UNIVERSITEIT
GENT

biblio.ugent.be

The UGent Institutional Repository is the electronic archiving and dissemination platform for all UGent research publications. Ghent University has implemented a mandate stipulating that all academic publications of UGent researchers should be deposited and archived in this repository. Except for items where current copyright restrictions apply, these papers are available in Open Access.

This item is the archived peer-reviewed author-version of:

Influence of Weak Labels for Emotion Recognition of Tweets

Olivier Janssens, Steven Verstockt, Erik Mannens, Sofie Van Hoecke, and Rik Van de Walle

In: Mining Intelligence and Knowledge Exploration, 108-118, 2014.

http://link.springer.com/chapter/10.1007%2F978-3-319-13817-6_12

To refer to or to cite this work, please use the citation to the published version:

Influence of Weak Labels for Emotion Recognition of Tweets

Olivier Janssens, Steven Verstockt, Erik Mannens,
Sofie Van Hoecke, and Rik Van de Walle

Multimedia Lab – Ghent University – iMinds,
Gaston Crommenlaan 8, 9050 Ledeberg-Ghent, Belgium
`{odjansse.janssens,steven.verstockt,erik.mannens`
`sofie.vanhoecke,rik.vandewalle}@ugent.be`
`http://www.mmlab.be/`

Abstract. Research on emotion recognition of tweets focuses on feature engineering or algorithm design, while dataset labels are barely questioned. Datasets of tweets are often labelled manually or via crowdsourcing, which results in strong labels. These methods are time intensive and can be expensive. Alternatively, tweet hashtags can be used as free, inexpensive weak labels. This paper investigates the impact of using weak labels compared to strong labels. The study uses two label sets for a corpus of tweets. The weakly annotated label set is created employing the hashtags of the tweets, while the strong label set is created by the use of crowdsourcing. Both label sets are used separately as input for five classification algorithms to determine the classification performance of the weak labels. The results indicate only a 9.25% decrease in f1-score when using weak labels. This performance decrease does not outweigh the benefits of having free labels.

Keywords: Emotion recognition, Twitter, Annotation, Crowdsourcing.

1 Introduction

Emotions influence everything in our life, e.g. relationships and decision making and are therefore analysed in many research projects. The automatic detection of emotions in text allows for a broad range of applications, such as forecasting movie box office revenues during the opening weekend [14], 3D facial expression rendering based on recognized emotions in text [5], stock prediction [4], and helping to understand consumer views towards a product [3].

The main goals of emotion recognition is to get implicit feedback about certain events, people's actions, services and products. For example, to see if a student has trouble learning, or finds the course too easy, it is possible to monitor the user's emotions while working on an e-learning platform. It is also possible to monitor the user's emotion during a game, allowing the difficulty to be adapted automatically to the user. It is also possible to monitor the user's opinion about products via social media in order to enhance recommender engine user profiles, so that a better recommender engine can be built.

Emotion recognition is either done with supervised machine learning methods, where a labelled dataset is required, or unsupervised machine learning methods, where no labels for the data are provided. Research shows that supervised methods tend to outperform unsupervised methods [10]. In order to gather labelled data, several methods are employed:

A first method is a strong labelling method because it consists of manually labelling the collected corpus. An example of this can be found in the work of Roberts et al. [13], where 7,000 tweets are manually labelled. This method has several disadvantages. First of all, annotating a large dataset takes a lot of time and effort. Secondly, because annotating data manually takes a lot of time, often the dataset will be small. Finally, since there is little data, mistakes will have a bigger impact on the classifiers' performance.

To create bigger datasets crowdsourcing can be used, which is also a strong labelling method and is employed regularly in machine learning research. Crowdsourcing services allow for data to be manually labelled by a large group of people. Crowdsourcing is an attractive solution since it can provide an annotated dataset possibly cheap and easy. Different methods exist to crowdsource labels. One can build his own application and attract annotators, but it is also possible to use existing services such as Mechanical Turk or Crowdflower.

