2014 IEEE International Conference on Data Mining Workshop

Semi-Supervised Method for Multi-Category Emotion Recognition in Tweets

Valentina Sintsova, Claudiu Musat, Pearl Pu
School of Computer and Communication Sciences
Swiss Federal Institute of Technology (EPFL)
Lausanne, Switzerland
{valentina.sintsova, claudiu-cristian.musat, pearl.pu}@epfl.ch

***Abstract*—Each tweet is limited to 140 characters. This constraint surprisingly makes Twitter a more spontaneous platform to express our emotions. Detecting emotions and correctly classifying them automatically is an increasingly important task if we want to understand how large groups of people feel about an event or relevant topic. However, constructing supervised classifiers can be a daunting task because of the high manual annotation costs. We propose constructing emotion classifiers with a minimal amount of initial knowledge (e.g. a general-purpose emotion lexicon) and using a semi-supervised learning method to extend it to correctly detect more emotional tweets within a specific domain. Additionally, we show that our algorithm, Balanced Weighted Voting (or BWV) is able to overcome the imbalanced distribution of emotions in the initial labeled data. Our validation experiments show that BWV improves the performance of three initial classifiers, at least in the specific domain of sports. Furthermore, its comparison with other two learning strategies reveals its superiority in terms of macro F1-score, as well as more stable performance among different emotion categories.**

***Keywords*-Emotion Recognition, Twitter, Semi-Supervised Learning, Text Mining, Natural Language Processing**

I. Introduction

The abundance of emotions we feel is reflected in our language. Their automatic recognition can help us build more sophisticated social and personal applications, including those that study social relations [1], enhance human-computer interaction [2], and summarize public reactions [3]. In this work, we model "emotions" as belonging to a finite number of categories and formulate a problem of a multi-category emotion recognition in text [4], [5]. For a given text sample, we aim to detect which emotion(s) from the given set it expresses.

While this problem has received substantial attention from the research community, constructing a universally applicable classifier remains an unsolved and complex task. One difficulty lies in the context-dependency of emotions: their linguistic expressions and causes vary with different domains (e.g. studying vs. sports), types of text (private messages, online statuses, or posts), and even the author's style. Furthermore, the set of suitable emotion categories varies as well, depending on the chosen domain and application. For example, *Love* would be a frequent emotion in interpersonal communications, while relatively rare in technical forums. Due to these intrinsic differences, a domain-independent classifier can have only limited accuracy on any given domain. Thus, we believe that the ability to build or adapt an emotion classifier for a specific domain will greatly enhance classification performance.

This paper proposes a novel semi-supervised method for this purpose. It leverages unlabeled data within a specific domain to extend the initial limited classifiers in order to capture the specificity of the target domain. We consider as the initial classifiers those that are either based on general-purpose emotion lexicons or trained on limited data within the domain. They are likely to have limited coverage of present emotions, and thus require further adaptation. Nevertheless, they contain prior knowledge, which the semi-supervised algorithm can potentially extend (under the assumption that domain-specific emotional expressions will appear within the text labeled by the initial classifier).

Our method is based on the idea of distant learning [5], [6]. Its overview is shown in Fig. 1. First, the *initial classifier* is applied to the unlabeled data to obtain the *pseudo-labeled data*. Second, this annotation is refined by choosing for each text the most prominent labels among those suggested. Third, the features, *n*-grams from the text, are extracted and filtered out. Then, the re-weighting techniques are applied to rebalance the annotation. Finally, the supervised learner

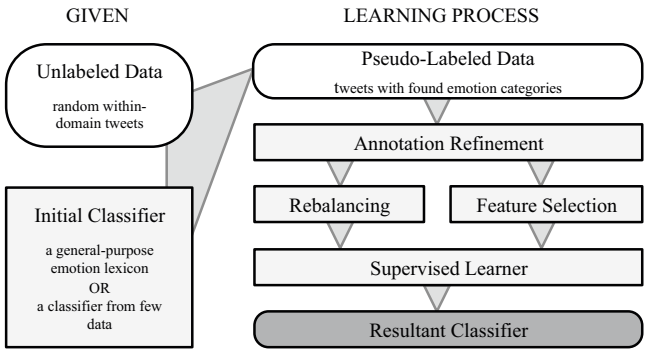

Figure 1. The framework for our semi-supervised learning method

978-1-4799-4274-9/14 $31.00 © 2014 IEEE
DOI 10.1109/ICDMW.2014.146

trains the *resultant classifier* using the rebalanced *pseudo-labeled data* and the *selected features*. This classifier can later be applied to other data within the domain.

A distant learning approach was shown to be effective for polarity [7], [8] and emotion recognition [5], [6]. However, it was tested only on domain-independent tweets. It was not clear *a priori* if such an approach would be beneficial or harmful when applied exclusively within one domain.

We test our semi-supervised method on sports events reactions on Twitter. We focus on the set of 20 emotion categories from the Geneva Emotion Wheel (GEW) [9], [10]. Its fine-granularity allows making more insightful discoveries about emotions. For this emotion set several initial classifiers are available; we consider three of them.

This paper makes the following contributions:

- We develop a semi-supervised method that uses unlabeled data to extend limited initial emotion classifiers for a specific domain.
- We design a learning algorithm, Balanced Weighted Voting, that addresses an imbalance of emotion labels in annotated data—a problem which is poorly studied for emotion classification.
- We experimentally show that this method results in substantially better classifier quality than that of the three initial ones: the relative increase of macro F1-score is between 24% and 105%.
- We compare this learning algorithm with two other commonly used supervised classifiers: Naïve Bayes and a PMI-based one. With Balanced Weighted Voting, we achieve not only better general performance (macro F1-score is higher in average on relative 33%), but also consistently better performance throughout many categories.

The paper is organized as follows: The following section reviews related work. Then, we present the formal description of our semi-supervised method and its steps. Next, we describe the experimental setup and present our results. The last sections conclude the paper and discuss future work.

II. Related Work

Multi-category emotion recognition in text is an increasingly popular sub-topic in sentiment analysis [11], [12] with many methods adapted from text polarity classification. The use of lexicons is one such adaptation. Just as sentiment lexicons store terms' polarities [13]–[15], emotion (or affective) lexicons provide term-emotion associations. Some list only terms directly expressing an emotion, such as "happy" for *Joy* (the GALC lexicon is an example [9]). Others contain additional terms linked to some emotional experience, such as "comfort" for *Joy* in WordNetAffect [16], "entertain" in NRC [17] or "visit friend" in EmoSenticNet [18]. Counting the number of lexicon terms appeared in the text for each emotion can be used as helpful features for various text classification problems [19], [20], including emotion classification itself [5], [21]. Rule-based algorithms go beyond simple keyword-spotting by taking into account sentence structure and syntactic features, such as presence of negations, intensifiers, or conjunctions [22], [23]. While such lexicon-based methods are unsupervised and can be applied to any domain, they do not cover the full variety of emotional expressions used in the language.

