Dystemo: Distant Supervision Method for Multi-Category Emotion Recognition in Tweets

VALENTINA SINTSOVA and PEARL PU, Swiss Federal Institute of Technology (EPFL)

Emotion recognition in text has become an important research objective. It involves building classifiers capable of detecting human emotions for a specific application, for example, analyzing reactions to product launches, monitoring emotions at sports events, or discerning opinions in political debates. Most successful approaches rely heavily on costly manual annotation. To alleviate this burden, we propose a distant supervision method—Dystemo—for automatically producing emotion classifiers from tweets labeled using existing or easy-to-produce emotion lexicons. The goal is to obtain emotion classifiers that work more accurately for specific applications than available emotion lexicons. The success of this method depends mainly on a novel classifier—Balanced Weighted Voting (BWV)—designed to overcome the imbalance in emotion distribution in the initial dataset, and on novel heuristics for detecting neutral tweets. We demonstrate how Dystemo works using Twitter data about sports events, a fine-grained 20-category emotion model, and three different initial emotion lexicons. Through a series of carefully designed experiments, we confirm that Dystemo is effective both in extending initial emotion lexicons of small coverage to find correctly more emotional tweets and in correcting emotion lexicons of low accuracy to perform more accurately.

CCS Concepts: • **Information systems** → **Sentiment analysis**; Web log analysis; • **Computing methodologies** → **Semi-supervised learning settings**; *Natural language processing*; *Information extraction;*

Additional Key Words and Phrases: Distant supervision, emotion recognition, natural language processing, semi-supervised learning, text mining, twitter

ACM Reference Format:
Valentina Sintsova and Pearl Pu. 2016. Dystemo: Distant supervision method for multi-category emotion recognition in tweets. ACM Trans. Intell. Syst. Technol. 8, 1, Article 13 (August 2016), 22 pages.
DOI: http://dx.doi.org/10.1145/2912147

1. INTRODUCTION

The abundance of emotions that human beings can feel is often reflected in our language. We use emotionally charged expressions (e.g., "Yay! We did it!") or explicit

This work was made possible by the research funding of EPFL.
Summary of differences from the published workshop paper [Sintsova et al. 2014]: Previously, we introduced a suggested framework for distant learning and provided a preliminary evaluation. The present paper describes a far more complete study of the suggested methodology. The following specific changes were made. In addition to the macro-metrics previously used for evaluation, we included metrics to evaluate the performance of classifiers at the micro-level. We replaced the original dataset for evaluation with one having a more realistic distribution of emotions. Additionally, we included a dataset with manual labels. These evaluation changes help to better estimate the extent to which the obtained classifiers are applicable to real data. We also include other supervised learning methods for comparison with the suggested Balanced Weighted Voting method, and validate the inclusion of neutral tweets in learning.
Authors' addresses: V. Sintsova and P. Pu, School of Computer and Communication Sciences, Swiss Federal Institute of Technology (EPFL), CH-1015, Lausanne, Switzerland; emails: {valentina.sintsova, pearl.pu}@epfl.ch.

ACM 2157-6904/2016/08-ART13 $15.00
DOI: http://dx.doi.org/10.1145/2912147

statements (e.g., "So happy now") to verbally describe our emotional experiences. The objective of emotion recognition is to detect which emotions are expressed in a given text sample. For example, when someone writes "Today was awesome," the system should conclude that the author is happy. Such automatic recognition of emotional statements can help us build more intelligent social and personal applications, including those that study online conversations [Quercia et al. 2012], enhance human–computer interaction [Picard and Klein 2002], and summarize public reactions to events or products [Mishne and de Rijke 2006; Kempter et al. 2014].

One successful approach is to treat emotion recognition as a text classification problem and to apply supervised machine-learning techniques [Aman and Szpakowicz 2007; Strapparava and Mihalcea 2008; Wang et al. 2012; Roberts et al. 2012]. However, supervised methods require considerable quantities of annotated data (i.e., text documents with known emotion labels). Such manual annotation is time-consuming due to difficulties involved in judging potentially ambiguous and subjective emotions. Annotation becomes even more challenging when the goal is to differentiate more fine-grained categories of emotion [Desmet 2012; Kempter et al. 2014].

Our research objective was to develop a method to automatically build fine-grained emotion classifiers in the absence of manually annotated data. In this endeavor, we have resorted to distant learning (also known as distant or weak supervision)—a type of semi-supervised learning used by many researchers in text classification [Go et al. 2009; Purver and Battersby 2012; Mintz et al. 2009]. The main idea is to train the classifiers on the data with automatically assigned emotion labels. In contrast to traditional semi-supervised learning, where classifiers are learned over partially annotated data (i.e., a mixture of annotated and unlabeled data) [Zhu 2005], the distant learning approach requires no manual annotation. Instead, annotated data are obtained automatically using some emotion labelers that are able to detect emotions of interest in the *subset* of available text documents.

In the domain of social media, researchers have successfully applied distant learning for topic-independent emotion recognition while using emoticons (e.g., :) or >:() and emotional hashtags (e.g., #*happy* or #*angry*) as initial labels [Yang et al. 2007; Wang et al. 2012; De Choudhury et al. 2012b; Mohammad 2012; Purver and Battersby 2012; Suttles and Ide 2013]. They are considered to summarize emotions in the corresponding texts. However, such content cues are not always present in adequate amounts within specialized topics of discussion (sports, politics, finance, or education) and are likely to be absent in text documents other than social media (reviews, news articles, or technical comments). Instead of relying on hashtags or emoticons for labeling, we aim to design and investigate a distant learning method that is more generally applicable.

To address this challenge, we suggest using terms from existing or easy-to-produce emotion lexicons as initial labelers. For instance, for any set of emotion categories, we can use a list of descriptive emotional terms (such as "proud" for *Pride*) and label texts according to the presence of these terms. Using such lexicon-based initial labelers ensures the generality of our methods: as they are not restricted to specific types of content cues, we can potentially detect emotional content within documents of any type or topic. A distant learning algorithm will then discover emotion associations of new terms based on their co-occurrences with given emotional terms. For example, it can recognize the phrase "well done" as an indicator of *Pride* emotion, if this phrase appears often enough together with known pride-related words, such as with the word "proud" in the text "So proud 2 be British! huge well done 2 all of Team GB! :D".

With this idea, we have developed Dystemo, a distant supervision method that generates fine-grained emotion classifiers from documents pseudo-labeled by some initial lexicon of limited coverage, accuracy, or both. We focus on recognizing emotions in tweets—short status updates from a popular social media website, Twitter. Twitter

contains discussions on a variety of topics and events, and provides an easy opportunity to collect large datasets. Two main novelties lead to the success of the proposed method. First, we suggest a new Balanced Weighted Voting (BWV) algorithm that incorporates per-category rebalancing coefficients while learning. This overcomes the intrinsic imbalance of emotion distribution in initial pseudo-labeled dataset, which if left untreated can cause classifier's bias toward dominant emotions. Second, using social media as a source of textual data allowed us to include simple heuristics for detecting non-emotional (or neutral) tweets. These tweets turned out to be indispensable for training classifiers to discern neutral tweets from emotional ones. Both of these novelties significantly increase the accuracy of final emotion classifiers.

We validate the suggested method on tweets in the field of sports events using the fine-grained model containing 20 emotion categories. We show that with Dystemo we obtain the final classifiers of substantially better quality than the three tested initial emotion lexicons (the relative increase of micro-F1 score is between 41% and 236% on the large hashtag-based ground-truth data). In comparison with other distant learning algorithms, Dystemo achieves the best micro-F1 scores with two out of the three initial lexicons on the hashtag-based data, and shows competitive performance on small manually annotated data.

In summary, to the best of our knowledge, Dystemo is the first distant learning method for producing fine-grained emotion classifiers without the help of manually labeled text, nor of structured content features such as emoticons or hashtags. It relies on terms from emotion lexicons instead. Our carefully designed experiments confirm the viability of this approach, at least within the domain of tweets.

This article is organized as follows. Section 2 reviews related work. Section 3 gives an overview of the proposed distant supervision method. Section 4 describes the setup for applying the method: which emotion model and data were used and which initial lexicons were considered. Section 5 introduces our method's evaluation framework, presenting ground-truth data, performance metrics, and other classifiers for comparison. Section 6 presents experimental results on the methodological validation. The last two sections discuss the findings and future work, followed by the conclusion.

2. RELATED WORK

Emotion recognition in text is an increasingly popular sub-topic in sentiment analysis. It aims to extract personal opinions, sentiments, and feelings expressed in text [Pang and Lee 2008; Liu 2012]. While borrowing many methods from polarity and multi-category text classification problems, emotion recognition has evolved into a distinct field of research due to the multiplicity of ways to express and discern emotions in language.