Another frequently used labelling method is a weak labelling method and consist of using the emotion hashtag, or an emotion linked to the hashtag, as label. This method requires the collection of tweets based on their emotion hashtags which are added by the author of the tweet. For example, in the work of Mohammad [10], the 6 basic emotions of Ekman [7] are used as hashtags (#anger, #disgust, #fear, #happy, #sadness, and #surprise) to query the twitter API to create a dataset. This method allows for a large dataset to be created without any costs in a short amount of time, though the assignment of these weak labels can be questioned [2]. For example, the hashtag can be used to indicate sarcasm, resulting in a hashtag that does not correspond with the true underlying emotion. Another problem when using a weak labelling method for tweets is that the hashtags need to be removed in order to create the ground truth. As a result the tweet might lose its emotional value. For example the tweet "*No modern family tonight #sad*" has no underlying emotion when the hashtag is removed. However, if the hashtags are kept in the dataset, data leakage [9] occurs which results in an unrealistically good classifier.

In this paper the classification performance impact of weakly labelled emotion labels for emotion recognition is investigated. The analysis is done on a corpus of tweets for which two label sets are constructed. The weak label set is annotated by the emotion linked with the author's emotional hashtags. The strong label set is constructed by the use of crowdsourcing using Crowdflower. Manual annotation is not included in this analysis because the corpus consists of 341,931 tweets, which is too large for manual annotation.

The remainder of this paper is as follows. Section 2 presents the related work in text based emotion recognition. Section 3 elaborates on the dataset creation, the preprocessing methods and the classification algorithms. Next, Section 4

shows the evaluation results for the different classification algorithms. Finally, section 5 lists the conclusions and the future work.

2   Related Work on Text Based Emotion Recognition

Current state-of-the-art emotion recognition techniques use lexicon based methods in order to classify text [10],[16]. Lexicon based techniques use dictionaries of words with pre-computed scores expressing the emotional value of the word. The advantage of this technique is that there is a known relation between the word and the emotion. However, the disadvantage of this method is that often no new features are extracted from the relevant dataset such as emoticons [15]. Opposed to lexicon based techniques there are learning based techniques that require the creation of a model by training a classifier with labelled examples [8]. This means that the vocabulary for the feature vectors are built based on the dataset, resulting in a vocabulary, which is more suitable for tweets.

3   Methodology

In order to study the impact of weakly labelled data on text based emotion recognition, first, a corpus together with two label sets is created. Second, the corpus is preprocessed so that it can be used by the classification algorithms. Finally, the corpus, together with one of the two label sets, is used as input for five classification algorithms for which several objective metrics are computed.

3.1   Dataset

The corpus consists of 341,931 English tweets, i.e. 171,485 tweets for the test set and 170,446 tweets for the training set. Every tweet is annotated with one of these emotions: 'anger', 'fear', 'joy', 'love', 'sadness', 'surprise' or 'thankfulness'. The dataset is a subset of the dataset collected by W. Wang et al. [17] where word lists for the emotions were constructed to query the Twitter Streaming API to collect tweets.

The strong label set for this corpus is created by Crowdflower. The job in Crowdflower is designed so that every tweet is labelled by at least 3 different people. Each person contributing to the job is given 10 test cases with known ground truth (which is not shown to the contributor) for training purposes before they could start the annotation job. Based on these test cases a trust score, which is the accuracy score achieved by a contributor on the set of test cases, is calculated. The trust score allows for Crowdflower to filter out bad annotators or automated programs to ensure that a high quality set of labels is created.

The weak label set for the corpus of tweets was collected by extracting the emotional hashtag from the tweet. For example the tweet "I hate when my mom compares me to my friends **#annoying**" is labelled as 'anger' since it contains the "#annoying" hashtag which can be found in the word list belonging to

'anger'. For more information on the dataset, such as the choice of emotions or how the dataset was split into a training set and test set, we refer to the work of W. Wang et al. [17]. Tweets with less than 5 words or URLs are discarded. Also the emotion hashtags themselves are removed in order to prevent data leakage.