Researchers adapted semi-supervised techniques to extend given lexicons (or term seeds). They define several metrics of term similarity, and then use them to cluster or classify new terms into emotion categories based on their similarity to those given. The original WordNetAffect [16] and one part of Synesketch lexicon [23] were built in this way, with similarity metrics defined using semantic relations, such as synonymy, from WordNet [24]. In construction of EmoSenticNet lexicon [18], [20], [25], additional term similarities were derived from term co-occurrences on the database of emotional experiences using Pointwise-Mutual Information (PMI) [26]. Other corpora used to construct emotion lexicons are web n-grams [27] and tweets [21]. For Twitter data, the following iterative algorithm can grow the lexicon of emotional hashtags: at each iteration, it learns an emotion classifier from the data extracted by the given hashtags, and applies it to the new tweets to discover new hashtags [28]. The main limitation is that such lexicon-growing methods were designed and evaluated to generate domain-independent resources; whereas the lack of domain-specific contextual knowledge and emotional expressions limits their application.

Supervised machine-learning algorithms are appropriate for training on domain data. For emotion recognition, researchers have experimented with different classifiers, such as NaïveBayes or SVM, and with various linguistic, stylistic, and syntactic features, such as n-grams, punctuation marks, parts of speech, and topics [4], [21], [29], [30]. However, supervised techniques require substantial annotated data, which are expensive to obtain for each domain.

With Twitter, researchers overcome the lack of annotated data by crawling the tweets with emotional hashtags and emoticons [5], [6], [21], [31]. Following the idea of distant learning—a kind of semi-supervised learning—such tweets serve as pseudo-annotated data and are used to train machine-learning classifiers in a supervised manner. This approach avoids costly manual annotation and allows relatively free choice of the emotion categories and domains to study. Yet, for a concrete domain, like sports or financial events, such pre-coded hashtags are likely to be found in only a limited amount of tweets. In this work, we investigate whether a distant learning approach is viable when applied within a restricted domain, and when initial pseudo-annotation is performed by the available emotion classifiers instead of seed keywords.

Multiple other algorithms were designed for semi-supervised learning ([32] gives an overview). One method,

applied to emotion recognition, represents the given text corpus in a reduced-dimensionality vector-space model and assigns emotions based on similarity to computed emotion vectors [33]. This method, however, is not easily scalable and does not allow classification of unseen data. For multi-category text classification, a commonly applied method is Naïve Bayes extended with the Expectation-Maximization procedure [34]. It first iteratively learns the parameters over the currently annotated data, and then re-annotates the data using those found parameters. For comparison, we also apply a Naïve Bayes classifier in our experiments, but start from the data pseudo-annotated by the given initial classifier.

Additionally, we review the advances of semi-supervised methods for binary polarity classification, a problem closely related to emotion recognition. Experiments show semi-supervised classifiers outperform supervised ones when few labeled data are available [35]. The idea of distant learning to train polarity classifiers from the pseudo-annotated data is successfully applied to Twitter data as well [7], [8]. Among other methods, iterative self-training approach was shown to be profitable [36]. To compare, we adapt one method used both for polarity and emotion classification: the classifier based on computing Pointwise-Mutual Information (PMI) between terms and emotion categories [21], [37].

On the whole, none of the related work studied how to apply the semi-supervised learning framework for multi-category emotion classification within a specific domain of tweets—the main problem we tackle in this paper. While previous emotion recognition methods were designed for a small set of categories, we design and validate the method that is able to cope with multiple emotion categories (a fine-grained problem). Also, we are, to the best of our knowledge, the first to deal with the problem of unbalanced emotion distribution present in a given corpus. We also compare the designed method with the two established algorithms from the related subject areas of polarity and text classification.

III. Semi-Supervised Method Details

We start by introducing the definitions used to describe the problem and the method suggested. We formulate the problem of emotion recognition as a multi-label classification task. Given the set of emotion categories $E = \{e_1, e_2, ..., e_{|E|}\}$, the classifier detects for a given document d—in our case a tweet—which emotion categories are expressed and outputs their label set $\{e_{i_k}\} \subseteq E$. If no emotion is present, the *Neutral* label e_0 is output.

We also define the *emotionality* of the text $\bar{p} = (p_1, p_2, \ldots, p_{|E|})$ as the distribution of the emotion categories expressed in the text, with $\sum_{i=1}^{|E|} p_i = 1$ and $\forall i\ p_i \geq 0$, where p_i is the weight of the ith emotion. The emotionality can be transformed into a multi-label by applying a technique adapted from the alpha-cut for fuzzy sets [38]. We denote this operator as $\mathfrak{A}: (\bar{p}, \alpha) \to 2^E$, where α defines a threshold on the emotion weight for the emotion to be included in the multi-label. $\mathfrak{A}(\bar{p}, \alpha)$ returns all the labels e_i that have the weight $p_i \geq \alpha \cdot p^*$, where $p^* = \max_i p_i$ is the maximal emotion weight within the distribution. Thus, all the labels with the weight close enough to the maximum weight are output. If $\alpha = 1$, only the labels with the maximum weight are output. For example, for the emotionality $(0, 0.2, 0.3, 0.5, 0, ..., 0)$ the multi-label $\{e_3, e_4\}$ would be found for $\alpha = 0.5$. In the opposite direction, a multi-label can be transformed into the emotionality by specifying the weights of present labels being 1 and then normalizing the distribution.

Our semi-supervised method is portrayed in Figure 1. It requires as an input some limited emotion classifier, taken as an *initial classifier* I, and data collected within a desired domain, considered as *unlabeled data* U. The first step is to apply this initial classifier I to the unlabeled data U in order to obtain the *pseudo-labeled data* L. We assume that the *initial classifier* returns the emotionality for the given text d, that is it assigns to the document $d \in U$ the emotionality $\bar{p}(d) = (p_1(d), p_2(d), \ldots, p_{|E|}(d))$. The pseudo-labeled data L contains the set of tweets with the mapped emotionalities. Those tweets where emotions were not found, i.e. the neutral ones, are not included in L. The pseudo-labeled data generated in this way are the entry point of the learning process described below.

The learning process starts from the *annotation refinement*. It is applied to each tweet individually. Given the parameter α, it sets to zero the weights of those emotions that would not be included in the multi-label: $e_i \notin \mathfrak{A}(\bar{p}, \alpha)$, and then normalizes the distribution. This eliminates emotions with relatively low weights. Whether to apply this refinement or not is also the parameter of the method.