Affective Linguistic Resources. Emotions can be detected by spotting the words used for expressing them, such as *happy*, *angry*, or *inspired*. Associations between linguistic terms and emotions are given in emotion (or affective) lexicons. These are similar to sentiment lexicons, which store terms' polarities for polarity classification and opinion mining, such as positive *good*, *great*, and *awesome* [Taboada et al. 2011; Thelwall et al. 2012]. Some emotion lexicons include only terms directly expressing an emotion, such as "happy" for *Happiness* (the GALC lexicon is an example [Scherer 2005]). Others contain terms indicative of an emotional experience, thus more indirectly expressing an emotion. Example terms linked to *Happiness* are "approval" in WordNet-Affect [Strapparava and Valitutti 2004], "entertain" in NRC [Mohammad and Turney 2013], and "visit friend" in EmoSenticNet [Poria et al. 2013]. Extracting features using emotion lexicons has shown promise for various text classification applications, including prediction of reviews' helpfulness [Martin and Pu 2014], personality detection

[Poria et al. 2014; Mohammad and Kiritchenko 2015], polarity classification [Bravo-Marquez et al. 2013; Carrillo-de Albornoz and Plaza 2013], and emotion classification itself [Mohammad 2012; Wang et al. 2012]. Rule-based algorithms go beyond simple term-spotting by taking into account syntactic structures, such as the presence of negations, intensity modifiers, and conjunctions [Neviarouskaya et al. 2011; Krcadinac et al. 2013]. While the above-mentioned lexicon-based methods can be applied to any textual data, they are unlikely to cover the full variety of emotional expressions used in language. This leaves room for more advanced methods to increase the quality of emotion recognition.

Semi-Supervised Extension of Emotion Lexicons. Researchers have developed semi-supervised techniques to extend initial general (but limited) emotion lexicons, considered as seeds. These methods define several metrics of term similarity and then use them to cluster new terms into emotion categories based on their similarity to the seeds. The original WordNetAffect lexicon [Strapparava and Valitutti 2004] and one part of the Synesketch lexicon [Krcadinac et al. 2013] were built in this way, starting from a small number of explicit emotional terms. Similarity metrics were defined using semantic relationships (such as synonymy). In the construction of the EmoSenticNet lexicon [Poria et al. 2012, 2013, 2014], additional term similarities were derived from term co-occurrences in the database of emotional experiences by using Pointwise-Mutual Information (PMI) [Turney and Littman 2003]. Other corpora used to construct emotion lexicons, using PMI-based scores and starting from a small number of seed emotional keywords or symbols, were web *n*-grams [Perrie et al. 2013], sentences from web-logs [Yang et al. 2007], and tweets [Mohammad 2012]. Such lexicon-growing methods can, therefore, increase the coverage of used emotional expressions. Instead of focusing on term-level emotion associations, our method aims at building document-level emotion classifiers. Nevertheless, for comparison, we adapt PMI-based computation of term emotion scores [Mohammad 2012] to be applied within our framework.

Supervised Emotion Recognition. Constructing emotion classifiers automatically is possible by applying supervised machine-learning algorithms over labeled data. Researchers have experimented with different classifiers (such as Naïve Bayes and SVM) and with various linguistic, stylistic, and syntactic features (such as *n*-grams, punctuation marks, parts of speech, and topics). These experiments were performed in different domains, including web-logs [Aman and Szpakowicz 2007], fairy tales [Alm et al. 2005], news headlines [Strapparava and Mihalcea 2008], and tweets [Mohammad 2012; Roberts et al. 2012]. However, such supervised techniques require substantial annotated data for training, which are expensive to obtain.

Distant Supervision in Emotion Recognition. With Twitter, many researchers overcome the lack of annotated data by crawling tweets with emotional hashtags, such as #*happy* or #*angry* [Mohammad 2012; Wang et al. 2012; De Choudhury et al. 2012b; Suttles and Ide 2013]. In accordance with the idea of distant supervision, such tweets serve as pseudo-labeled data and are used to train machine-learning classifiers in a supervised manner. Yet, only a small fraction of tweets is likely to contain such hashtags, making questionable the application of these restrictive heuristics throughout different datasets. In the present work, instead of using hashtags for the pseudo-labeling of tweets, we propose using more applicable initial labelers based on terms from a given emotion lexicon. The data labeled based on emotional hashtags are used only for automatically validating the constructed emotion classifiers.

Building emotion classifiers using a limited set of emotional terms and unlabeled data has been attempted before. One method is to represent the given text corpus in a

reduced-dimensionality vector-space model and assign emotions based on similarities to computed emotion vectors [Kim et al. 2010; Danisman and Alpkocak 2008].

These methods were validated for a small set of emotion categories, whereas we design a methodology capable of dealing with a much more complete set of emotion categories. Moreover, those methods disregarded the treatment of neutral tweets (i.e., tweets without emotions), while we design and successfully apply heuristics to help classifiers recognize neutral tweets.

Semi-Supervised Learning for Other Tasks. Many other algorithms have been designed for semi-supervised learning (Zhu [2005] gives an overview). For multi-category text classification, a commonly applied method is Naïve Bayes with the Expectation-Maximization procedure [Nigam et al. 2000]. It iteratively repeats two actions: first, it learns the parameters over the annotated data; second, it re-annotates the data using the learned parameters. In our experiments, we also applied a Naïve Bayes as one of the compared classifiers, but starting from the data that were pseudo-annotated by a given initial labeler.

We also review the advances in semi-supervised methods for polarity classification—a problem closely related to emotion recognition. Experiments show semi-supervised classifiers outperform supervised ones when few labeled data are available [Wiegand and Klakow 2009]. The idea of distant learning for building polarity classifiers has been successfully applied to Twitter data as well, where researchers use emoticons and hashtags as the sentiment pseudo-labels (positive or negative) and identify neutral tweets using objective hashtags or as tweets from the news websites [Go et al. 2009; Pak and Paroubek 2010; Kouloumpis et al. 2011]. Among other methods, an iterative self-training approach has been shown to be effective [Qiu et al. 2009]. To apply binary polarity classification methods to our multi-category emotion classification problem, we first split it into multiple independent binary classification problems, each distinguishing one emotion category from all the others. This setup allowed us testing machine-learning classifiers suitable for binary applications.

Overall, no related work has studied how to apply the distant supervision framework for multi-category emotion classification when neither manual labels nor labels from a content structure (e.g., hashtags) are accessible. This is the main problem tackled in this article.

3. DISTANT SUPERVISION METHOD—DYSTEMO

We first introduce the definitions used to describe the problem and our suggested method. We formulate the problem of emotion recognition as a multi-label classification task. Given the set of emotion categories $E = \{e_1, e_2, \ldots, e_{|E|}\}$, the classifier detects which emotion categories are expressed in a given document d—in our case, a tweet—and produces their label set $Y_d = \{e_{i_k}\} \subseteq E$. In order to separate neutral from emotional documents, this method uses the extended set of categories $E^0 = \{e_0\} \cup E$, where e_0 represents the *Neutral* label. If e_0 is within the multi-label output Y_d (i.e., $e_0 \in Y_d$) or if no emotion is present (i.e., $Y_d = \varnothing$), we assign a *Neutral* category e_0 alone ($Y_d = \{e_0\}$).

We also define the *emotionality* of the text $\vec{p} = (p_0, p_1, p_2, \ldots, p_{|E|})$ as the distribution of the emotion categories expressed in the text, with $\sum_{i=0}^{|E|} p_i = 1$ and $\forall i \; p_i \geq 0$, where p_i is the weight of the ith emotion. Emotionality can be transformed into a multi-label by applying a technique adapted from the alpha-cut for fuzzy sets [Bojadziev and Bojadziev 1995]. We denote this operator as $\mathfrak{A} : (\vec{p}, \alpha) \rightarrow 2^{E^0}$, where α defines a threshold on the emotion weight for the emotion to be included in the multi-label. $\mathfrak{A}(\vec{p}, \alpha)$ returns all the labels e_i that have the weight $p_i \geq \alpha \cdot p^*$, where $p^* = \max_i p_i$ is the maximum emotion weight within the distribution $\vec{p}$. Thus, all the labels with a weight close enough to the maximum weight are output. If $\alpha = 1$, only the labels with

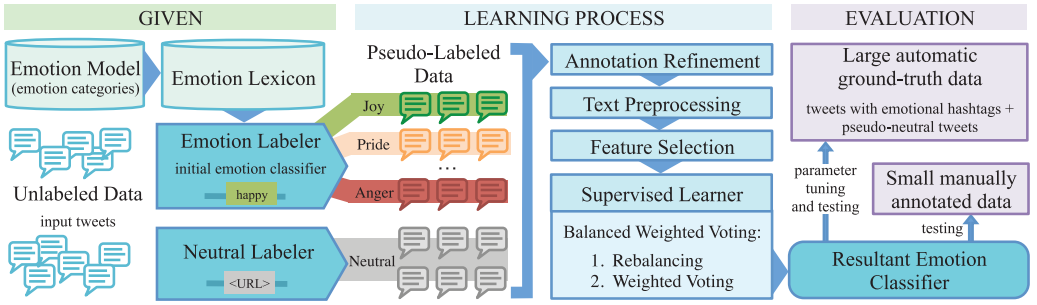

Fig. 1. The framework for our distant supervision method.

the maximum weight are output. For example, for the emotionality ($p_2 = 0.2,\ p_3 = 0.3,\ p_4 = 0.5,\ \forall i \neq 2, 3, 4\ p_i = 0$) the multi-label $\{e_3, e_4\}$ would be found with $\alpha = 0.5$. In the opposite direction, a multi-label Y_d can be transformed into the emotionality by specifying the weight of each label in Y_d as $\frac{1}{|Y_d|}$.

