suttles & Ide, 2013), with the following differences:

- (Suttles & Ide, 2013) use emoticons and emoji
- They used the emotion categories of (Plutchik, 1980), as opposed to those of (Ekman, 1993).
- They used sets of hashtags rather than a single hashtag in order to increase recall. This expansion is fully justified in light of the fact that at the time a search interface to Twitter was not available for tweet selection. However, it is likely that it introduced noise (the authors do not report the set of hashtags used).

¹ As calculated from the figures provided in Table 1 of

(Strapparava & Mihalcea, 2007, p72).

- In the normalization phase they eliminate the emotional word which caused the selection of the tweet. In our case we perform elimination only when the word is outside a syntactic context, as systematic elimination would break syntactic semi-wellformedness, thus negatively affecting symbolic analysis.

Overall, the dataset resulting from their approach is larger by roughly one order of magnitude. As for the proportions among different categories, they are unfortunately incomparable, given the difference in the two adopted emotion taxonomies.

3.2 A gold standard corpus for tweet relevance

Evaluating a system on a corpus composed only of emotional tweets hides the most challenging problem in real life applications, namely, the one of distinguishing, in a continuous stream of tweets, emotional tweets from non-emotional ones. In order to evaluate our system on this task, we collected a set of 1,250 tweets containing the keyword “iPhone 4” posted in the period from October 1st 2015 21:55:36 CEST to October 2nd 2015 12:05:32 CEST. These were manually annotated by a single English native speaker, according to their emotional content – emotional or not emotional. The annotator was instructed to mark as emotional all tweets that contained:

- an emotional lexical item,
- an emotional emoticon,
- an emotional internet slang expression,
- an emotional hashtag.

These annotation criteria are all based on the linguistic and other surface forms of the text and as such, make for an annotation task that relies as little as possible on the subjective judgement of the annotator. In the resulting corpus, the “Emotion Tweet Corpus for Relevance” (henceforth, ETCC) we noticed that only 9% of the tweets had some lexically realized emotional content, according to the annotators. We also observed that only 4% of this 9% had an emotional hashtag.²

4. The Machine Learning Approach

The first test we ran, on the ETCC corpus, was to use a classifier to discern the emotion expressed in each tweet. The corpus was split via random sampling into 80% training and 20% test. We used a multiclass linear classifier associated with a Quasi Newton minimizer, under the Stanford NLP implementation. We paid particular attention to the feature selection process, and after several tests the best results were obtained with the following set of features:

- **Word:** the surface word form without normalization.

- **Lemma:** lemma and POS tag, as resulting from POS disambiguation.
- **Noun phrase:** we use the output of the dependency grammar to produce all possible well-formed noun phrases. Noun phrases are passed to the classifier both as sequences of word forms and sequences of lemmas.
- **Dependencies:** a certain subset of grammatical dependencies is passed to the classifier as a set of triples. For instance (verb,SUBJ,noun), (verb,OBJ,noun), (noun,MOD,adj), etc. where parts of speech are obviously replaced by the relevant lemma (e.g. (hate,SUBJ,I), (have,OBJ,money)). As the grammar we use produces Stanford-style dependencies (Marneffe & Manning, 2008), the dependency features are close to a semantic representation.

For each tweet, the classifier assigns a probability for each emotion (the total probability mass being 1) and each tweet is assigned the emotion with the highest probability.

5. The Symbolic Approach

Our symbolic approach to emotion annotation was carried out using our in-house system, SentiMiner, developed within the company over several years (Maurel et al, 2007; Bittar et al, 2014) as part of HOLMES, a hybrid NLP platform (Dini et al, 2013). Processing consists of 3 main stages that integrate into the usual pre-processing pipeline (sentence-detection, tokenization, part-of-speech tagging, morphological analysis and lemmatization, dependency parsing). These stages are: lexical tagging, token-based regular expression annotation, and dependency graph transformation. Each of these is described below.