The next step is to *select features* over which the classifier will be learned. We use 1-, 2-,..., n-grams as features. We keep only those that appeared K or more times in the pseudo-labeled dataset L. Among them, we select features that are indicative of emotions by estimating their polarity. For this, we compute a term's semantic orientation using the Pointwise-Mutual Information (PMI) [26]. We first identify the polarity label (l^+ or l^-) of each tweet $d \in L$ as $sign\left(\sum_{i \in E^+} p_i(d) - \sum_{i \in E^-} p_i(d)\right)$, where $E^+ \subset E$ and $E^- \subset E$ are the sets of positive and negative emotions correspondingly. Then the semantic orientation SO of a term t is computed by

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

where V is the used vocabulary of terms, $freq(l^{\pm})$ is the number of positive (l^+) or negative (l^-) tweets, while

$freq(t, l^{\pm})$ is the number of tweets with the term t, which are either positive or negative. Smoothing is used in the formula: we add 1 to each term frequency computation, and $|V|$ to class frequency computation (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 representation of the documents from L.

Having the documents (i.e. tweets) with the associated emotionalities and their feature representation, we can now learn a final classifier in a supervised manner. We apply Balanced Weighted Voting as such a *supervised learner*. Its choice also defines how the *resultant classifier* will work.

A. Balanced Weighted Voting

This algorithm is based on the emotion lexicon, in which each feature (n-gram entry in our case) has the associated emotionality of the term $\bar{w}(t) = (w_1(t), w_2(t), \ldots, w_{|E|}(t))$, where $w_i(t)$ is the weight of the term t for the emotion i. To compute the emotionality of the tweet document $\bar{p}(d)$, we search for the terms in the lexicon within its text, sum the weights of the found lexicon entries and normalize the vector. If no lexicon terms were found, the *Neutral* label is returned. Otherwise, the output is an emotion multi-label obtained from the found emotionality with the operator $\mathfrak{A}(\bar{p}(d), \alpha_0)$, where α_0 is the algorithm parameter. This weight-based application structure is similar to the one used in [39]. However, the learning process is different.

For learning, we know the emotionality for each tweet d $\bar{p}(d) = \left(p_1(d), p_2(d), \ldots, p_{|E|}(d)\right)$ returned by the initial classifier. In Balanced Weighted Voting (BWV), we first balance the distribution of the 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

$$w_i(t) = \frac{\sum\limits_{d:t\in d} c_i \cdot p_i(d)}{\sum\limits_{i}\sum\limits_{d:t\in d} c_i \cdot p_i(d)} \tag{2}$$

After the preliminary evaluations with several rebalancing options, we define the coefficient for the i-th emotion as $c_i = \frac{1}{\sum\limits_{d} |d| \cdot p_i(d)}$, where $|d|$ represents the number of extracted features in the tweet d. This means that the emotion categories are balanced based on the number of features appeared in each of them, as was inspired by [40].

The original Weighted Voting approach [39] is different from BWV only in that it lacks the rebalancing coefficients c_i. This makes it project the distribution of emotions in the annotated data onto the emotionality of each term. A term obtains a higher weight for the emotion which appears more often. Thus, the created lexicon is biased towards more dominant emotions. Our approach, Balanced Weighted Voting, involves re-weighting the emotional assignments of the tweets to cope with the skewness of the distribution. This re-weighting process is equivalent to the re-sampling approaches applied to cope with class imbalances for the classification problems [41].

IV. Experiments

In our evaluation, we focus on the domain of sports events reactions in Twitter. This domain was chosen because it contains various emotions with domain-specific emotional expressions; and because it was already studied in the context of multi-category emotion recognition, resulting in the availability of a limited within-domain classifier [39].

A. Emotion Model

In continuation with the previous work [39], we use the same 20 emotion categories from Geneva Emotion Wheel (GEW, v. 2.0). This model was developed in the psychological research in order to systematize self-reports on emotional experience [9]. The categories are enumerated in Table I.

GEW has multiple advantages. Whereas common sets of basic emotions, such as Ekman's [42] or Plutchik's [43], contain up to 8 categories, the higher granularity of GEW allows discovering more insightful details about emotional reactions. Moreover, it contains as many positive emotions as negative ones (10)—a rare characteristic for emotion models. Compared with the OCC model [44] (another fine-grained categorization model with 22 categories differentiated based on cognitive attribution of emotion-invoking factors), we believe that GEW emotions are more likely to be distinguished without extracting cognitive attributes. Another alternative could be the 24 primary emotions from the Hourglass of Emotions [45], the advanced representation of Plutchik's emotion wheel distinguishing 4 affective dimensions and specifying 6 emotion levels in each. However, this model lacks cognitive-based emotions such as *Pride*, *Envy*, or *Pity*, while *Pride*, for example, was shown previously to be dominant in the domain of sports events reactions [39].

B. Data Description

Our data consist of Twitter posts collected during the two weeks of the 2012 Olympic Games by querying Olympic-related keywords, such as "Olympic" or "London2012". This resulted in 33.2 million tweets written in English. We use different subsets of those tweets in the experiments.

1) Unlabeled Data: We randomly chose 250,000 tweets as the unlabeled data within the semi-supervised framework. They are filtered: we included only the tweets containing more than 3 words (not counting hashtags), that are not a retweet and which have no URLs present. We also avoided including tweets with duplicate text.

2) Data Labeled with Emotional Hashtags: Based on the GALC lexicon [9], we define the set of 167 emotional hashtags for all GEW emotion categories.[1] By extracting the tweets containing those hashtags at the end of the text, we generate the pseudo-annotated tweets used for tuning the meta-parameters of the algorithms and for the final tests. We again exclude the tweets with URLs, retweets, or repeated tweets. The distribution of the refined version of this dataset is given in Table I (Full set). In a similar approach, but with fewer emotions, Wang et al. [5] evaluated precision of tweet-emotion associations to be 93.16%, which we consider of substantial quality. We also exclude tweets where several emotional hashtags were found. As these data are intended for testing the algorithms' outputs, the hashtags used for generation of the labels are removed from the texts.

3) Presumably-Neutral Data: Successful emotion recognition also implies an effective distinction between the *Neutral* and *Emotional* categories. While we do not introduce a hierarchical classification [46], we consider *Neutral* as a separate class e_0. Classifiers then work with the extended set of categories $E^0 = E \cup e_0$. If e_0 is within the multi-label output, we output only *Neutral* category. This avoids constructing classifiers that detect emotions in all tweets.

To identify presumably-neutral tweets, we assume that the presence of a URL can indicate less emotional tweets, such as news or information sharing. We extracted such tweets and observed that to enforce tweet neutrality, we should in addition avoid presence of usernames (which makes sharing personal) and emoticons (which explicitly indicates the presence of emotions). In the observation of 100 tweets we discovered 19 emotional ones, which we consider to be acceptable for such heuristic labeling.

Among all found presumably-neutral tweets, we randomly select $250,000$ to supplement the unlabaled data. An initial classifier is also applied to these tweets and we exclude the ones with detected emotions. All others appear as assigned to the Neutral class. We name these data N. We also extract some additional presumably-neutral tweets for testing purposes as described below.