3.1. Method Input

Figure 1 shows an overview of our distant learning method, Dystemo.[1] It aims at building an emotion classifier for detecting emotions of the specified category set within a specific dataset of tweets (e.g., those on a certain topic). Correspondingly, as an input, it requires Twitter data collected for a desired application, denoted *unlabeled data* U, and emotion model specifying which category set E to recognize. The method also requires *emotion* and *neutral labelers*. The core of an emotion labeler is an emotion lexicon containing associations of linguistic expressions (terms) to the emotion categories of interest. Then, the *emotion labeler* is a simple initial emotion classifier assigning emotions to tweets based on the occurrence of terms from the given lexicon. The *neutral labeler* in its turn aims at identifying neutral tweets. It is essential to have neutral tweets in the training set. Otherwise, we risk obtaining classifiers that identify almost every tweet as emotional (it will be shown in the experimentation section), which is unacceptable for a successful emotion recognition system. We suggest simple heuristics for detecting neutral tweets, namely based on the presence of URLs in the tweet and absence of potential emotional cues (they are described in detail in Section 4.4).

3.2. Initialization of Learning Process

The learning process starts with applying both *emotion* and *neutral labelers* to unlabeled data U to obtain the *pseudo-labeled data* L. We assume that the emotion labeler returns the emotionality $\vec{p}(d)$ for a given document $d \in U$, while the neutral labeler assigns a tweet to a neutral class e_0 by setting $p_0(d)$ to 1.0. Tweets detected by the neutral labeler are referred to as pseudo-neutral and are not considered to be labeled by the emotion labeler. Tweets from U where the emotion labeler found no emotion are not included in L, because they could be classified as neutral due to the lack of information about emotional expressions in the initial emotion lexicon. Overall, the pseudo-labeled data L comprise the set of tweets with mapped emotionalities, one part found by the emotion labeler, and another—by the neutral labeler.

The first step of actual learning process is *annotation refinement*. It is essential to apply it when emotion lexicons assign weights to terms, as we need to eliminate

[1]The source code along with the resultant emotion classifiers are publicly available for research purposes at http://hci.epfl.ch/dystemo.

annotations of emotions with relatively low weights. The refinement is applied to each tweet individually. Given the parameter α_{ref}, it sets to zero the weights of those emotions that would not be included in the multi-label: $e_i \notin \mathfrak{A}(\vec{p}, \alpha_{ref})$, and then normalizes the distribution. Whether or not to apply this refinement is a parameter of the method.

The second step is to *preprocess the texts* of the tweets used in learning. This includes extraction of emoticons and punctuation marks as separate tokens, lower-case transformation, and normalization of elongations. More details are given in the appendix.

The third step is to *extract* and *select features* over which the classifier will be learned. We use 1-, 2-, ..., n-grams as features. We exclude the n-grams containing only stopwords and mark n-grams as negated if a negation word is detected up to two words before them. Also, we only retain the n-grams that appeared K or more times in the pseudo-labeled dataset L. From these, we select terms that are indicative of emotions by estimating their polarity. We compute a term's semantic orientation using Pointwise-Mutual Information (PMI) [Turney and Littman 2003]. First, the polarity label (l^+ or l^-) of each tweet $d \in L$ is identified as $\text{sign}(\sum_{i \in E^+} p_i(d) - \sum_{i \in E^-} p_i(d))$, where $E^+ \subset E$ and $E^- \subset E$ are the corresponding sets of positive and negative emotions. Then, the semantic orientation $SO(t)$ of a term t is computed as

$$pmi(t, l^+) - pmi(t, l^-) = \log \frac{P(t, l^+)P(l^-)}{P(t, l^-)P(l^+)} = \log \left[\frac{1 + freq(t, l^+)}{1 + freq(t, l^-)} \cdot \frac{|V| + freq(l^-)}{|V| + freq(l^+)} \right], \quad (1)$$

where V is the set of extracted terms, $freq(l^{\pm})$ is the number of positive (l^+) or negative (l^-) tweets, and $freq(t, l^{\pm})$ is the number of tweets with the term t, which are either positive or negative. The formula uses smoothing: we add 1 to each term frequency computation, and $|V|$ to class frequency computations in order to compensate for the additions to term frequencies. The higher the absolute value of $SO(t)$, the more confident we are that the term t has strong polarity and is thus potentially emotional. We filter out the features that have an absolute score $|SO(t)|$ lower than a threshold τ. The remaining features are used for the feature representation of the tweets. As tweets are short, the terms' presence is used for features' values, instead of their frequency.

With the tweets represented as feature vectors and their associated emotionalities, the final resultant classifier can now be learned in a supervised manner. We apply BWV as the *supervised learner*. Its choice also defines how the *resultant classifier* will work.

3.3. Supervised Learner—Balanced Weighted Voting (BWV)

The BWV algorithm is a supervised learner that produces a lexicon of terms with the associated emotionalities based on their occurrences in pseudo-labeled data L. It takes as an input the list of terms (in our case, n-grams from the feature selection process), and for each term t computes its emotionality $\vec{w}(t) = (w_0(t), w_1(t), w_2(t), \ldots, w_{|E|}(t))$, where $w_i(t)$ is the weight of the term t for the emotion i.

For learning, we know the emotionality of each tweet $d \in L$, $\vec{p}(d) = (p_0(d), p_1(d), p_2(d), \ldots, p_{|E|}(d))$. In BWV, we first balance the distribution of emotions: we compute the rebalancing coefficient c_i for each emotion and multiply by it the corresponding emotion weight for each tweet. We then compute the weights of emotions for a term t as the normalized sum of rebalanced tweet emotion weights:

$$w_i(t) = \frac{\sum_{d:t \in d} c_i \cdot p_i(d)}{\sum_j \sum_{d:t \in d} c_j \cdot p_j(d)} \quad (2)$$

We define the coefficient for the i-th emotion as $c_i = -\log \frac{\sum_{d \in L} p_i(d)}{|L|}$. Using a logarithm in the formula allows penalizing the emotion categories appearing more often without overestimating the weights of under-represented emotion categories.

This algorithm is inspired by the simple Weighted Voting (WV) approach used by Sintsova et al. [2013]. The original WV differs from BWV in that it lacks the rebalancing coefficients c_i. As a result, the lexicon created is biased towards dominant emotions: the more often an emotion appears in the labeled data, the greater its weight will be in the emotionalities of the terms. The BWV approach involves reweighting process of the emotional assignments of tweets, which is similar to the resampling approaches designed to cope with class imbalances for classification problems [Japkowicz 2000].

The lexicon constructed via BWV learner from pseudo-labeled data is the basis for the *resultant emotion classifier*. It is applied to the tweets as follows. To compute the emotionality of a tweet $\vec{p}(d)$, we search for the lexicon terms within its text, sum the emotionalities of the lexicon entries found, and normalize the vector. If no lexicon terms are found, the *Neutral* label is returned. When lexicon terms are found, the output is an emotion multi-label obtained from the computed emotionality with the operator $\mathfrak{A}(\vec{p}(d), \alpha_0)$, where α_0 is the parameter of the algorithm.

3.4. Parameter Tuning and Automatic Evaluation

Our distant learning method involves multiple parameters, for example, the refinement parameter α_{ref} or the length n of n-gram features. To find its optimal parameters, we need to perform parameter tuning. For this, we suggest using automatically generated set of ground-truth tweets labeled based on the presence of emotional hashtags. The used emotional hashtags are explicit descriptive words for the chosen emotions, such as #*happy* for *Happiness*. In a study of users' moods, De Choudhury et al. [2012a] found that an emotional hashtag at the end of a text corresponded to the author's mood in 83% of tweets. We considered this evaluation to be the indicator of the good enough quality for using such emotional hashtags as ground-truth labels for automatic evaluation and parameter tuning. As our emotion recognition system also should be able to recognize tweets without emotions, we additionally include pseudo-neutral tweets in these constructed data. Overall, having such large ground-truth data allows for an automated way to set the parameters of our method and to validate its performance.

4. SETUP FOR METHOD APPLICATION

We present here a potential scenario of developing an emotion classifier for a new set of emotion categories to be detected within a specific topic of discussions. This section describes the data, emotion model, and initial labalers used, providing the details of how to apply our distant learning method in the real application.

4.1. Data for Application

We focus on the domain of fans' Twitter reactions to sports events. This domain was chosen because it contains various emotions with domain-specific emotional expressions, allowing us to see whether our method can adjust a general classifier to a target domain. Furthermore, it was studied in our previous work, providing access to the small within-domain emotion lexicon [Sintsova et al. 2013].