Lexical tagging with gazetteers: The main lexical resource used in SentiMiner is a gazetteer of emotions (1,577 lemmas) automatically extracted from the WordNet Affect database. A mapping between the WordNet Affect emotions and the 6 basic emotions used for this experiment was established. Classes of emotions that did not have a coherent mapping were discarded. Classes that had multiple mappings were split. The resulting gazetteer used for this experiment contains 1,302 lemmas. Furthermore, a separate gazetteer of internet slang terms and their corresponding emotions (e.g. LOL = JOY, WTF = SURPRISE), containing 416 entries, was also used. This gazetteer was created from a list of internet slang terms³ from which non-emotional terms were discarded, and each remaining term was manually attributed one of the 6 emotion values. A third gazetteer was used to disambiguate lemmas that are

² This step is also performed by (Suttles & Ide, 2013). However, they do not provide the percentage of relevant tweets, making comparison difficult.

³ From the website <http://www.noslang.com/dictionary/>, consulted 10 October 2015.

only emotions when present with specific parts-of-speech. For example, “like” is an emotion as a verb, but not as a preposition; “close” is an emotion as an adjective or adverb, but not as a verb, etc. This gazetteer contains 1,547 emotion lemmas with their possible parts-of-speech. After part-of-speech tagging, all lexical items in the input text that have a lemma in one of the emotion gazetteers are tagged with their corresponding emotion and possible parts-of-speech.

Token-based regular expression annotation (cf Stanford TokensRegex (Chang & Manning, 2014)): After lexical tagging, a set of token-based grammar rules are first applied to correct emotion annotations based on possible parts-of-speech. For example, the preposition *like* in *John eats like a pig* is not an emotion and is discarded. A further set of rules is used to process certain multi-word expressions that are able to be dealt with by a regular grammar without a deep syntactic analysis. These rules remove emotions from certain contexts, e.g. *close minded* (sic), *with respect to*, etc.

Dependency graph transformation grammar (using the Stanford Semgrep engine (Chambers et al., 2007)): After dependency parsing, the final step is the sequential application of a set of graph transformation grammars that mark relations between emotion words and their arguments. These grammars have access to all annotations added during previous processing. Certain rules are used to remove emotions from specific contexts, for example in the scope of a modal operator (e.g. *You would be astonished*, *You should be happy*, etc.), to remove emotions from common interjections (e.g. *Good night*, *Happy Birthday/New Year/Anniversary*, *Merry Christmas*), in certain expressions (e.g. *You have got to be kidding me*), and so on. Other rules add an EMOTION relation between an emotion word and its syntactic argument (e.g. *John is angry* – EMOTION(*angry*,*John*), *This is a frightening book* – EMOTION(*frightening*,*book*), *John has sympathy for Mary* – EMOTION(*sympathy*,*John*), *John’s sympathy for Mary* – EMOTION(*sympathy*,*John*), etc.). Furthermore, our grammar assigns one of two relations to indicate the status of the experiencer with respect to the emotion (causative or stative).

- c. *John* is a *shy* person.
- d. This *film* impresses me.

For example, in sentence c), the grammar marks EXPERIENCER_STAT(*shy*,*John*), indicating that *shy* is a state of its subject, while in d) the grammar assigns EXPERIENCER_CAUSE(*film*,*impresses*), indicating that the subject of the emotion word *impresses* is the cause of the emotion. Although these two relations are output by our system, they were not used for the purposes of the current experiments.

The annotated relations, aside from the two just mentioned, mark the presence of an emotion in the final output. The final emotion is assigned to a given tweet, firstly, according to the number of occurrences found. If all detected emotions occur in equal numbers, the first one (from left to right) is assigned.

6. The Symbolic Approach

6.1 Tweet emotion detection

For this task, we determined a baseline against which to gauge the performance of our classifiers by calculating precision, recall and F-score for each emotion in the ETCC corpus according to the simple presence or absence of the appropriate emotion hashtag in the tweet text (e.g. “anger” in the ANGER tweets were considered true positives, “anger” absent from an ANGER tweet was a false negative, and so on). It is clear that this method emphasized precision over recall, as it emulates a grep-style filtering program. However, it does set a more challenging baseline than random assignment. Baseline figures are presented in Table 2.

	Precision	Recall	F-score
Anger	0.96	0.37	0.53
Disgust	1.00	0.33	0.49
Fear	0.98	0.17	0.28
Joy	0.78	0.62	0.69
Sadness	0.99	0.32	0.48
Surprise	0.98	0.28	0.43
Average	0.95	0.35	0.49

Table 2: Baseline evaluation figures for emotion classification.