4) Preprocessing Steps: All the collected tweets are first preprocessed. We replace each emoticon by a distinct placeholder to ensure their correct extraction as separate tokens. We also replace usernames with a placeholder and remove stop-words. All the texts are converted to lower-case. Punctuation marks are included as separate tokens. We also delete hashtag symbols (#) from the text.

C. Validation and Test Sets

We construct a validation set to tune the meta-parameters of the algorithm and the test set to evaluate the resultant classifiers. We include both emotional tweets with hashtags and presumably-neutral tweets. In the collected hashtag-based dataset, we observe the large skewness of the emotion distribution (see Table I). Some classes are present only in a few tweets (as few as only 17 tweets for *Contempt*). As we believe that an emotion classifier should be able to correctly distinguish emotions for any given distribution, we exclude the influence of the given skewness on the evaluation by balancing the dataset between emotion classes. Thus, our test dataset is balanced, with a similar number of tweets per emotion, as is the dataset for validation.

We randomly chose a maximum of 200 tweets for each class, including Neutral, to include in the test dataset (avoiding tweets from the unlabeled data). We follow the same process for the separate validation set. When there are fewer than 400 tweets in a class, we split them between test and validation sets, preserving round numbers (10, 50, 100 or 150) whenever possible. Overall, the test set has $3,360$ tweets, and the validation set has $3,184$ tweets. Table I presents the distribution of emotions in both datasets.

TABLE I. DATASETS STATISTICS: PER-CATEGORY NUMBER OF TWEETS DETECTED WITH EMOTIONAL HASHTAGS WITHIN SPORTS DOMAIN

Emotion category	Full set	Validation set	Test set
Involvement/Interest	1669	200	200
Amusement/Laughter	195	50	100
Pride/Elation	22172	200	200
Happiness/Joy	3497	200	200
Pleasure/Enjoyment	364	150	150
Love/Tenderness	8278	200	200
Awe/Wonderment	251	100	100
Relief/Disburned	134	50	50
Surprise/Astonishment	1665	200	200
Nostalgia/Longing	567	200	200
Pity/Compassion	93	30	50
Sadness/Despair	10555	200	200
Worry/Fear	1296	200	200
Shame/Embarrassment	2317	200	200
Guilt/Remorse	276	100	100
Regret/Disappointment	2754	200	200
Envy/Jealousy	4390	200	200
Disgust/Repulsion	403	100	200
Contempt/Scorn	17	4	10
Anger/Irritation	2985	200	200
Neutral	-	200	200
Total	63878	3184	3360

D. Initial Classifiers

The initial classifier is the starting point of the learning process in our semi-supervised framework. We consider two domain-independent emotion classifiers: one lexicon of explicit emotional terms (GALC) and one machine-learning classifier trained on general data (MNB-Hash). We also take one domain-specific classifier constructed over the small annotated dataset using human computation (OlympLex).

1) GALC: The GALC is the domain-independent emotion lexicon of the unigram stems explicitly expressing an emotion, e.g. "happ*" for *Happiness/Joy*. It was developed along with GEW to automatically classify survey

[1]The list of used hashtags is available at http://hci.epfl.ch/emotions-in-olympic-tweets/galc-emotion-hashtags-file

responses [9]. It contains 279 stemmed terms for 36 emotion categories (covering all 20 GEW categories). To compute the emotionality of a document using this lexicon, we sum the number of terms found for each emotion (excluding the negated ones) and normalize the obtained vector.

2) OlympLex: This domain-specific emotion lexicon was obtained over the annotation of the tweets about sports events in crowdsourcing settings. It contains the emotion indicators selected from those tweets by the labelers as well as related user-entered emotional expressions [39]. This emotion lexicon allocates an emotionality for each of its 3193 terms (from unigrams to 5-grams). The average of those emotionalities for the terms found within the tweet text (excluding the ones negated or covered by other found terms) is the emotionality of the tweet.

3) MNB-Hash: We also include the state-of-the-art domain-independent classifier: we train a Multinomial Naïve Bayes model over the large dataset of tweets collected with the emotional hashtags. For this, we collected 5 million English tweets with one of the emotional hashtags (the same as used before for the test data labeling). After the selection of tweets where hashtags appear in the end, the deletion of retweets, short tweets and duplicates, we had 669,216 tweets suitable for training the model. We then added the 250,000 neutral tweets extracted in our Olympic data (dataset N). We used the unigrams and bigrams as features and applied the same preprocessing steps to the text. This approach achieved 57.7% accuracy in the 10-fold cross validation, with a macro F1-score of 29.8%. For comparison, in similar supervised settings other researchers achieved a macro F1-score of 25.3% for 11 emotional states [6], and 53.5% for 7 emotions [5].

E. Tuning Meta-Parameters

We choose the optimal learning meta-parameters of our method by tuning it over the validation dataset. This process is separate for each initial classifier. To increase the computation speed, we used only 100,000 unlabeled tweets and 100,000 presumably-neutral tweets in these experiments. The following parameters are varied:

- the length n of n-grams features: from 1 to 5;
- the minimum occurrence of n-grams $K = 5$ (fixed);
- the threshold τ of feature selection: 0 (no selection), 0.1, 0.2, 0.3, 0.7, and 1.0;
- α used in the operator $\mathfrak{A}$: 0.5, 0.7, 0.9, and 1.0;
- whether the annotation refinement is applied (with the parameter α specified earlier);
- α_0 of the multi-label selection for output: either $\alpha_0 = \alpha$ (considering α as a general characteristic of the problem) or $\alpha_0 = 1$ (outputting only dominant emotions).

For each set of parameters, we recorded the performance of the corresponding algorithm instance with macro-precision, macro-recall and macro-F1 scores over the emotion categories (excluding *Neutral* and *Contempt*, which was under-represented in the dataset). The review of these results revealed several possible behaviors of the algorithm. To cover all of them, we apply three strategies to define the best parameters. In the first setting, we maximize macro F1-score, as it is usually considered to capture the best trade-off between precision and recall (*F1-based* settings S_{F1}). In the second setting, we maximize macro-precision, because it is harder to optimize than recall, which can be increased by outputting more emotion labels for a tweet (*Precision-based* settings S_P). In the third setting, we maximize macro F1-score while ensuring that precision is greater than recall (*Centered* settings S_C). This excludes the cases where the maximum of F1-score is achieved with a low precision.

As a result, we identified the three best performing learning parameters of our algorithms separately for each initial classifier. All chosen parameters are described in the Appendix, Table VI.

F. Validation of Improvement over the Initial Classifiers

We now present and discuss the test results of the tuned instances of our semi-supervised method. For each of the three initial classifiers (GALC, OlympLex, and MNB-Hash), we run the learning process of the Balanced Weighted Voting (*BalancedWV*) under three sets of meta-parameters chosen from the validation experiments (S_{F1}, S_P, and S_C). In this process, all the 250,000 unlabeled tweets and 250,000 of presumably-neutral tweets were accessible for learning. The resultant classifiers were then evaluated on the test dataset. Table II presents these results.