Our data consist of 33.2 million English Twitter posts collected over 2 weeks during the 2012 Olympic Games by querying Olympic-related keywords, such as "Olympic" or "London2012". We apply prior data filtering in order to select the tweets most useful for learning: we use only tweets containing at least three words (disregarding hashtags and usernames) to increase the probability of learning additional terms, and exclude retweets and tweets with duplicate text to avoid overfitting.

4.2. Emotion Model

Consistent with our previous work [Sintsova et al. 2013, 2014], we use the 20 emotion categories from the Geneva Emotion Wheel (GEW, v. 2.0), a model developed in

psychological research to systematically summarize self-reported emotional experience [Scherer 2005]. The categories are enumerated as labels on the horizontal axis in Figure 2. Each emotion category is represented by two common emotion names to emphasize its family nature (e.g., *Happiness*/*Joy*). We will use the first of them for a shorter reference.

The GEW has multiple advantages. Whereas common sets of basic emotions, such as from Ekman [1992] or Plutchik [2001], contain up to 8 categories, the GEW's 20 categories provide a more accurate approximation of the full range of emotions that humans are capable of experiencing. Such a fine-grained model allows us to discover more insightful details about emotional reactions. Compared to the OCC model [Ortony et al. 1988] containing 22 categories differentiated based on cognitive attribution of factors evoking emotion, we believe that the GEW emotions are more likely to be distinguished correctly based solely on lexical terms (e.g., it can be difficult to distinguish *Gratification* from *Satisfaction* without proper context modeling). Another alternative were the 24 primary emotions of the Hourglass of Emotions [Cambria et al. 2012], an advanced representation of Plutchik's emotion wheel distinguishing 4 affective dimensions and specifying 6 levels of emotion in each. However, that model lacks cognitive-based emotions such as *Pride* or *Pity*, thus precluding their analysis in sports events.

4.3. Emotion Labelers

Three initial emotion lexicons are taken as emotion labelers for our distant learning method. Two are topic-independent: one lexicon of explicit emotional terms (GALC) and one weighted lexicon learned from general Twitter data (PMI-Hash). We also take one domain-specific lexicon built using human computation for analyzing reactions to sporting events on Twitter (OlympLex).

4.3.1. GALC. GALC is a domain-independent emotion lexicon of the unigram stems explicitly expressing an emotion, for example, "happ∗" for *Happiness*. It was developed along with the GEW, for automatically classifying free-format survey responses into emotion categories [Scherer 2005]. GALC contains 279 stemmed terms for 36 emotion categories. From these, we use 212 stems associated with 20 GEW categories. To avoid dealing with pattern-based detection of lexicon terms, we replaced them with their instances detected in 15,000,000 random general tweets. From these, we excluded some frequent misallocations, such as "functional" mapped for "fun*". This process resulted in 1,027 terms associated with emotions. To compute a document's emotionality using this lexicon, we sum the number of terms found for each emotion (excluding negated terms) and normalize the obtained vector.

4.3.2. OlympLex. This domain-specific emotion lexicon was obtained by annotating tweets about sports events in crowdsourcing settings. It contains the emotion indicators selected from those tweets by the annotators, as well as related user-entered emotional expressions [Sintsova et al. 2013]. This emotion lexicon allocates a GEW-based emotionality for each of its 3,193 terms (from unigrams to 5-grams). Furthermore, we removed 94 frequent terms related to a description of the Olympics rather than emotions, such as "event". The average of the emotionalities of terms found in the tweet text (excluding negated terms) is the emotionality of the whole tweet.

4.3.3. PMI-Hash. We also generate a topic-independent Twitter-specific emotion lexicon using the PMI-based method [Mohammad 2012]. It computes emotion weights of terms using tweets with emotional hashtags. We defined a set of 167 emotional hashtags for all the 20 emotion categories of GEW based on the GALC lexicon [Scherer 2005] described earlier. The English tweets with those hashtags were collected via streaming API without any further restrictions. From them, we randomly selected

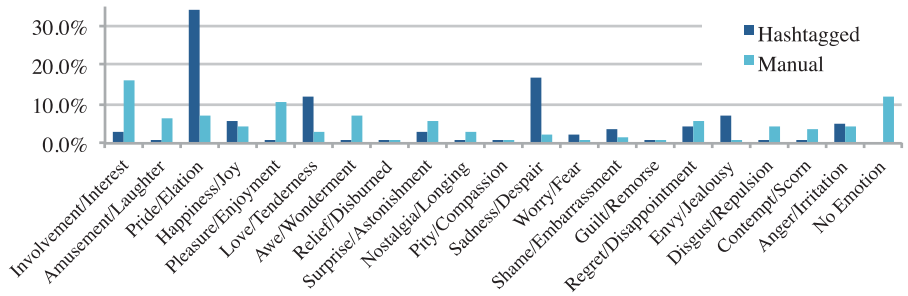

Fig. 2. Distribution of emotions found in the hashtagged dataset and manual annotations.

500,000 tweets where the specified hashtags appear at the end, excluding retweets, short tweets, tweets with several emotional hashtags, and duplicates. We applied the same preprocessing and n-gram extraction steps as in our method, and used unigrams and bigrams as terms for the lexicon. The weights of these terms are computed via the *PMI-Based* learner. It computes the strength of association $SoA(t, e_i)$ of term t to the emotion e_i as the difference in PMI of term t toward the presence and absence of emotion e_i. The formula (1) is used again while considering the presence of emotion e_i^+ as a positive class and an absence of emotion e_i^- as a negative class. The positive values are saved as term emotion weights, that is, $w_i(t) = \max(0, SoA(t, e_i))$. In total 85,530 terms are extracted. When applying this lexicon to the text, we sum the weights of found lexicon terms and normalize the resultant vector to obtain an emotionality of the text.

4.4. Neutral Labeler

The neutral labeler aims to find the tweets with a high probability of being neutral. To define the heuristics of such labeling, we assume that the presence of a URL indicates less emotional tweets, such as news or information sharing. We extracted such tweets and observed that to enforce tweet neutrality, we should avoid the presence of usernames and personal pronouns (which makes sharing more personal), emoticons, and other emotional cues. We exclude tweets that contain explicit emotional terms from the GALC lexicon, intensity shifters (exclamation marks, elongations, intensifier and downtoner words), and strong subjective terms (from MPQA Subjectivity Lexicon [Wilson et al. 2005]). The examples of such identified neutral tweets are "Sports Debates and Olympic Coverage <URL>" and "read more: history of the Olympic torch, flame, and relay <URL>."

5. EVALUATION METHODOLOGY

This section describes ground-truth data used for the evaluation of the obtained classifiers, as well as how we tune and evaluate the resultant emotion classifiers. Also, it introduces other classifiers used for comparison with BWV.

5.1. Ground-Truth Data

5.1.1. Large Automatically Labeled Data. Following the idea introduced in Section 3.4, we generate the pseudo-annotated tweets for evaluating the quality of built classifiers in an automatic way. We extracted from the full dataset of pre-filtered Olympic tweets (no short tweets, retweets, or duplicates) those that contain emotional hashtags at the end of a text (we used the same 167 hashtags that were used earlier while building PMI-Hash). We additionally excluded tweets featuring several emotional hashtags. This procedure resulted in 52,218 tweets labeled with emotional hashtags, that is, only

0.16% of the full dataset of 33.2 million tweets. The distribution of emotion categories is given in Figure 2. As these data are intended for testing the algorithms' outputs, "labeling" hashtags were removed from the texts.

The second half of the automatic ground-truth data consist of pseudo-neutral tweets, that is, tweets detected by the introduced neutral labeler. We randomly selected the same number of such tweets (52,218) for inclusion in the evaluation set. The URLs were removed from their texts. We decided to use the same number of pseudo-neutral tweets as hashtagged tweets because the real proportion of emotional to non-emotional tweets is unknown and may vary between datasets or dataset subsets.

We split these automatic data into a validation set S_V to tune the algorithm's meta-parameters and test set S_T to evaluate the resultant classifiers in 1:2 proportion, that is 34,802 tweets for S_V and 69,634 tweets for S_T. This process preserved emotion distribution, meaning tweets for each emotion category were split proportionally, including pseudo-neutral tweets.

5.1.2. Manually Annotated Data. We asked human annotators to annotate 600 Olympic tweets (again pre-filtered, without overlap with S_V). To ensure the presence of multiple emotions, we avoided using only random tweets. Instead, we selected three types of tweets for annotation: 200 random tweets, 200 tweets with emotional hashtags (10 per each emotion category, with removed emotional hashtags), and 200 pseudo-neutral tweets (with removed URLs). Every tweet was labeled by two annotators. They were asked to provide up to three emotion labels per tweet, with one marked as dominant. They could also choose to label *Other emotion* or *No emotion*. Additionally, we asked them to mark if a tweet's emotion is ambiguous or if the text is unclear. We excluded such tweets to have a dataset of higher quality, resulting in 492 tweets available for evaluation. The Fleiss Kappa [Fleiss 1971] of paired dominant labels is 0.31, showing a fair agreement. We also computed what proportion of the tweets have partial agreement: we counted that in 58.3% of tweets, the dominant label from one annotator is within the full set of labels from another annotator. We found that disagreement comes frequently while discerning whether the tweets is emotional or not (19.3% of tweets). We asked for the third annotation of such tweets and excluded an annotation in disagreement with other two regarding whether the tweet is emotional or not.

