Fine-Grained Emotion Recognition in Olympic Tweets Based on Human Computation

Valentina Sintsova[a, b] **Claudiu Musat**[a] **Pearl Pu**[b]
[a]Artificial Intelligence Laboratory [b]Human Computer Interaction Group
School of Computer and Communication Sciences
Swiss Federal Institute of Technology (EPFL)
CH-1015, Lausanne, Switzerland
{valentina.sintsova, claudiu-cristian.musat, pearl.pu}@epfl.ch

Abstract

In this paper, we detail a method for domain specific, multi-category emotion recognition, based on human computation. We create an Amazon Mechanical Turk[1] task that elicits emotion labels and phrase-emotion associations from the participants. Using the proposed method, we create an emotion lexicon, compatible with the 20 emotion categories of the Geneva Emotion Wheel. GEW is the first computational resource that can be used to assign emotion labels with such a high level of granularity. Our emotion annotation method also produced a corpus of emotion labeled sports tweets. We compared the cross-validated version of the lexicon with existing resources for both the positive/negative and multi-emotion classification problems. We show that the presented domain-targeted lexicon outperforms the existing general purpose ones in both settings. The performance gains are most pronounced for the fine-grained emotion classification, where we achieve an accuracy twice higher than the benchmark.[2]

1 Introduction

Social media platforms such as Twitter.com have become a common way for people to share opinions and emotions. Sports events are traditionally accompanied by strong emotions and the 2012 summer Olympic Games in London were not an exception. In this paper we describe methods to analyze and data mine the emotional content of tweets about this event using human computation. Our goal is to create an emotion recognition method, capable of classifying domain specific emotions with a high emotion granularity. In the stated case, domain specificity refers not only to the sport event, but also to the Twitter environment.

We focus on the categorical representation of emotions because it allows a more fine-grained analysis and it is more natural for humans. In daily life we use emotion names to describe specific feelings rather than give numerical evaluations or specify polarity. So far, the multi-item emotion classification problem has received much less attention.

One reason is that high quality training corpora are difficult to construct largely due to the cost of human annotators. Further, if emotion representation is not carefully designed, the annotator agreement can be very low. The higher the number of considered emotions is, the more difficult it is for humans to agree on a label for a given text. Low quality labeling leads to difficulties in extracting powerful classification features. This problem is further compounded in parsimonious environments, like Twitter, where the short text leads to a lack of emotional cues. All this presents challenges in developing a high-quality emotion recognition system operating with a fine-grained emotion category set within a chosen domain.

In this paper, we show how to tackle the above challenges through human computation, using an online labor market such as the Amazon Mechanical Turk or AMT (Snow et al., 2008). To overcome the possible difficulties in annotation we employ a well-designed emotion assessment tool, the Geneva

[1] www.mturk.com

[2] The corpus and the lexicon are available upon email request

Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 12–20,
Atlanta, Georgia, 14 June 2013.

Emotion Wheel (GEW) (Scherer, 2005). Having 20 separate emotion categories, it provides a desirable high level of emotion granularity. In a given task, we show the annotators the tweets, related to the aforementioned sports event, and ask them to classify the tweets' emotional content into one of the provided emotion categories. The action sequence requires them to both label the tweets and to specify the textual constructs that support their decision. We view the selected textual constructs as probable classification features. The proposed method thus simultaneously produces *an emotion annotated corpus* of tweets and creates an *emotion lexicon*. The resulting weighted emotion lexicon is a list of phrases indicative of emotion presence. It consists solely of ones selected by respondents, while their weights were learnt based on their occurrence in the constructed Sports-Related Emotion Corpus (SREC).

We show that the human-based lexicon is well suited for the particularities of the chosen environment, and also for an emotion model with a high number of categories. Firstly, we show that domain specificity matters, and that non-specialists, using their common sense, can extract features that are useful in classification. We use the resulting lexicon, $OlympLex$, in a binary polarity classification problem on the domain data and show that it outperforms several traditional lexicons.

In multi-emotion classification, we show that it is highly accurate in classifying tweets into 20 emotion categories of the Geneva Emotion Wheel (GEW) (Scherer, 2005). As a baseline for comparison we use the Geneva wheel compatible lexicon, the Geneva Affect Label Coder (GALC) (Scherer, 2005). The experiments show that $OlympLex$ significantly outperforms this baseline.

Such a detailed emotion representation allows us to create an accurate description of the sentiment the chosen event evokes in its viewers. For instance, we find that *Pride* is the dominant emotion, and that it is 2.3 times more prevalent than *Anger*.