We again compute macro F1-score (F1), macro-precision (P) and macro-recall (R). The performance of the initial classifiers without semi-supervised learning is reported as *Initial*. For each semi-supervised algorithm, we also report how significantly its performance metrics are different from the ones of corresponding initial classifiers. We use Wilcoxon signed rank test for this goal. An asterisk indicates a p-value of 0.05 or lower, and two asterisks indicates a p-value of 0.01 or lower (The same notation is applied for all tables). We also adapt a random baseline (*Random*) to estimate the difficulty of the problem: it decides independently for each emotion if it is present or not with the probability provided by the emotion distribution in the dataset. *Random* has macro F1-score of 4.8% for the test data, as estimated by 1000 runs.

The results show that under all three settings our semi-supervised method improves the macro F1-score of the initial classifiers. The increase is statistically significant in all cases except one, with a minimum relative increase of 24.4%. The largest improvements (and thus highest F1-scores) are achieved by the F1-based setting S_{F1}: 26.8% when started from MNB-Hash, 76.1% for OlympLex, and 105% for GALC (with the maximum achieved F1-score of 20.5% with MNB-Hash as a start). These findings confirm our hypothesis that unlabeled data can be leveraged to improve the performance of initial classifiers within a domain.

TABLE II. VALIDATION OF PERFORMANCE IMPROVEMENT AFTER THE APPLICATION OF BALANCED WEIGHTED VOTING. MACRO METRICS ARE REPORTED.

	Algorithm	Setting	GALC			OlympLex			MNB-Hash		
			P	***R***	***F1***	***P***	***R***	***F1***	***P***	***R***	***F1***
	Initial	-	**19.7**	4.8	6.7	**18.2**	9.7	9.2	**25.3**	13.6	16.2
BWV-S_{F1}	BalancedWV	S_{F1}	11.0**	**19.2****	**13.7****	14.5	**20.3****	**16.2****	17.0**	**27.3****	**20.5****
BWV-S_P	BalancedWV	S_P	16.3	6.8**	8.9*	16.8	10.6	12.1	22.8	19.4**	20.1*
BWV-S_C	BalancedWV	S_C	13.3	13.5**	13.1**	17.0	15.4	15.0**	22.2	19.8**	20.2*
Improvement of BWV-S_C over Initial					95%			63%			25%
Improvement of BWV-S_C over Random					182%			224%			337%

This improved performance can be explained for all settings by the increase of macro-recall, statistically significant in most cases. Unfortunately, it leads to the lower macro-precision. This decrease is statistically significant only for the S_{F1} setting, despite it having the highest F1-score on all initial classifiers. At the same time, the recall increase for this setting can be explained by outputting multiple labels for a tweet (average is 2 for this setting, while around 1 for other two settings). Thus, the F1-based setting can be preferred only if lower precision is not an issue for the application.

The decrease in macro-precision is statistically insignificant for two other settings, S_P and S_C (p-value > 0.05). Although their results are comparable when started from MNB-Hash (p-value for F1 is 0.768), from GALC or OlympLex, the F1-score (and recall) of S_P is lower than that of S_C (p-value for F1 with OlympLex is 0.029). Therefore, we see the Centered settings S_C as the most suitable for real applications when both recall and precision are of equal importance.

Overall, our evaluation indicates that with the described semi-supervised method we are able to make the initial classifiers more suitable for an application within a chosen domain. This is achieved by correctly detecting more emotional documents (increased recall), while maintaining a comparable level of precision in most cases.

G. Validation of Rebalancing Process

The suggested method, Balanced Weighted Voting, originates from Weighted Voting, which does not introduce the rebalancing coefficients c_i (as described in Section III). We test what effect the rebalancing process brings. We first run the equivalent validation experiments to find the best parameters for Weighted Voting. Then, we evaluate the obtained classifiers for the same three parameter settings on the test dataset, starting from the three initial classifiers. We present the test results of Weighted Voting when started from GALC in Table III. The performance patterns are similar for the other two initial classifiers.

We observe that without rebalancing, Weighted Voting was not able even to increase the F1-score of the initial classifier. While the precision remains comparable (p-value $>$ 0.05), the recall decreases significantly (p-value $= 0.001$ for S_{F1}). This means that such an algorithm without rebalancing is not suitable for our semi-supervised learning, at least not under the same framework.

H. Comparison with Other Methods

We compare our Balanced Weighed Voting classifier with the other two classifiers imported from related subject areas of text classification and sentiment analysis.

1) Naïve Bayes: We consider the machine-learning classifier most widely used for text classification tasks—Multinomial Naïve Bayes. As tweets are short, we interpret the presence of the terms as features, instead of their frequency. While the original classifier was designed for the hard classification requiring label input, we adapted it to our settings of soft classification with emotionality input by following the process described in [34]. The conditional probability of the term t given the emotion class is computed as:

$$P(t|e_i) = \frac{1 + \sum\limits_{d:t\in d} p_i(d)}{|V| + \sum\limits_{s\in V}\sum\limits_{d:s\in d} p_i(d)}, \tag{3}$$

where $|V|$ is the size of the feature vocabulary. The classification procedure is unchanged: we output the class(es) that have the highest probability conditioned on the tweet text $P(e_i|d)$. However, we consider the found vector $P(e_i|d)$ as the emotionality, and output the multi-label using the α_0-based operator $\mathfrak{A}$.

2) Pointwise-Mutual Information Classifier: The PMI-based classifier is used in polarity and emotion classification within semi-supervised settings [21], [26]. It splits the problem into $|E|$ independent binary classification tasks: a classifier for ith emotion decides if it is present (class e_i^+) or not (e_i^-). For learning, we transform the given emotionality $\bar{p}$ of each document into the multi-label format, using the operator $\mathfrak{A}$ with the parameter α (the same as the annotation

TABLE III. SHOWING THE IMPROVEMENT OF REBALANCING USED IN BALANCED WEIGHTED VOTING, WHEN STARTED FROM GALC. MACRO METRICS ARE REPORTED.

	Algorithm	Setting	GALC		
			P	***R***	***F1***
	Initial	-	19.7	4.8	6.7
WV-S_{F1}	Weighted Voting	S_{F1}	**20.7**	3.6**	5.3**
WV-S_P	Weighted Voting	S_P	19.3	2.8**	4.4**
WV-S_C	Weighted Voting	S_C	20.1	3.4**	5.1**
BWV-S_C	BalancedWV	S_C	13.3	**13.5****	**13.1****
Improvement of BWV-S_C over WV-S_{F1}					147%

TABLE IV. COMPARING BALANCED WEIGHTED VOTING WITH OTHER ALGORITHMS. MACRO METRICS ARE REPORTED.