In order to prepare the ground-truth dataset for testing, we assign to a tweet an emotion multi-label that includes two chosen dominant emotion categories and all other agreed categories from two annotators. The average number of labels per tweet is 1.71, showing the multiplicity of emotional experience and the need to treat this problem as multi-label classification. The distribution of outputted labels is shown in Figure 2. We name this evaluation set S_M.

5.2. Performance Metrics

We record the performance of the corresponding algorithm instances using multiple evaluation metrics suitable for multi-label classification [Tsoumakas and Katakis 2007; Sokolova and Lapalme 2009]. We compute both macro- and micro-versions of precision, recall, and F1-score, as well as accuracy.

Let T_i be the set of tweets where the emotion e_i is present according to the ground truth, O_i be the set of tweets that a classifier outputs as belonging to emotion e_i, and $C_i = T_i \cap O_i$ be the set of tweets correctly classified as belonging to emotion e_i. Then, for emotion e_i, recall is $R_i = \frac{|C_i|}{|T_i|}$, precision is $P_i = \frac{|C_i|}{|O_i|}$, and F1-score is $F_i = \frac{2P_iR_i}{P_i+R_i}$.

To compute macro-recall (*macro-R*), macro-precision (*macro-P*), and macro-F1 score (*macro-F*1), we average those values between emotion categories. Thus, $macro\text{-}R = \frac{1}{|E|}\sum_{i=1}^{|E|} R_i$, $macro\text{-}P = \frac{1}{|E|}\sum_{i=1}^{|E|} P_i$, $macro\text{-}F1 = \frac{1}{|E|}\sum_{i=1}^{|E|} F_i$. It is noteworthy that the

Neutral category e_0 is excluded from this averaging, as it is not the focus of the emotion recognition system. We also exclude the *Contempt* category, which was underrepresented in the dataset. The benefit of using macro-scores is that they assign equal importance to each emotion category, regardless of their distribution.

We compute micro-recall (*micro-R*), micro-precision (*micro-P*), and micro-F1 score (*micro-F*1) using the formulas for recall, precision, and F1-score with the total number of true labels, outputted labels, and correctly detected labels for all emotion categories. That is, $micro\text{-}P = \frac{\sum_i |C_i|}{\sum_i |O_i|}$, $micro\text{-}R = \frac{\sum_i |C_i|}{\sum_i |T_i|}$, and $micro\text{-}F1 = \frac{2\, micro\text{-}P \cdot micro\text{-}R}{micro\text{-}P + micro\text{-}R}$. The labels for the *Neutral* category are again excluded. In contrast to macro-metrics, micro-metrics take into account the distribution of emotions in the dataset. Thus, they provide an estimation of how well an evaluated classifier can detect emotions while giving more weight to the most frequently appearing emotions.

We also evaluate the accuracy of classifiers. In the context of multi-label classification, the accuracy $A(d)$ for a tweet d is defined as the Jaccard measure between the set of its true labels $T(d)$ and the set of labels $O(d)$ that a classifier outputs for it, that is, $A(d) = \frac{|T(d) \cap O(d)|}{|T(d) \cup O(d)|}$. The overall accuracy A is the mean of $A(d)$ for all tweets in the dataset D: $A = \frac{1}{|D|} \sum_{d \in D} A(d)$. Accuracy evaluates how applicable the classifier is, in general, over the dataset, as it checks its performance at the per-document level, while also evaluating its ability to separate the neutral category from other emotions.

5.3. Comparison with Other Methods

Using the same distant learning framework, we compare the BWV classifier with the five other supervised classifiers used for emotion recognition and text classification. To apply them instead of BWV, we transform the format of the pseudo-labeled data from emotionalities into multi-labels: to each emotionality $\vec{p}$ we apply operator $\mathfrak{A}(\vec{p}, \alpha_{ref})$ with the parameter α_{ref} specified for the annotation refinement. We consider two ways to address such multi-label classification using standard machine-learning classifiers: Multi-Class (mcl) and One-vs.-Rest (1vR) transformations [Tsoumakas and Katakis 2007].

5.3.1. Multi-Class Transformation (mcl). This approach transforms the given multi-label classification problem into a multi-class problem: each document d with a multi-label $Y = \{e_{i_k}\} \subseteq E^0$ yields $|Y|$ documents in the new training set, one for each label $e_{i_k} \in Y$. We consider two classifiers: Multinomial Naïve Bayes (mcl-MNB) and Logistic Regression (mcl-LogReg), implemented using WEKA [Hall et al. 2009] and LibLINEAR software [Fan et al. 2008], respectively. Both of them return probabilistic output, which is treated as an emotionality of the text and is transformed back into the multi-label using the operator $\mathfrak{A}$ with the parameter α_0 again.

5.3.2. One-vs.-Rest Transformation (1vR). This approach transforms the given multi-label problem into $|E^0|$ independent binary classification tasks, one for each emotion category. A classifier for emotion e_i decides if it is present (class e_i^+) or not (e_i^-). We again evaluate Multinomial Naïve Bayes (1vR-MNB) and Logistic Regression (1vR-LogReg) classifiers in these settings (but for binary classification). As both classifiers support the probabilistic output, we specified that multi-label output for a text d should contain only those emotions e_i for which the probability of its presence is higher than some threshold r (a new parameter) (i.e., when $P(e_i^+|d) > r$). We also applied an additional per-category feature selection with this transformation. To select features for emotion e_i, we used the term's strength of association to that emotion $SoA(t, e_i)$ computed in the same way as for PMI-based learner (described in Section 4.3.3). The terms that have an

absolute score $|SoA(t, e_i)|$ lower than a threshold θ are filtered out. For simplification, θ is fixed to the same value for all emotion categories.

5.3.3. PMI-Based Learner. We additionally compare our method with the PMI-based learner used for generation of emotion lexicon from pseudo-labeled data by Mohammad [2012] and described in detail in Section 4.3.3, where it is used for generating PMI-Hash emotion lexicon. We only include here the threshold θ to filter out low values of $|SoA(t, e_i)|$, similar to per-category feature selection in 1vR transformation. The outputted emotionality of the tweets is transformed into multi-label output using the operator $\mathfrak{A}$ with the parameter α_0 again.

5.3.4. Random Baseline. We also adapt a random baseline (*Random*) to estimate the problem's difficulty: it decides independently whether or not each emotion is present, with probability defined by the emotion distribution in the test dataset. Performance scores are averaged over 1,000 runs.

5.4. Input Data, Parameter Tuning, and Testing

In our experiments, instead of applying emotion labelers to all the available tweets, we use only N_U random pre-filtered tweets (no retweets, duplicates, or short tweets) due to our limited computational resources. At the same time, as our Neutral Labeler applies more restrictive heuristics, we could apply it to all the available pre-filtered tweets, and use in the experiments the same amount N_U of pseudo-neutral tweets. Balancing the amount of pseudo-neutral and potentially emotional tweets in learning process allows us to give the same detection priority to both of these classes. All unlabeled data used in learning are disjoint from any considered ground-truth data.

To find the optimal parameters of each algorithm, we perform parameter tuning separately for each initial emotion labeler. The data for learning are built with $N_U = 100{,}000$, and validation set S_V is used for recording the performance. Among the obtained results, we find a set of parameters that yields the highest micro-F1 score on S_V. We chose to maximize the micro-F1 score because it was found to lead to a better balance between micro-precision and recall. The parameter space explored and optimal parameters chosen are described in the appendix. The learning process for building final classifiers to test uses larger data built with $N_U = 500{,}000$. The obtained classifiers are then evaluated on automatic test set S_T and manual test set S_M.

6. EVALUATION RESULTS

This section presents the results of the tuned distant learning algorithms on the test datasets. We compare the performance of the resultant classifiers with the baseline performances of the initial emotion labelers applied without distant learning or the neutral labeler. They are reported as *Initial*. Further, we report how significantly each algorithm's performance metrics differ from those of the corresponding initial labeler, as estimated by randomization tests [Yeh 2000]. One asterisk * indicates a p-value ≤ 0.05; two asterisks ** indicate a p-value ≤ 0.01.