Because the weak label set and strong label set come from different sources, the distribution of labels are different as can be seen in Figure 1a and 1b. Both label sets were split according to the training and test set of the corpus.
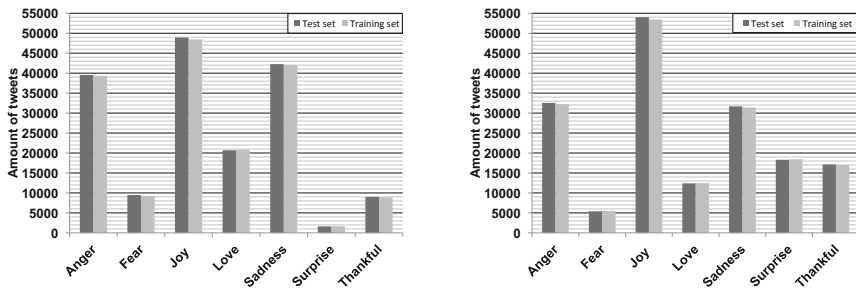
(a) Distribution of the weak labels constructed by the emotional hashtags

(b) Distribution of the strong labels constructed by crowdsourcing

Fig. 1. Label distributions

The training set is representative for the test set for both label sets as they have a similar distribution. However, there are some differences between the distribution of the strong labels and weak labels. For example, the weak label set has less tweets in the category 'surprise' compared to the strong label set. There are also more tweets associated with 'thankfulness' in the strong label set compared to the weak label set.

In order to improve the quality of the strong label set further, an additional filtering process is applied. This process uses the confidence score provided by Crowdflower. The confidence score describes the agreement between annotators and is calculated as follows: different users annotate the same tweet with one of the provided labels. In order to get a confidence score, all the trust scores of the users who voted for the same label for a tweet are added together. This number is then divided by the sum of all the trust scores of all the user who annotated that tweet. As a result the confidence score is always between 0 and 1.

Because of this confidence score, it is possible to filter out tweets and their respective labels which have no underlying emotion or are ambiguous. In the test set part of the corpus, only tweets are kept with a confidence score larger than 66%, which indicates that 2 of the 3 annotators agree on the emotion of a tweet. For the training set, tests with different ranges of confidences scores were done e.g.: >50%; >70%; >90%. The best result is achieved when tweets with a confidence score equal to 100% are kept in the training set because this results in less noise in the features as only the best features remain.

The resulting distributions of the reduced label sets can be found in Figures 2a and 2b. As result of the filter process based on the confidence score, the reduced training set (48,985 tweets) is smaller than the reduced test set (114,273 tweets) Overall it can still be stated that for both the reduced strong label set and the reduced weak label set, the test set distribution is similar to the training set distribution, although downscaled.
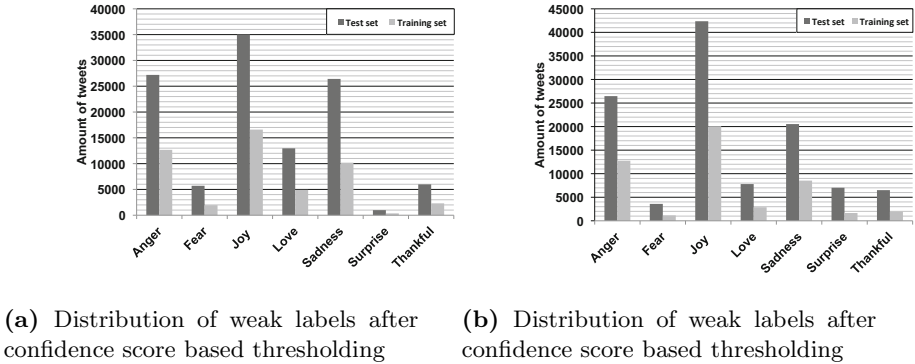