	Algorithm	Setting	GALC			OlympLex			MNB-Hash		
			P	***R***	***F1***	***P***	***R***	***F1***	***P***	***R***	***F1***
	Initial	-	**19.7**	4.8	6.7	**18.2**	9.7	9.2	25.3	13.6	16.2
NB-$S_{F1(C)}$	NaïveBayes	$S_{F1(C)}$	14.5	12.7*	10.1	16.9	16.1**	11.0	**28.4**	15.1	15.8
NB-S_P	NaïveBayes	S_P	15.9	6.5	7.6	16.5	10.9	8.2*	27.1	14.7	15.4
PMI-S_{F1}	PMI-based	S_{F1}	6.8**	37.0**	11.1**	9.2**	**30.7****	13.3*	12.6**	**27.1****	16.3
PMI-S_P	PMI-based	S_P	8.1**	29.4**	10.7*	8.6**	26.1**	10.6	16.1*	18.3**	14.9
PMI-S_C	PMI-based	S_C	6.9**	**41.3****	11.1*	9.4**	28.9**	13.3*	11.7**	26.2**	15.1
BWV-S_C	BalancedWV	S_C	13.3	13.5**	**13.1****	17.0	15.4	**15.0****	22.2	19.8**	**20.2***
Improvement	BWV-S_C vs. NB-$S_{F1(C)}$				29%			37%			28%
Improvement	BWV-S_C vs. PMI-S_{F1}				18%			13%			24%

refinement parameter). The ith binary classifier is learned by computing the strength of association (SoA) of each term with the ith emotion as the PMI difference towards the e_i^+ and e_i^- classes. The formula (1) is used again, but while considering emotion presence e_i^+ as a positive class and emotion absence e_i^- as a negative. We also add an extra parameter θ for this algorithm. It is used for feature selection at a per-category level: the ith classifier takes only those features that have a strength of association above this threshold, i.e. $|SoA(t,e_i)| \geq \theta$. Then, the emotion e_i is detected as present in a tweet if the sum of the scores $SoA(t,e_i)$ of the feature terms found in the text is positive.

3) Test Methodology: Both of these classifiers are taken as supervised learners within the semi-supervised framework (see Figure 1). For each, we perform validation experiments with the same validation set, parameter space, and the unlabeled data, as for Balanced Weighted Voting. We only add an extra tuning of the parameter θ for the PMI-based classifier, with evaluated possible values being 0, 0.1, 0.2, 0.3, 0.7, and 1.0. Similarly, we choose three sets of meta-parameters (S_{F1} for F1-based, S_P for Precision-based, and S_C for Centered settings). However, S_C coincides with S_{F1} for the NaïveBayes classifier—we name them $S_{F1(C)}$. We then learn the classifiers using the full data (250,000 of unlabeled tweet and 250,000 of pseudo-labeled ones). The results of the learned classifiers on the test dataset are shown in Table IV. We again present macro-Precision (P), macro-Recall (R) and macro-F1 score (F1) for evaluation.

4) Comparative Results: First, the results show that both baseline algorithms are also able to improve the F1-score of an initial classifier (under some settings). However, these improvements, if present, are statistically significant only for the PMI-based classifier. It achieves this high F1-score by optimizing recall with a large sacrifice of precision (with statistically significant changes in all settings). Such a recall-favoring behavior is a result of making an independent decision for each emotion, which makes it prone to output more emotion labels per tweet (4.45 in average for all cases). The Naïve Bayes classifier has a different behavior— in all cases, its macro-precision is higher than macro-recall, and in case of MNB-Hash it is even higher than that of the initial classifier (it is not statistically significant, though). Yet, its F1-scores are mostly comparable to those of the initial classifiers (no significant difference at significance level 0.05). For both baseline algorithms, the highest F1-scores are achieved with the F1-based setting S_{F1} and when started from MNB-Hash.

In any case, the F1-scores of Balanced Weighted Voting with Centered setting S_C are higher for all initial classifiers. And this result is achieved with lower average difference between macro-precision and macro-recall (1.4% versus the minimal 5.3% for NB-$S_{F1(C)}$).

We further compare the test results of the best-performing settings for all classifiers at per-category level (Table V). We take MNB-Hash as an initial classifier, because the highest F1-scores are achieved with it. Its per-category performance is also presented in order to clarify the starting point of improvement. The following best-performing settings are used in comparison: S_C for Balanced Weighted Voting, $S_{F1(C)}$ for Naïve Bayes, and S_{F1} for the PMI-based classifier.

We discover that BWV-S_C not only achieves the highest macro F1-score, but also has the highest F1-scores for most of the categories (for 10 out of 20 emotions, while only for 4 for NB-$S_{F1(C)}$). For those categories on which it has F1-score lower than of another semi-supervised classifier, the absolute difference to the highest is no greater than 3.4%. In addition, BWV-S_C has a smaller average difference between precision and recall than either NB-$S_{F1(C)}$ or PMI-S_{F1} (6.1% vs. 16.7% or 15.4%). This means that its performance is also more stable in terms of the balance between precision and recall among emotion categories.

The per-category comparison also shows that the low macro F1-score of NB-$S_{F1(C)}$ is mostly due to its poor results (F1-score $< 5\%$) on those categories that have been found with an insufficient recall by the initial classifier ($R \geq 6\%$). At the same time, both BWV-S_C and PMI-S_{F1} can overcome this problem at least for some of these categories (for example, while MNB-Hash initially had a F1-score of 0% on *Amusement*, BWV-S_C has 17.1%).

In comparison with two other learning algorithms, Naïve Bayes and PMI-based, our method, Balanced Weighted Voting, achieves better test performance not only in terms of macro F1-score, but also in terms of the balance between precision and recall both at macro and per-category levels.

TABLE V. PER-CATEGORY RESULTS OF THE CLASSIFIERS, WHEN STARTED FROM MNB-HASH.