6.1. Improvement over the Initial Emotion Labelers

Table I presents the results on the test dataset S_T. They show that our proposed BWV method substantially improves the quality of initial emotion labelers on all of the main performance metrics: macro-F1, accuracy, and micro-F1 score. The only exception is a lower macro-F1 score when starting from OlympLex, but this result is insignificant (p-value $= 0.065$). The largest improvements are observed for micro-F1 scores: 41% when started from PMI-Hash, 53% from OlympLex, and 236% from GALC. The highest micro-F1 score is 40.6% with PMI-Hash as the input emotion labeler. The minimum relative increase in accuracy is 10.6% (with GALC). These findings confirm that BWV

Table I. Evaluating Distant Supervision Algorithms on Automatic Test Data S_T

Emotion Labeler	Algorithm	macro				A	micro				rank
		P	R	F1			P	R	F1		
-	Random	2.5	1.3	1.7		41.6	8.7	4.4	5.8		
GALC	Initial	20.6	3.6	4.8		52.2	23.6	5.1	8.4		
	mcl-MNB	**21.4**	12.2**	**10.3****	↑	**62.0****	↑	**30.6****	28.1**	**29.3****	↑	1
	mcl-LogReg	7.5**	**23.9****	8.9**	↑	43.1**	↓	9.6**	30.4**	14.6**	↑	6
	1vR-MNB	11.8**	17.1**	9.7**	↑	57.0**	↑	16.9**	**34.6****	22.7**	↑	4
	1vR-LogReg	12.1**	8.8**	8.1**	↑	54.4**	↑	22.0**	20.9**	21.5**	↑	5
	PMI-based	12.7**	10.2**	9.3**	↑	53.1**	↑	28.0**	26.4**	27.2**	↑	3
	BWV	16.8**	11.5**	9.8**	↑	57.8**	↑	27.2*	29.1**	28.2**	↑	2
Olymp-Lex	Initial	11.4	9.7	7.1		47.4	19.3	19.3	19.3		
	mcl-MNB	**19.7****	11.2**	6.8	↓	58.5**	↑	26.3**	27.0**	26.7**	↑	3
	mcl-LogReg	9.1**	12.4**	7.6**	↑	42.9**	↓	16.1**	21.6**	18.4**	↓	6
	1vR-MNB	19.4**	12.3**	7.3	↑	58.9**	↑	23.3**	28.3**	25.6**	↑	4
	1vR-LogReg	11.1*	**16.5****	**9.8****	↑	51.3**	↑	17.1**	27.9**	21.2**	↑	5
	PMI-based	15.8**	9.6	7.3	↑	58.8**	↑	28.3**	26.0**	27.1**	↑	2
	BWV	17.8**	9.4	6.7	↓	**59.4****	↑	**29.9****	**29.2****	**29.5****	↑	1
PMI-Hash	Initial	12.1	17.0	11.5		23.7	21.8	42.0	28.7		
	mcl-MNB	22.8**	15.9**	13.1**	↑	64.4**	↑	37.6**	43.0**	40.1**	↑	3
	mcl-LogReg	14.4**	18.7**	14.8**	↑	52.7**	↑	30.9**	41.8	35.5**	↑	6
	1vR-MNB	19.9**	16.7	14.2**	↑	**64.6****	↑	37.5**	43.3**	40.2**	↑	2
	1vR-LogReg	17.6**	**18.9****	**16.2****	↑	60.6**	↑	35.4**	42.2	38.5**	↑	5
	PMI-based	22.3**	15.6**	14.4**	↑	63.8**	↑	**38.5****	41.3**	39.9**	↑	4
	BWV	**29.3****	15.5**	13.1**	↑	64.1**	↑	37.3**	**44.4****	**40.6****	↑	1

All performance scores are percentages. The results of learned classifiers are compared with those of the corresponding initial labelers. One asterisk * indicates a p-value ≤ 0.05; two asterisks ** indicate a p-value ≤ 0.01.

can build emotion classifiers that are far more accurate than the existent emotion labelers. The experiments also show that the other algorithms applied within the same distant learning framework can improve the performance of initial classifiers too.

To our surprise, we observed greater macro- and micro-precision from the classifiers obtained through distant learning than from the initial classifiers. The best micro-precision of BWV is 37.3% starting from PMI-Hash. It is 71% better than that of the initial PMI-Hash lexicon. Based on our previous experiments [Sintsova et al. 2014], we expected that a distant learning approach would only improve the classifiers' recall by finding more emotional expressions. However, this increase in precision indicates that, in many cases, the distant learning process corrects the terms' emotion distributions.

6.2. Comparison of BWV and Other Supervised Classifiers

To further compare the distant learning algorithms, we rank their performance using the micro-F1 score (as it produces a more stable ranking for different emotion labelers).

Mcl-LogReg performs worst, both in terms of micro-F1 and accuracy. This is due to its lower precision, possibly because it finds more tweets to be emotional ($\geq$62% for all input emotion labelers) than other classifiers ($\leq$54%), thus making more mistakes on neutral tweets.

1vR-LogReg is the next worst, with the moderate micro-F1 scores. Although 1vR-LogReg achieves the highest macro-F1 scores for OlympLex and PMI-Hash, these are accompanied by relatively low macro-precision (in comparison to BWV), which is undesirable for the real-world applications.

Table II. Improvements Due to the Specific Charasteristics of Dystemo

	Parameters		macro				micro		
Algorithm	Blc	Neut	P	R	F1	A	P	R	F1
Initial	-	-	12.1	**17.0**	11.5	23.7	21.8	42.0	28.7
BWV	Log	Incl	**29.3****	15.5**	**13.1**** ↑	**64.1**** ↑	37.3**	**44.4****	**40.6**** ↑
Alternative parameters									
WV	No		25.9 **	11.2**	12.1** ↑	63.4** ↑	**45.3****	31.5**	37.1** ↑
BWV		No	16.0**	14.1**	11.1* ↓	34.0** ↑	25.9**	40.2**	31.5** ↑

The results are on automatic test set S_T with PMI-Hash as initial emotion labeler. All performance scores are percentages. Results of learned classifiers are compared with those of the corresponding initial labeler.

The third and forth ranks in the aggregated performance are shared between 1vR-MNB and PMI-based methods. 1vR-MNB performs best with PMI-Hash, achieving the top accuracy and high micro-F1 score; whereas PMI-based systematically increases accuracy, micro-precision and micro-F1 scores for all three emotion labelers.

The two classifiers with the highest ranks are mcl-MNB and BWV. When starting from GALC, mlc-MNB's performance is superior to BWV for all metrics except micro-recall. However, BWV produces the highest micro-F1 scores starting from the OlympLex and PMI-Hash lexicons. Moreover, its 40.6% micro-F1 score when starting from MNB-Hash is the highest score achieved in all our experiments, indicating that BWV was the most appropriate for real-world applications of emotion recognition in tweets.

6.3. Effects of Choosing Initial Emotion Labeler

The three evaluated initial emotion labelers differ not only in their basic performance but also in their results in conjunction with the distant learning method. Due to its explicit nature, the GALC lexicon has relatively high macro- and micro-precision, but low recall. Distant learning can improve its performance by increasing recall—it discovers new emotional terms that co-appear with the given terms in the unlabeled data. OlympLex's precision and recall are close to each other due to its higher coverage of emotion terms used in the sports domain. This lexicon's size is moderate, but our method can still discover new terms indicative of emotion and increase both micro- and macro-precision, probably because of a better adjustment of the distribution in emotion categories and better separation of the most frequent categories. Finally, PMI-Hash shows the highest macro- and micro-F1 scores of all the initial emotion labelers, yet its accuracy is the lowest and its recall is almost twice as large as precision. The PMI-Hash has this behavior because it was trained on data without neutral tweets, and thus it classifies most tweets (96%) as belonging to an emotion category and has low accuracy for neutral tweets. The distant learning approach successfully helps overcome this problem and increases PMI-Hash's precision up to the level of its recall.

Overall, this evaluation indicates that the described distant learning method is able to adjust all three initial emotion lexicons to an application domain. This is validated by a statistically significant increase in accuracy and micro-F1 score.

6.4. Variations in Dystemo Configuration

Rebalancing Process. The suggested BWV learning method originates from Weighted Voting (WV), which does not introduce rebalancing coefficients c_i (described in Section 3.3). Table II shows the benefits of having the rebalancing process. It compares BWV with WV on test set S_T while using PMI-hash as emotion labeler (the parameters of WV were tuned separately).

Table III. Evaluating Distant Supervision Algorithms on Manual Test Data S_M

Emot. Labeler	Algorithm	A	micro-P	micro-R	micro-F1	Coverage	#Labels
GALC	Initial	25.6	**50.0**	5.4	9.8	14.2	1.14
	mcl-MNB	**30.0**** ↑	30.5** ↓	**12.2**** ↑	**17.4**** ↑	51.2	1.17
	BWV	28.9** ↑	27.8** ↓	11.1** ↑	15.9** ↑	51.8	1.16
OlympLex	Initial	32.7	**42.5**	16.4	**23.6**	54.5	1.06
	mcl-MNB	**34.9**** ↑	39.7 ↓	**16.6** ↑	23.5 ↓	63.0	1.00
	BWV	33.6** ↑	37.7 ↓	13.9* ↓	20.4* ↓	55.5	1.00
PMI-Hash	Initial	12.2	19.7	**12.9**	15.6	98.0	1.00
	mcl-MNB	28.4** ↑	**27.5**** ↑	**12.9** –	**17.5** ↑	65.9	1.06
	BWV	**28.5**** ↑	25.9** ↑	11.8 ↓	16.2 ↑	60.4	1.13

All scores are percentages, except for the average number of emotion labels #Labels. Results of learned classifiers are compared with those of the corresponding initial labelers.