(a) Distribution of weak labels after confidence score based thresholding

(b) Distribution of weak labels after confidence score based thresholding

Fig. 2. Label distributions of reduced sets

Another insight is given by Table 1. Since there is one corpus and two sets of labels, it is possible to determine how the label sets differ. The table illustrates how the labels from the weak label set are redistributed in the strong label set. For example, from all the tweets labelled as 'joy' in the weak label set; 70.80% are labelled as 'joy' in the strong label set. Some other notable facts are:

Table 1. Redistribution of the weak label set to the strong label set

		From weak labels						
		Joy	Fear	Sadness	Thankfulness	Anger	Surprise	Love
	Joy	70.80%	5.78%	3.53%	22.89%	1.62%	16.73%	19.63%
	Fear	3.53%	69.77%	2.94%	1.66%	1.37%	5.53%	1.32%
	Sadness	5.61%	12.97%	74.33%	4.72%	16.06%	40.84%	6.12%
To strong labels	Thankfulness	4.31%	0.59%	0.56%	53.22%	0.20%	1.36%	6.59%
	Anger	2.74%	9.13%	15.06%	2.79%	79.28%	20.40%	2.21%
	Surprise	0.70%	0.21%	0.23%	0.35%	0.18%	8.12%	0.30%
	Love	12.31%	1.56%	3.34%	14.37%	1.30%	7.03%	63.83%
	Sum	100%	100%	100%	100%	100%	100%	100%

– A large part (12.31%) of tweets labelled as 'joy' according to the hashtag, are labelled as 'love' by crowdsourcing. Conversely, a large part (19.63%) of tweets labelled as 'love' according to the hashtag, are labelled as 'joy' by crowdsourcing.

- A large part (15.06%) of tweets labelled as 'sadness' according to the hashtag, are labelled as 'anger' by crowdsourcing. Conversely, a large part (16.06%) of tweets labelled as 'anger' according to the hashtag, are labelled as 'sadness' by crowdsourcing.
- A large part (12.97%) of tweets labelled as 'fear' according to the hashtag, are labelled as 'sadness' by crowdsourcing.
- Almost half of all tweets labelled as 'thankfulness' according to hashtag, are labelled as other emotions by crowdsourcing, with 'joy' being the largest one.
- Tweets annotated with the emotion 'surprise' according to the hashtag are distributed amongst the other emotions by crowdsourcing. Only 8.12% receives the 'surprise' label and a large part is labelled as 'sadness' (40.84%). Since there is only a 8.12% overlap in the label sets for the category 'surprise', it can be stated that the category 'surprise' in the weak label set contains of a lot of noise. Because the tweets were initially collected using words in lists linked to the categories [17], it is our belief that the noise stems from the chosen words in the word list for 'surprise'. For example the tweet "Got to see him today #unexpected" has the weak label 'surprise' since it was collected by the hashtag "#unexpected" which is listed in the 'surprise' word list. Nevertheless the tweet is labelled with the strong label 'joy' by crowdsourcing. Another example is the tweet "5-1 Edmonton over Blackhawks nearing end of the 1st Period #ASTONISHED". This tweet received the weak label 'surprise' because "#astonished" is present in the word list for 'surprise', nevertheless it is labelled as 'sadness' by crowdsourcing. Because of these observation it is our belief that the collection of the tweets based on certain hasthags present in the word list for the 'surprise' don't always return unambiguous tweets.

To summarize, 31.26% of the tweets are labelled differently by crowdsourcing compared to the weak label set. Generally, tweets in both label sets are labelled with the same valence, i.e., negative valence or positive valence. 'Joy', 'love', 'thankfulness' have a positive connotation, 'fear', 'sadness', 'anger' have a negative connotation and 'surprise' is neutral. 89.72% of the tweets with a positive connotation according to the weak label set also have a positive connotation in the strong label set. 93.83% of the tweets with a negative connotation according to the weak label set also have a negative connotation in the strong label set. These numbers indicate that only a small fraction of the labels differ when it comes to valence. The differences between the label sets will have an impact on the classification results, thus supporting the research question of this paper to investigate the impact of weak emotions labels on the classification results compared to strong emotion labels.