	Initial: MNB-Hash			BWV-S_C			NB-$S_{F1(C)}$			PMI-S_{F1}		
Category	***P***	***R***	***F1***	***P***	***R***	***F1***	***P***	***R***	***F1***	***P***	***R***	***F1***
Involvement / Interest	37.7	10.0	15.8	25.0	15.5	**19.1**	41.1	11.5	18.0	9.7	10.0	9.8
Amusement / Laughter	0.0	0.0	0.0	20.0	15.0	**17.1**	100.0	1.0	2.0	4.5	7.0	5.5
Pride / Elation	24.7	31.5	27.7	31.5	37.0	34.0	30.9	41.0	**35.3**	17.0	44.0	24.5
Happiness / Joy	35.9	23.5	28.4	29.8	26.5	28.0	32.2	28.0	29.9	25.0	40.0	**30.8**
Pleasure / Enjoyment	0.0	0.0	0.0	8.6	4.7	6.1	-	0.0	-	11.6	6.7	**8.5**
Love / Tenderness	24.1	19.5	**21.5**	23.7	11.5	15.5	20.8	16.5	18.4	13.7	20.5	16.4
Awe / Wonderment	0.0	0.0	0.0	9.1	8.0	**8.5**	-	0.0	-	8.0	6.0	6.9
Relief / Disburned	28.6	8.0	12.5	11.3	30.0	**16.4**	30.0	6.0	10.0	7.5	24.0	11.4
Surprise / Astonishment	30.8	2.0	3.8	27.1	13.0	**17.6**	50.0	1.0	2.0	12.8	11.5	12.1
Nostalgia / Longing	50.9	27.0	35.3	35.9	44.5	**39.7**	43.2	27.0	33.2	19.4	33.5	24.5
Pity / Compassion	0.0	0.0	0.0	1.7	4.0	**2.4**	-	0.0	-	0.0	0.0	-
Sadness / Despair	26.5	35.0	30.2	29.6	29.5	29.6	26.1	44.5	**32.9**	14.7	64.0	23.9
Worry / Fear	37.4	20.0	**26.1**	20.5	16.5	18.3	23.9	16.0	19.2	12.4	24.0	16.4
Shame / Embarrassment	29.1	16.0	20.6	22.2	17.5	19.6	22.8	22.0	**22.4**	14.6	41.5	21.6
Guilt / Remorse	37.5	6.0	10.3	20.8	16.0	**18.1**	0.0	0.0	-	9.6	23.0	13.5
Regret / Disappointment	19.3	10.5	13.6	24.2	16.0	19.3	17.5	11.0	13.5	14.2	43.5	**21.5**
Envy / Jealousy	32.6	22.0	26.3	32.3	27.0	29.4	35.5	30.5	**32.8**	17.0	41.0	24.0
Disgust / Repulsion	37.5	6.0	10.3	18.2	17.5	**17.9**	41.7	2.5	4.7	12.9	19.0	15.4
Contempt / Scorn	0.0	0.0	0.0	0.0	0.0	-	-	0.0	-	-	0.0	-
Anger / Irritation	28.6	22.0	24.9	30.2	26.0	**28.0**	23.9	28.0	25.8	14.2	55.0	22.5
Neutral	11.0	93.5	19.7	17.6	46.5	**25.5**	11.9	83.5	20.9	10.6	61.0	18.0
Macro (of emotions)	25.3	13.6	16.2	22.2	19.8	**20.2**	28.4	15.1	15.8	12.6	27.1	16.3

V. DISCUSSION AND FUTURE WORK

Our experiments showed that semi-supervised approaches can achieve substantial improvements over the initial classifiers. The best achieved performance is a 20.5% macro F1-score, which reflects the difficulty of the considered multi-category emotion recognition problem: its fine-granularity combined with subjectivity of classification, lack of annotated data and low random baseline. While the larger number of emotions (20 against the 6 or 8 basic emotions more commonly used in emotion recognition literature [4], [21]) makes the classification more difficult, it brings the opportunity to select those emotions that are more suitable for the domain or application. One future research direction is to investigate how the choice of categories affects the results. We expect that we can achieve higher recognition quality with emotions that are more separable and that have many descriptive expressions.

Another direction for future work is to determine whether our results can be generalized for other domains (e.g. the reactions to other public events such as awards or elections). It would be also interesting to study if the discovered benefits of semi-supervised learning hold for larger or smaller amounts of unlabeled data, and whether better performance can be achieved by other feature selection methods. It is also possible that a hierarchical approach to the classification problem [46] could bring further improvements.

VI. CONCLUSION

With this paper, we believe we are the first to study a semi-supervised learning method for multi-category emotion recognition in tweets from a specific domain. We describe a method that, starting from an existent but limited initial classifier (e.g. a general-purpose emotion lexicon), constructs a novel classifier that is able to correctly detect more domain-specific emotional tweets and thus is more suitable to apply within the chosen domain. Using sports tweets, we validate this approach experimentally on the three different initial classifiers. In all three cases, the proposed semi-supervised method, Balanced Weighted Voting, improves the macro F1-score, with a relative increase between 24% and 105%. Our further experiments suggest that rebalancing the initially labeled data prior to training the classifier is an essential step for the success of our method. Finally, in comparison with other two learning algorithms (Naïve Bayes and PMI-based), Balanced Weighted Voting achieves the highest final macro F1-score, with consistently-high F1-scores throughout the emotion categories and with less difference between precision and recall both at macro and per-category levels.

REFERENCES

[1] D. Quercia, L. Capra, and J. Crowcroft, "The social world of Twitter: Topics, geography, and emotions." in *Proc. ICWSM*, 2012.

[2] R. W. Picard and J. Klein, "Computers that recognise and respond to user emotion: theoretical and practical implications," *Interacting with computers*, vol. 14, no. 2, pp. 141–169, 2002.

[3] G. Mishne and M. de Rijke, "MoodViews: Tools for blog mood analysis," in *AAAI Spring Symp.: Comput. Approaches to Analyzing Weblogs*, 2006, pp. 153–154.

[4] S. Aman and S. Szpakowicz, "Identifying expressions of emotion in text," in *Text, Speech and Dialogue*. Springer, 2007, pp. 196–205.

[5] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Harnessing Twitter "big data" for automatic emotion identification," in *Proc. Int. Conf. on Soc. Comput. (SocialCom)*. IEEE, 2012, pp. 587–592.

[6] M. De Choudhury, M. Gamon, and S. Counts, "Happy, nervous or surprised? Classification of human affective states in social media," in *Proc. ICWSM*, 2012.

[7] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," Standord, Tech. Rep., 2009.

[8] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *Proc. LREC*, 2010.

[9] K. R. Scherer, "What are emotions? And how can they be measured?" *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[10] K. Scherer, V. Shuman, J. Fontaine, and C. Soriano, "The GRID meets the wheel: assessing emotional feeling via self-report," *Components of emotional meaning: a sourcebook*, pp. 281–298, 2013.

[11] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. and trends in Inform. Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[12] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[13] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comput. linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[14] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," *J. Amer. Soc. for Inform. Sci. Technol.*, vol. 63, no. 1, pp. 163–173, 2012.

[15] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. ICWSM*, 2014.

[16] C. Strapparava and A. Valitutti, "WordNet Affect: an affective extension of WordNet." in *Proc. LREC*, vol. 4, 2004, pp. 1083–1086.

[17] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Comput. Intell.*, vol. 29, no. 3, pp. 436–465, 2013.

[18] S. Poria, A. Gelbukh, A. Hussain, D. Das, and S. Bandyopadhyay, "Enhanced SenticNet with affective labels for concept-based opinion mining," *IEEE Intell. Syst.*, pp. 31–38, 2013.