Evaluation results for the ML classification of tweets (Table 3) on the ETCC corpus shows varying performance across emotion types, reflecting the amount of data available for training for each emotion (cf. Table 1). The classifier achieved an average improvement in F-score of 9% points over the baseline.

	Precision	Recall	F-score
Anger	0.53	0.46	0.49
Disgust	0.66	0.72	0.69
Fear	0.61	0.65	0.63
Joy	0.63	0.6	0.62
Sadness	0.54	0.37	0.44
Surprise	0.62	0.61	0.62
Average	0.6	0.57	0.58

Table 3: Evaluation of emotion classification of tweets via ML classifier.

Evaluation results for the classification of tweets using the symbolic classifier are presented in Table 4. Figures are significantly lower than those obtained via machine learning (17% points lower F-score) and also lower than the proposed baseline (8% points lower F-score). The relatively low performance of the symbolic classifier can be explained by the fact that the system was not developed for this particular type of corpus (it was initially developed to extract emotional responses – to products or brands etc. – provided in user-generated feedback). Indeed, the symbolic classifier proves less robust when faced with texts from a different domain than that for which it was developed.

	Precision	Recall	F-score
Anger	0.75	0.33	0.46
Disgust	0.76	0.24	0.37
Fear	0.72	0.35	0.47

Joy	0.26	0.68	0.37
Sadness	0.24	0.37	0.29
Surprise	0.84	0.37	0.52
Average	0.60	0.39	0.41

Table 4: Evaluation of emotion classification of tweets via symbolic classifier.

6.2 Tweet relevance

To evaluate the performance of the classifier in detecting emotional tweets in the ETCR corpus, we ran the classifier with differing score thresholds for emotion detection. For example, with a threshold set at 0.4, at least one emotion must have a score above 0.4 for the tweet to be classified as emotional. The reasoning behind this is that, for a tweet with no emotional content, one would expect the classifier to attribute equal scores to all emotions and, in the case of emotional ones, different scores for each emotion. By varying the threshold score for emotion classification we hoped to determine the optimal score for detecting the emotional relevance of a tweet. The graph in **Erreur ! Source du renvoi introuvable.** shows the F-score results of the evaluation for each of the tested thresholds. As expected, lower score thresholds favored recall, with relatively low precision, while precision, although still mediocre, was better at higher thresholds. Best performance for the classifier was an F-score of 0.26.

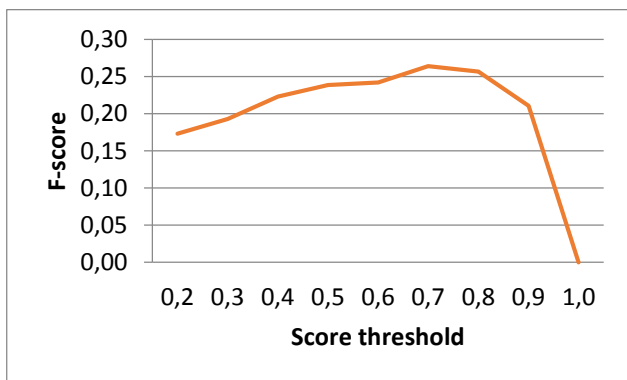


Figure 1: F-score for the detection of emotional versus non-emotional tweets with respect to minimum score threshold.

Notice that always assigning the non-emotional category to tweets in the corpus provides the baseline figures shown in Table 5.

	Precision	Recall	F-score
Emotional	0	0	0
Non-emotional	0.91	1.00	0.95
Average	0.45	0.50	0.48

Table 5: Baseline figures for detecting emotional versus non-emotional tweets.

Evaluation figures for the symbolic methods to detecting emotional versus non-emotional tweets (Table 6) show a

⁴ In absolute terms, our results for distinguishing emotional from non-emotional tweets are similar to the averaged figures of (Suttles & Ide, 2013). Surprisingly, however the results they obtain are derived from the application of a

major improvement over the above baseline (0.72 versus 0.48 average F-score). The figures also show a major improvement over those obtained by the classifier (F-score of 0.48 versus 0.26 for detection of emotional tweets).⁴

	Precision	Recall	F-score
Emotional	0.54	0.43	0.48
Non-emotional	0.94	0.96	0.95
Average	0.74	0.69	0.72

Table 6: Figures for detecting emotional versus non-emotional tweets with symbolic methods.