We observe that without rebalancing, WV is inferior to BWV for micro- and macro-F1 scores. Although WV showes the highest micro-precision, its recall is significantly lower than the initial labelers. With OlympLex and GALC as start points, WV's macro-F1 scores are even lower than those of the initial emotion labelers. This means that WV without rebalancing is unsuitable for distant learning, at least not within our method.

Using Neutral Tweets during Learning. One part of pseudo-labeled data for learning comprises pseudo-neutral tweets. We investigate if adding them is helpful by learning additionally the BWV classifier without including the pseudo-neutral tweets in the learning process (with the parameters retuned accordingly). Its results on the test set S_T are indicated with parameter *Neut=No* in Table II. It is noteworthy that, as S_T includes neutral tweets, the classifier's ability to recognize them is evaluated too.

We find that without neutral tweets BWV performs worse than with them in all metrics. This is because without exposure to neutral tweets during learning, resultant classifiers tend to classify most test tweets as emotional (up to 86%), even though it adapted higher feature selection threshold τ. This results in many errors on neutral tweets. Similar behavior is observed when using the other two initial labelers (OlympLex and GALC), but results are aggravated by a significant decrease in accuracy.

6.5. Validation of Distant Learning on Manually Annotated Data

Testing algorithms on large ground-truth data S_T allowed us to automatically find the best parameters of the algorithms and cover more feature terms in evaluation. However, testing on manual data is essential to understand how the quality of classifier will be perceived in human's eye. Thus, we confirm the positive effects of distant learning on small manually annotated data, S_M, described in Section 5.1.2. Table III presents the results of this test with two distant learning methods, BWV and mcl-MNB. We compare these two methods because they ranked high on automatic test data S_T. Notice that we do not report macro-scores because for many categories there are not enough tweets to obtain conclusive per-category metrics. However, we additionally report coverage of the methods which estimates how many tweets were detected as emotional, and the average number of emotion labels found in tweets classified as emotional.

When initial emotion labelers are GALC and PMI-Hash, the effects of distant learning algorithms remain similar to those discovered with automatic test data S_T: applying distant learning increases the accuracy and micro-F1 scores of initial labelers, due to recall increase for GALC (along with coverage increase) and precision increase for PMI-Hash. However, the improvements are smaller. This can be attributed to the fact that our manual annotation is less skewed toward dominant categories and requires from classifiers to perform better across more categories. Also, while comparing the

performance of mcl-MNB and BWV algorithms starting from GALC and PMI-Hash, we can observe that mcl-MNB slightly outperforms BWV. However, we did not find a significant difference in their micro-F1 scores.

With OlympLex as a starting labeler, we obtain different effects. While both distant learning methods still increase the accuracy over the initial OlympLex labeler, neither mlc-MNB nor BWV improve the micro-F1 score despite previously observed significant increase. However, already on automatic data we have observed an insignificant decrease in their macro-F1 scores. This can signify the need to optimize for both macro- and micro-F1 scores in the parameter tuning process. Moreover, this evaluation shows that OlympLex, built using manual annotations of tweets, performs best on manually annotated data. This reveals the difficulty to improve emotion lexicons of better quality via distant learning and the need for more advanced methods in such cases.

We also observe all resultant classifiers have lower micro-recall scores on manual test set S_M compared to those scores on automatic set S_T. This can be due to the higher average number of emotion labels per tweet in the manual ground-truth (1.71 in S_M versus 1.0 in S_T). This means that, to achieve better recall scores, resultant classifiers have to find correctly more emotion labels per tweet. However, all the final classifiers return only up to 1.17 emotion labels per tweet. With OlympLex as an initial labeler, both mcl-MNB and BWV learn to return exactly one label per tweet. This leaves room for potentially better optimization of the classifiers' output parameter α_0.

Overall, our method, Dystemo applied with BWV as a learning algorithm is shown to be effective in extending initial emotion lexicons of small coverage to find more emotional tweets (coverage is 264% more and recall is 105% higher for GALC lexicon). Additionally, it can improve coarse emotion lexicons to perform more accurately (accuracy is 133% higher for PMI-Hash lexicon).

7. DISCUSSION AND FUTURE WORK

The present work showed that applying distant learning with emotion lexicons as initial labelers is a viable approach for building application-specific emotion classifiers. Experiments show that the resultant classifiers are able to achieve micro-F1 scores between 15.9% and 40.6% while recognizing 20 emotions. Previous work reported similar scores when fewer emotion categories were used, for example, Mohammad [2012] achieved a micro-F1 score of 49.9% for six basic emotion categories in cross-validation on hashtagged tweets and 43.7% on news headlines. Our classifiers deal with more emotion categories, and thus the performance baseline for guessing randomly is much lower (5.8% for 20 emotions versus 16.7% for 6). This means the F1-scores of our method are more difficult to achieve given the challenging nature of the problem.

The suggested method was proven to be beneficial while using as an input three different kinds of initial lexicons. The performance of the resultant classifiers seems to vary depending on the amount of pseudo-labeled emotional data discovered by initial emotion lexicons. It would be interesting for future studies to examine what quantity of unlabeled data is required for the successful distant learning process. Moreover, we observe that the initial lexicons can have different best-detected categories. This can motivate future research in aggregating the classifiers obtained via distant learning from different initial lexicons in order to build the classifier having a better quality.

We confirmed the contribution of the main components specific to our Dystemo method. The rebalancing, introduced in Balanced Weighted Voting (BWV) learner, leads to the relative increase of micro-F1 score by 9.2%. Techniques for balancing training data have never been tested for emotion recognition before. Applying other rebalancing techniques [Batista et al. 2004] and testing how rebalancing processes help other learning algorithms for emotion recognition could be interesting avenues for future research as well. Another distinguishing property of our method is inclusion of

novel heuristics to identify neutral tweets for learning. Our experiments show that this is essential to avoid constructing classifiers that find emotions in almost every tweet: when starting from PMI-Hash, accuracy grows from 34% to 64.1%. While a distant supervision over pseudo-neutral tweets was already proposed in the context of polarity classification [Pak and Paroubek 2010; Kouloumpis et al. 2011], for the problem of emotion classification, a *Neutral* category was only studied when training data were labeled manually, for example by Neviarouskaya et al. [2011].

By comparing the suggested BWV learning method with other more advanced supervised classifiers, we show that even a simple lexicon-based classifier can achieve competitive performance. Yet, the additional advantage of BWV is that it produces an emotion lexicon, where each term (*n*-gram) is associated with an emotion distribution (called emotionality). This property opens a large perspective for potential future applications and improvements, such as extracting lexicon-based features for machine-learning classifiers [Mohammad 2012; Wang et al. 2012].

While investigating the viability of distant learning starting from emotion lexicons, we used relatively simple features for classification, namely *n*-grams appearance, and simple feature aggregation techniques, that is averaging the distributions of appeared *n*-grams. We further review what mistakes our classifiers repeatedly made due to these simplifications. Many seem to appear in the tweets where emotional sense is captured within spans of texts longer than *n*-grams. Examples are "Why is <x> always on when I want to watch <y>?" and "<x>'s hopes for medal in <y> dashed." Our method would potentially benefit from incorporating more developed techniques of representing emotional meaning in text, such as parsing semantic concepts [Poria et al. 2014] or extracting main emotional parts [Shaheen et al. 2014]. Similarly, modeling semantic compositionality could help to better aggregate detected lexicon features into tweet-level emotions. An example solution can involve treating emotions in composite phrases using hand-coded rules [Neviarouskaya et al. 2011] or deep neural network representations [Socher et al. 2013; Severyn and Moschitti 2015]. Another source of mistakes is the lack of proper modeling of contextual modifiers that can change the emotional meaning of terms. In the future, we plan to include better treatment of such linguistic modifiers as negations (e.g., "lose interest") and downtoners (e.g., "least favorite") while applying both final and initial classifiers [Carrillo-de Albornoz and Plaza 2013]. Finally, we observe tweets with the mixture of positive and negative emotions (e.g., "unlucky <x>, we are still proud of you"). Learning from them is likely to cause erroneous associations of terms to positive and negative emotions simultaneously. Future work should address how to limit the scope of corresponding emotion descriptions in the text, for example, based on annotating parts of the texts with off-the-shelf polarity classifiers, such as SentiStrength [Thelwall et al. 2012].

The distant learning method developed and analyzed in this article is potentially valuable to many domains of textual emotion analysis lacking easily accessible labels. Further studies are required to determine whether these results can be generalized to those domains (e.g., reactions to other public events such as awards or elections, product reviews, or posts in support forums) with their corresponding sets of emotions.

8. CONCLUSION

This manuscript presents an in-depth study of Dystemo—a distant learning method for multi-category emotion recognition in tweets. Instead of defining heuristics for detecting tweets with specific emotions based on hashtags or emoticons, we argue for the use of existing or easy-to-produce emotion lexicons as a starting point. We describe a method that can either extend an initial lexicon to cover more emotional terms and expressions, or refine it to detect emotional tweets more correctly. Both improvements make the novel classifiers more suitable for the chosen application.