3.2 Preprocessing

In this paper a learning based technique is used to classify the tweets. The corpus is preprocessed first in order to transform it into feature vectors. Preprocessing consists of stemming, which reduces every word of every tweet to their stem if

possible. An example: kneeling, kneeled, kneels are all reduced to kneel. As the feature space, when using learning based methods, can grow large very quickly, reducing conjugations of the same word to their stem reduces this feature space. In this paper, the Porter stemmer [18] is used because it outperforms other widely used stemmers such as the Paice-Husk [11] and Lovins [1] stemmers.

The next step transforms the tweets to feature vectors.

3.3   Feature Extraction Methods

The feature extraction methods transform words and sentences into a numerical representation which can be used by the classification algorithms. In this research, the combination of N-grams and TF-IDF feature extraction methods is used. As will be motivated below, this will preserve syntactic patterns in text and help solve class imbalance respectively.

N-gram Features: In the field of computational linguistics, an N-gram is a contiguous sequence of N items from a given sequence of text. An N-gram of size 1 is referred to as a "unigram", size 2 is a "bigram", size 3 is a "trigram". An N-gram can be any combination of letters, words or base pairs according to the application. In this paper, a combination of 1,2 and 3 grams of words is used and passed to the Term Frequency-Inverse Document Frequency algorithm.

Term Frequency-Inverse Document Frequency: One of the broadly used feature representation methods is the bag of words representation together with word occurrence [6], equalizing a feature value to the number of times it occurs. This method is fairly straightforward and easy, though it has a major downside as longer documents have higher average count values. In this paper, the tweets are transformed to feature vectors by using the bag of word representation together with word occurrence. The number of occurrences of each word are then normalized to create the so called Term Frequencies. Additionally, weights for words that occur in many documents are downscaled making them less dominant than those that occur only in a small portion of the corpus. The combination of term frequencies and downscaling weights of dominant words is called Term Frequency-Inverse Document Frequency (TF-IDF).

By composing the vocabulary for the feature vectors based on the dataset, the feature vectors will be very high dimensional. In this case the feature vectors consist of 48,036 features. However by combining TF-IDF and the N-grams method, a feature vector will be very sparse. This can be beneficial if the used algorithms can work with sparse data.

3.4   Classification Algorithms

At this point the tweets have been transformed to feature vectors. In this step the corpus and label sets are used as input for various machine learning algorithms. In this paper, five different classification algorithms are compared to ensure objective determination of the impact of the different label sets for the corpus.

The algorithms compared here are the ones frequently used in text classification as they deliver good results and work very fast. [12]

1. **SGD:** A linear classifier which uses stochastic gradient descent with the modified huber loss. SGD is a method where the gradient of the loss function is estimated each sample at a time. This means that not all the samples have to be loaded into the memory at once. This results in an algorithm that can be used on large datasets that normally do not fit in the memory. The choice of the loss function influences how the data will be handled. For example the modified huber loss reduces the effect of the outliers much more than the log loss function. Additionally, L2 regularization is added to reduce overfitting.
2. **SVM:** A linear support vector machine from the LIBLINEAR library. This SVM is optimised to work with datasets with a large number of samples.
3. **MNB:** Multinomial Naive Bayes, the Naive Bayes algorithm implemented for data where the features are assumed to have a multinomial distribution.
4. **NC:** Nearest Centroid classifier, which is known as the Rocchio classifier when using TF-IDF feature vectors for text classification. In a nearest centroid classifier, every class is represented by a centroid and every test sample is classified to the class with the nearest centroid.
5. **Ridge** A linear classifier that uses regularized least-squares.

Since linear classifiers are designed for binary classification problems, the one versus all method is used to combine them for the multi-class problem presented here. The classification results can be found in the section below.