7. Conclusion & Future Work

We described two corpora – ETCC for tweet emotion classification and ETCR for tweet emotion relevance – created in the hope of providing evaluation data that avoids problems of over-subjectivity. This data was used to train and test a machine learning classifier, and test a rule-based classifier in the tasks of emotion classification and tweet relevance with respect to emotion. Evaluation results showed that a symbolic approach outperforms a classifier for determining the relevance (emotional versus non-emotional) of tweets, while the ML classifier is better adapted to the task of tweet emotion classification. A logical step for future work would be to combine both approaches in a hybrid system to process a stream of tweets – retrieving pertinent tweets via a symbolic system and classifying them with a ML classifier. Ensuring the quality of the annotated data would also be beneficial: a manual verification and correction of the ETCC “silver standard” corpus would provide higher quality data and checking inter-annotator agreement for the ETCR corpus would provide a means of verifying the relatively “objective” nature of the annotation task we defined.

8. References

- Alm, C.O., Roth, D. and Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of HLT/EMNLP*, Vancouver, BC, Canada.
- Bittar, A., Dini, L., Maurel, S. and Ruhlmann, M. (2014). The Dangerous Myth of the Star System. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, May 2014, Reykjavik, Iceland.
- Chambers, N., Cer, D., Grenager, T., Hall, D., Kiddon, C., MacCartney, B., de Marneffe, M.-C., Ramage, D., Yeh, E. and Manning, C.D. (2007). Learning Alignments and Leveraging Natural Logic. *Proceedings of the Workshop on Textual Entailment and Paraphrasing*, pp. 165–170.
- Chang, A.X. and Manning, C.D. (2014). TokensRegex: Defining cascaded regular expressions over tokens. *Stanford University Technical Report*.

Dickinson, T. (2015). Scraping Tweets Directly from Twitter. <http://tomkdickinson.co.uk/2015/01/scraping->

threshold based Machine Learning technique similar to the one we described here, for which our performance figures are lower.

tweets-directly-from-tweeters-search-page-part-1.
Consulted 27 September 2015.

Dini, L., Bittar, A. and Ruhlmann, M. (2013). Approches hybrides pour l'analyse de recettes de cuisine DEFT, TALN-RECITAL 2013. In *Proceedings of Défi fouille de textes*, June 2013, Les Sables d'Olonne, France.

Marneffe (de) M.-C., Manning C. D. (2008). The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pp.1-8

Ekman, P. (1993). Facial Expression and Emotion. *American Psychologist* 48(4):384–392.

Maurel, S., Curtoni, P. and Dini, L. (2007). Classification d'opinions par méthodes symbolique, statistique et hybride. In *Proceedings of Défi fouille de textes*, July 2007, Grenoble, France.

Mohammad, S. (2012). #Emotional Tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*Sem)*, June 2012, Montreal, Canada.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Malta.

Plutchik, R. (1980). Emotion: Theory, research, and experience. In *Theories of Emotion*, vol. 1. Academic Press, New York, NY, USA.

Strapparava, C. and Mihalcea, R. (2007). SemEval-2007 Task 14: Affective Text, in *Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval 2007)*, June 2007, Prague, Czech Republic.

Suttles, J., Ide, N. (2013). Distant Supervision for Emotion Classification with Discrete Binary Values. In Gelbukh, A. (ed.), *Computational Linguistics and Intelligent Text*, Lecture Notes in Computer Science, Springer, 121-136.

Valitutti, R. (2004). WordNet-Affect: an Affective Extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Vo, H. and Collier, N. (2013). Twitter Emotion Analysis in Earthquake Situations. *International Journal of Computational Linguistics and Applications*, 4(1):159-173.

Wiebe, J., Wilson, T. and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, volume 39(2-3):165-210.

Zhuang, Y. (2014). Building a complete Tweet index. <https://blog.twitter.com/2014/building-a-complete-tweet-index>. Consulted 28 September 2015.