Using sports tweets as a dataset, we have shown a detailed validation process involving three different initial emotion lexicons for the classification of twenty emotion categories. The proposed distant learning method, applied with a novel supervised learner—Balanced Weighted Voting—improves the micro F1-score in all three cases, with relative increases between 41% and 236%. Subsequent experiments suggest that rebalancing initially labeled data is an essential step in our method's success. Among other contributions, we introduce heuristics to automatically find neutral tweets and show the importance of including them in the learning process.

To the best of our knowledge, Dystemo is the first framework to produce domain-specific emotion classifiers without using costly manual labeling or special content cues such as hashtags. Because of these properties, Dystemo is more general than other existing methods. Researchers and practitioners can easily adapt Dystemo's emotion model to the requirements of their specific domain, and start building an optimal classifier using the procedure described in this paper.

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for providing valuable suggestions in the process of improving this paper. We are also grateful to Marina Boia for collecting the Twitter data in question and to Dr. Claudiu Musat for helpful discussions. Additionally, we thank all the participants who annotated the tweets for their time and effort.

REFERENCES

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. ACL, 579–586.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *Text, Speech and Dialogue*. Springer, 196–205.

Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter* 6, 1 (2004), 20–29.

George Bojadziev and Maria Bojadziev. 1995. *Fuzzy Sets, Fuzzy Logic, Applications*. Vol. 5. World Scientific.

Felipe Bravo-Marquez, Marcelo Mendoza, and Barbara Poblete. 2013. Combining strengths, emotions and polarities for boosting Twitter sentiment analysis. In *Proceedings of the 2nd International Workshop on Issues of Sentiment Discovery and Opinion Mining*. ACM, 2.

Erik Cambria, Andrew Livingstone, and Amir Hussain. 2012. The hourglass of emotions. In *Cognitive Behavioural Systems*. Springer, 144–157.

Jorge Carrillo-de Albornoz and Laura Plaza. 2013. An emotion-based model of negation, intensifiers, and modality for polarity and intensity classification. *Journal of the American Society for Information Science and Technology* 64, 8 (2013), 1618–1633.

Taner Danisman and Adil Alpkocak. 2008. Feeler: Emotion classification of text using vector space model. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, Vol. 1. 53.

Munmun De Choudhury, Scott Counts, and Michael Gamon. 2012a. Not all moods are created equal! Exploring human emotional states in social media. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Munmun De Choudhury, Michael Gamon, and Scott Counts. 2012b. Happy, nervous or surprised? Classification of human affective states in social media. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Pieter M. A. Desmet. 2012. Faces of product pleasure: 25 positive emotions in human-product interactions. *International Journal of Design* 6, 2, (2012).

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion* 6, 3–4 (1992), 169–200.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9 (2008), 1871–1874.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 5 (1971), 378.

Alec Go, Richa Bhayani, and Lei Huang. 2009. *Twitter Sentiment Classification using Distant Supervision*. Technical Report. Standord.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter* 11, 1 (2009), 10–18.

Nathalie Japkowicz. 2000. The class imbalance problem: Significance and strategies. In *Proceedings of the International Conference on Artificial Intelligence (ICAI)*, Vol. 1. Citeseer, 111–117.

Renato Kempter, Valentina Sintsova, Claudiu Musat, and Pearl Pu. 2014. EmotionWatch: Visualizing fine-grained emotions in event-related tweets. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Sunghwan MacKim, Alessandro Valitutti, and Rafael A. Calvo. 2010. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings NAACL-HLT 2010 Workshop on Computing Approaches to Analysis and Generation of Emotion in Text*. ACL, 62–70.

Efthymios Kouloumpis, Theresa Wilson, and Johanna D. Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)* 11 (2011), 538–541.

Uros Krcadinac, Philippe Pasquier, Jelena Jovanovic, and Vladan Devedzic. 2013. Synesketch: An open source library for sentence-based emotion recognition. *IEEE Transactions on Affective Computing* 4, 3 (July 2013), 312–325.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* 5, 1 (2012), 1–167.

Lionel Martin and Pearl Pu. 2014. Prediction of helpful reviews using emotions extraction. In *Proceedings of 28th AAAI Conference on Artificial Intelligence (AAAI)*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Vol. 2. ACL, 1003–1011.

Gilad Mishne and Maarten de Rijke. 2006. MoodViews: Tools for blog mood analysis. In *AAAI Spring Symp.: Computing Approaches to Analyzing Weblogs*. 153–154.

Saif M. Mohammad. 2012. #Emotional tweets. In *Proceedings 1st Joint Conference on Lexical and Computing Semantics (*SEM)*. ACL, 246–255.

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence* 31, 2 (2015), 301–326.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computing Intelligence* 29, 3 (2013), 436–465.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2011. Affect analysis model: Novel rule-based approach to affect sensing from text. *Natural Language Engineering* 17, 1 (2011), 95–135.

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning* 39, 2–3 (2000), 103–134.

Andrew Ortony, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press, New York, NY.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2, 1–2 (2008), 1–135.

Jessica Perrie, Aminul Islam, Evangelos Milios, and Vlado Keselj. 2013. Using google n-grams to expand word-emotion association lexicon. In *Computing Linguistics and Intelligent Text Process.* Springer, 137–148.

Rosalind W. Picard and Jonathan Klein. 2002. Computers that recognise and respond to user emotion: Theoretical and practical implications. *Interacting with Computers* 14, 2 (2002), 141–169.

Robert Plutchik. 2001. The nature of emotions. *American Scientist* 89, 4 (2001), 344–350.

Soujanya Poria, Basant Agarwal, Alexander Gelbukh, Amir Hussain, and Newton Howard. 2014. Dependency-based semantic parsing for concept-level text analysis. In *Computational Linguistics and Intelligent Text Processing*. Springer, 113–127.

Soujanya Poria, Alexander Gelbukh, Erik Cambria, Dipankar Das, and Sivaji Bandyopadhyay. 2012. Enriching SenticNet polarity scores through semi-supervised fuzzy clustering. In *Proceedings of SENTIRE, 12th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 709–716.

Soujanya Poria, Alexander Gelbukh, Erik Cambria, Amir Hussain, and Guang-Bin Huang. 2014. EmoSenticSpace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems* 69 (Special Issue on Big Social Data Analysis) (2014), 108–123.

Soujanya Poria, Alexander Gelbukh, Amir Hussain, Dipankar Das, and Sivaji Bandyopadhyay. 2013. Enhanced SenticNet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems* (2013), 31–38.

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. ACL, 482–491.

Likun Qiu, Weishi Zhang, Changjian Hu, and Kai Zhao. 2009. SELC: A self-supervised model for sentiment classification. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*. ACM, 929–936.

Daniele Quercia, Licia Capra, and Jon Crowcroft. 2012. The social world of twitter: Topics, geography, and emotions. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. EmpaTweet: Annotating and detecting emotions on twitter. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*. 3806–3813.

Klaus R. Scherer. 2005. What are emotions? And how can they be measured? *Social Science Information* 44, 4 (2005), 695–729.

Aliaksei Severyn and Alessandro Moschitti. 2015. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 959–962.

Shadi Shaheen, Wassim El-Hajj, Hazem Hajj, and Shady Elbassuoni. 2014. Emotion recognition from text based on automatically generated rules. In *Proceedings of the Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction (SENTIRE)*.

Valentina Sintsova, Claudiu Musat, and Pearl Pu. 2013. Fine-grained emotion recognition in olympic tweets based on human computation. In *Proceedings of the NAACL-HLT WASSA*. ACL, 12–20.

Valentina Sintsova, Claudiu Musat, and Pearl Pu. 2014. Semi-supervised method for multi-category emotion recognition in tweets. In *Proceedings of the Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction (SENTIRE)*.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Vol. 1631. Citeseer, 1642.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45, 4 (2009), 427–437.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposuim on Applied Computing*. ACM, 1556–1560.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet affect: An affective extension of WordNet. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Vol. 4. 1083–1086.

Jared Suttles and Nancy Ide. 2013. Distant supervision for emotion classification with discrete binary values. In *Computing Linguistics and Intelligent Text Process.* Springer, 121–136.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computing Liguistics* 37, 2 (2011), 267–307.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology* 63, 1 (2012), 163–173.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)* 3, 3 (2007), 1–13.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)* 21, 4 (2003), 315–346.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2012. Harnessing twitter "big data" for automatic emotion identification. In *Proceedings of the International Conference on Social Computing (SocialCom)*. IEEE, 587–592.

Michael Wiegand and Dietrich Klakow. 2009. Predictive features in semi-supervised learning for polarity classification and the role of adjectives. In *Proceedings of NoDaLiDa*. 198–205.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 347–354.

Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007. Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. ACL, 133–136.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th Conference on Computational Linguistics*, Vol. 2. ACL, 947–953.

Xiaojin Zhu. 2005. Semi-supervised learning literature survey. Computer Sciences Technical Report 1530. University of Wisconsin–Madison.

Received August 2015; revised January 2016; accepted April 2016