4 Results

The results are subdivided in 2 subsections. The first subsection elaborates on the classification results by the comparison of different objective metrics. The second subsection discusses the confusion matrices of the best classifier.
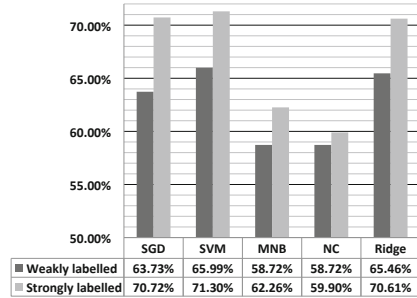
4.1 Classification Metrics

For the comparison of the five classification algorithms, three metrics are compared, namely weighted precision, weighted recall and weighted f1-score. The most important metric here is the weighted f1-score because it incorporates the recall and the precision score, and a good classifier will maximize both. Since the classes are highly imbalanced in both label sets, the accuracy score is not included. For example, if the classifier labels all the samples as the majority class, a good accuracy score would still be achieved, though in fact the classifier was not able to learn the underlying pattern of the data.

The results can be found in Figure 3a, 3b and 3c respectively. For almost every algorithm and metric, the tests with the strong labels delivers better results. The best result for the weighted f1-metric is given by the SGD algorithm with a modified huber loss where a difference of 6.53% between the results using the weakly labelled label set and the strongly labelled label set is noticed. By taking
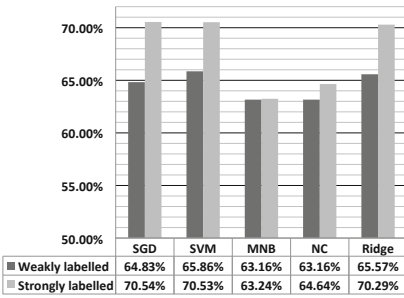
(a) Weighted precision

(b) Weighted recall
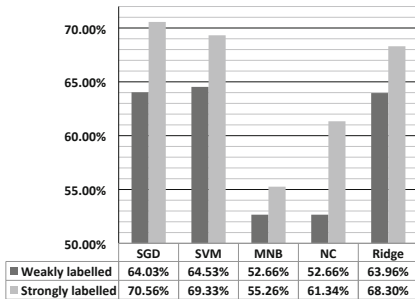
(c) Weighted f1-score

Fig. 3. Classification metrics comparison of the five classification algorithms

the strong labels as the ground truth of the dataset, a decrease of 9.25% of f1-score occurs when using weakly labelled data.

4.2   Confusion Matrices

To see if the algorithms confuse any emotions, which often is the case when there is a significant imbalance in the label sets, confusion matrices of the SGD classifier for both label sets tests are presented in Figure 4a and 4b.

Both classifiers display very little confusion, this can be attributed to the fact that tweets with possible ambiguity and/or which have no underlying emotion are filtered out using the confidence score.

In our previous work [8], a dataset of 498,885 tweets, annotated with the emotion hashtags provided by the authors of the tweets was used. In that case, no confidence score could be calculated to filter out tweets. The results showed that the classifiers confused some emotions, as it was possible that tweets were not annotated with the correct emotion, or did not have an underlying emotion at all, or were ambiguous. The classifiers of this paper do not show this shortcoming.
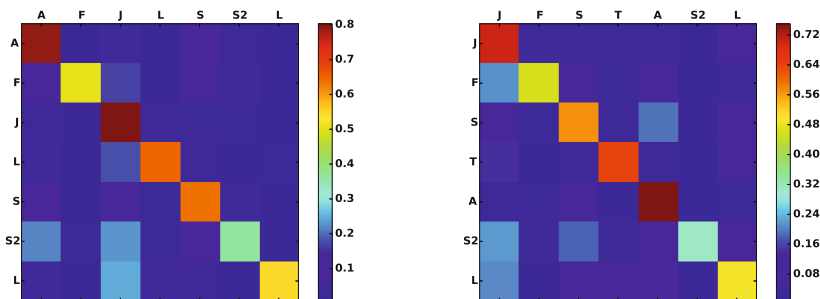
(a) SGD classifier with modified huber loss when using the stong labels. A='anger', F='fear', J='joy', L='love', S='sadness', S2='surprise', T='thankfulness'

(b) SGD classifier with modified huber loss when using the weak labels. A='anger', F='fear', J='joy', L='love', S='sadness', S2='surprise', T='thankfulness'

Fig. 4. Confusion matrices

5   Conclusion and Future Work

In this paper the impact of weak emotion labels on classification results is studied by comparing the classification results of five classification algorithms using both a weak label set and a strong label set. Both label sets are created for the same corpus of tweets. However one set is created by using the emotion hashtag of the tweet, and the other set is constructed by a crowdsourcing service. A filter process is applied on the corpus to eliminate tweets and their corresponding labels in both label sets which have a low confidence.