[19] L. Martin and P. Pu, "Prediction of helpful reviews using emotions extraction," in *Proc. AAAI*, 2014.

[20] S. Poria, A. Gelbukh, E. Cambria, A. Hussain, and G.-B. Huang, "EmoSenticSpace: A novel framework for affective common-sense reasoning," *Knowledge-Based Systems*, vol. 69 (Special Issue on Big Social Data Analysis), pp. 108–123, 2014.

[21] S. M. Mohammad, "#Emotional tweets," in *Proc. 1st Joint Conf. on Lexical and Comput. Semantics (*SEM)*. ACL, 2012, pp. 246–255.

[22] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Affect analysis model: novel rule-based approach to affect sensing from text," *Natural Language Engineering*, vol. 17, no. 1, pp. 95–135, 2011.

[23] U. Krcadinac, P. Pasquier, J. Jovanovic, and V. Devedzic, "Synesketch: An open source library for sentence-based emotion recognition," *IEEE Trans. on Affective Comput.*, vol. 4, no. 3, pp. 312–325, July 2013.

[24] "About WordNet," http://wordnet.princeton.edu, 2010.

[25] S. Poria, A. Gelbukh, E. Cambria, D. Das, and S. Bandyopadhyay, "Enriching SenticNet polarity scores through semi-supervised fuzzy clustering," in *Proc. SENTIRE, 12th Int. Conf. on DataData Mining Workshops (ICDMW)*. IEEE, 2012, pp. 709–716.

[26] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Trans. on Inform. Syst. (TOIS)*, vol. 21, no. 4, pp. 315–346, 2003.

[27] J. Perrie, A. Islam, E. Milios, and V. Keselj, "Using google n-grams to expand word-emotion association lexicon," in *Comput. Linguistics and Intell. Text Process.* Springer, 2013, pp. 137–148.

[28] A. Qadir and E. Riloff, "Bootstrapped learning of emotion hashtags # hashtags4you," *Proc. of NAACL-HLT WASSA 2013*, pp. 2–11, 2013.

[29] K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu, "EmpaTweet: Annotating and detecting emotions on Twitter." in *Proc. LREC*, 2012, pp. 3806–3813.

[30] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proc. 2008 ACM Symp. on Appl. Comput.* ACM, 2008, pp. 1556–1560.

[31] J. Suttles and N. Ide, "Distant supervision for emotion classification with discrete binary values," in *Comput. Linguistics and Intell. Text Process.* Springer, 2013, pp. 121–136.

[32] X. Zhu, "Semi-supervised learning literature survey," 2005.

[33] S. M. Kim, A. Valitutti, and R. A. Calvo, "Evaluation of unsupervised emotion models to textual affect recognition," in *Proc. NAACL-HLT 2010 Workshop on Comput. Approaches to Anal. and Generation of Emotion in Text*. ACL, 2010, pp. 62–70.

[34] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine learning*, vol. 39, no. 2-3, pp. 103–134, 2000.

[35] M. Wiegand and D. Klakow, "Predictive features in semi-supervised learning for polarity classification and the role of adjectives," in *Proc. NoDaLiDa*, 2009, pp. 198–205.

[36] L. Qiu, W. Zhang, C. Hu, and K. Zhao, "SELC: A self-supervised model for sentiment classification," in *Proc. CIKM*. ACM, 2009, pp. 929–936.

[37] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets," in *Proc. SemEval*. ACL, 2013, pp. 321–327.

[38] G. Bojadziev and M. Bojadziev, *Fuzzy sets, fuzzy logic, applications*. World Scientific, 1995, vol. 5.

[39] V. Sintsova, C. Musat, and P. Pu, "Fine-grained emotion recognition in olympic tweets based on human computation," in *Proc. NAACL-HLT WASSA*. ACL, 2013, pp. 12–20.

[40] E. Frank and R. R. Bouckaert, "Naive bayes for text classification with unbalanced classes," in *Knowledge Discovery in Databases: PKDD 2006*. Springer, 2006, pp. 503–510.

[41] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *Proc. ICAI*, vol. 1. Citeseer, 2000, pp. 111–117.

[42] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[43] R. Plutchik, "The nature of emotions," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[44] A. Ortony, G. L. Clore, and A. Collins, "The cognitive structure of emotions." 1988.

[45] E. Cambria, A. Livingstone, and A. Hussain, "The hourglass of emotions," in *Cogn. Behav. Syst.* Springer, 2012, pp. 144–157.

[46] D. Ghazi, D. Inkpen, and S. Szpakowicz, "Hierarchical versus flat classification of emotions in text," in *Proc. NAACL-HLT 2010 Workshop on Comput. Approaches to Anal. and Generation of Emotion in Text*. ACL, 2010, pp. 140–146.

APPENDIX

TABLE VI. THE PARAMETERS OF THE ALGORITHMS CHOSEN UNDER DIFFERENT SETTINGS

Algorithm	Initial Classifier	n	τ	α	α_0	Extra
BWV-S_{F1}	GALC	5	0.1	0.7	0.7	
BWV-S_P	GALC	1	1.0	0.7	1.0	
BWV-S_C	GALC	5	0.7	0.9	0.9	
BWV-S_{F1}	OlympLex	2	0.1	-	0.9	
BWV-S_P	OlympLex	1	1.0	0.5	1.0	
BWV-S_C	OlympLex	2	0.3	-	1.0	
BWV-S_{F1}	MNB-Hash	2	0.3	-	0.7	
BWV-S_P	MNB-Hash	2	0.1	-	1.0	
BWV-S_C	MNB-Hash	2	0.3	-	1.0	
NB-S_{F1}	GALC	3	0.1	-	0.5	
NB-S_P	GALC	1	0.3	0.9	0.9	
NB-S_{F1}	OlympLex	3	0.2	-	0.5	
NB-S_P	OlympLex	1	0.1	0.9	0.9	
NB-S_{F1}	MNB-Hash	3	0.1	-	0.9	
NB-S_P	MNB-Hash	5	0.1	1.0	1.0	
PMI-S_{F1}	GALC	1	0.7	0.7	0.7	$\theta = 0.3$
PMI-S_P	GALC	4	0.7	-	0.7	$\theta = 1.0$
PMI-S_C	GALC	1	0.1	-	0.5	$\theta = 0.7$
PMI-S_{F1}	OlympLex	1	0.7	0.5	0.5	$\theta = 0.7$
PMI-S_P	OlympLex	2	0.2	0.9	0.9	$\theta = 0.7$
PMI-S_C	OlympLex	1	0.2	-	0.7	$\theta = 1.0$
PMI-S_{F1}	MNB-Hash	2	1.0	0.5	0.5	$\theta = 0.7$
PMI-S_P	MNB-Hash	2	0.1	0.7	0.7	$\theta = 1.0$
PMI-S_C	MNB-Hash	5	1.0	0.5	0.5	$\theta = 0.3$