The results show that, when using weak labels, there is only a 9.25 % decrease in f1-score compared to the results when using strong labels. This disadvantage does not outweigh the benefits of weak labels, i.e. it is available free of charge and requires almost no extra work to collect. Also, since hashtags and tweets are published together, a continues stream of labelled data is created. This is useful for online learning algorithms, such as SGD, since it gives access to a vast amount of labelled data, which can result in improved classification results [17].

Also, it should be noted that emotion recognition of tweets remains a difficult task. The best weighted f1-score for the strong labels is 70,56%, leaving room for improvement.

Future work will consist of exploring new methods to create weak emotion label sets that better approximate the results of a strong emotion label sets. A possible improvement is to gathering tweets based on a combination of words related to an emotion instead of just the hashtag.

References

1. Lovins, J.B.: Development of a stemming algorithm. Mechanical Translation and Computational Linguistics 11, 22–31 (1968)
2. Bergamo, A., Torresani, L.: Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. Advances in Neural Information Processing Systems 23, 181–189 (2010)
3. Bollen, J., Mao, H., Pepe, A.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: ICWSM, pp. 450–453 (2011)
4. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science 2(1), 1–8 (2011)
5. Calix, R.A., Mallepudi, S.A., Chen, B.C.B., Knapp, G.M.: Emotion Recognition in Text for 3-D Facial Expression Rendering. IEEE Transactions on Multimedia 12(6), 544–551 (2010)
6. Chaffar, S., Inkpen, D.: Using a heterogeneous dataset for emotion analysis in text. In: Butz, C., Lingras, P. (eds.) Canadian AI 2011. LNCS, vol. 6657, pp. 62–67. Springer, Heidelberg (2011)
7. Ekman, P.: Basic emotions. Handbook of Cognition and Emotion, vol. 98. John Wiley & Sons (1999)
8. Janssens, O., Slembrouck, M., Verstockt, S., Hoecke, S.V., Walle, R.V.D.: Real-time Emotion Classification of Tweets. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1430–1431 (2013)
9. Kaufman, S., Rosset, S.: Leakage in data mining: Formulation, detection, and avoidance. In: 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 556–563 (2012)
10. Mohammad, S.: #Emotional Tweets. In: First Joint Conference on Lexical and Computational Semantics, pp. 246–255. Association for Computational Linguistics, Montréal (2012)
11. Paice, C., Husk, G.: Another Stemmer. ACM SIGIR Forum 24, 56–61 (1990)
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
13. Roberts, K., Roach, M.A., Johnson, J.: EmpaTweet: Annotating and Detecting Emotions on Twitter. In: LREC, pp. 3806–3813 (2012)
14. Rui, H., Whinston, A.: Designing a social-broadcasting-based business intelligence system. ACM Transactions on Management Information Systems 2(4) (2011)
15. Suttles, J., Ide, N.: Distant supervision for emotion classification with discrete binary values. In: Gelbukh, A. (ed.) CICLing 2013, Part II. LNCS, vol. 7817, pp. 121–136. Springer, Heidelberg (2013)
16. Wang, A., Hoang, C., Kan, M.: Perspectives on crowdsourcing annotations for natural language processing. Language Resources and Evaluation 47(1), 9–31 (2013)
17. Wang, W., Chen, L., Thirunarayan, K., Sheth, A.P.: Harnessing Twitter "Big Data" for Automatic Emotion Identification. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, pp. 587–592 (2012)
18. Willett, P.: The Porter stemming algorithm: then and now (2006), http://dx.doi.org/10.1108/00330330610681295