2 Related Work

GEW Emotion Representation Model In our work we used the emotion categories from the Geneva Emotion Wheel (GEW, version 2.0). GEW was developed as a tool for obtaining self-reports of emotional experience with a goal to structure the exhaustive list of possible emotion names used in free-format self-reports with minimal loss in expressibility. It presents 20 (10 positive/10 negative) emotion categories frequently answered in free-format self-reports as main options. Each emotion category is represented by two common emotion names to emphasize its family nature (e.g. *Happiness/Joy*[3]). These categories are arranged on the circle following the underlying 2-dimensional space of valence (positive-negative) and control (high-low). Several levels of intensity for each emotion category are presented as answer options. Also, 2 other answers are possible: *No emotion* and *Other emotion* with free-format input in the latter case.

Compared to raw dimensional models where emotion states are described as points in space (e.g. Pleasure-Arousal-Dominance model, PAD (Mehrabian, 1996)) GEW has an advantage of categorical representation where emotion state is described in terms of discrete set of emotion names. It allows humans to measure their emotions in terms of emotion names they accustomed to instead of unnatural numerical measurements. Among commonly used emotion categories sets GEW categories are the most fine-grained, compared, for instance, to Ekman's (1992) or Plutchik's (1980) basic emotions. While these models have been popular in emotion recognition research, their main shortcoming is their limited items. In sports events, fans and spectators not only feel strong emotions, but also likely want to express them in multitudes of expressions. *Pride/Elation*, *Envy/Jealousy* are just two examples that are missed in those models with basic emotions.

Lexical Resources Emotion recognition is closely related to the positive/negative sentiment classification. In a traditional approach the units defining the polarity of the text are polarity-bearing terms. A list of such terms with corresponding polarity label or score forms a polarity lexicon. Commonly used examples of polarity lexicons include GI (Stone et al., 1968), Bing Liu's lexicon (Hu and Liu, 2004), and OpinionFinder (Wilson et al., 2009).

Similarly, emotion lexicons can be defined as lists of terms bearing emotions with their corre-

[3]In the paper text we often use one name per category for brevity reasons

sponding emotion information. Depending on the construction methods, they can be separated into those that constructed manually (GALC (Scherer, 2005)), semi-automatically (WordNet-Affect (Strapparava and Valitutti, 2004)) or via human computation (ANEW (Bradley and Lang, 1999), NRC (Mohammad and Turney, 2010; Mohammad and Turney, 2012)). Our work is most closely related to the NRC lexicon which was also extracted via human computation on AMT. The authors developed a task where, for a given term, the annotators rated to what extent the term is associated to each emotion of Plutchik's set. In contrast, in our work, we harvest emotional labels and features in context. The terms are associated with emotions in the context of the tweet they appear in. We use the approach suggested by (Aman and Szpakowicz, 2007) where humans are asked to select an excerpt of the text expressing emotion. Moreover, we ask the annotators for additional interchangeable, emotional expressions for the same situation. Lexicons obtained from unsupervised learning methods using automatically annotated Twitter data (Mohammad, 2012) have also been proposed, but their performance has been shown to be inferior to benchmarks such as NRC.

The underlying emotion representation model differs from one emotion lexicon to another. For instance, ANEW uses the PAD dimensions, Plutchik's basic categories are used by NRC and Ekman's categories in WordNet-Affect. However, such representations do not provide a sufficient emotion granularity level. There is only one lexicon which incorporates GEW emotion model: the GALC (Scherer, 2005) lexicon. It contains 279 unigram stems (e.g. *happ**) explicitly expressing one of 36 emotion categories (covering all GEW categories). We use therefore this lexicon for benchmarking.

The main differences of our lexicon compared to its predecessors lie in the usage of new fine-grained emotion set, new methods of human computation employed in its construction and specificity to the context of Twitter posts and sport-related emotions.

3 Emotional Labeling and Emotion Feature Elicitation

We created a Human Computation method, using the online labor market (Amazon Mechanical Turk or AMT) to simultaneously accomplish two goals. The first is to have a reliable, human annotation of the emotions within a text corpus. The second is to enable the respondents to provide us with the features needed to construct an emotion lexicon. In this section we describe the processes of data selection, annotation, and refinement, as well as provide the statistical description of the obtained data.

3.1 Data Collection

Our goal is to analyze the emotions of the spectators of Olympic games. We consider the tweets about the Olympics posted during the 2012 Olympic games as a data source for this analysis. We assume that the same emotions are expressed in the same way for all the sports. We thus narrow the scope of our analysis to a single sport – gymnastics.

Traditionally, the gymnastics teams from the USA have strong bid for victory. Thus, we assume that a large group of English-speaking nation may be interested in it. Then, gymnastics is a dynamic type of sport where each moment of performance can play a crucial role in final results, enhancing the emotional experience in audience. Also, it is less common than, for instance, running or swimming, thus the occurrence of this term in tweets, at the time of the Olympics, will more likely signal a reference to the Olympic gymnasts.

We used the hashtag *#gymnastics* (hashtags represent topics in tweets) to obtain the tweets related to the gymnastic competitions during the Olympics time resulting in $199,730$ such tweets. An emotional example is *"Well done #gymnastics we have a SILVER yeayyyyyyyy!!!! Wohoooo"*.

3.2 Annotation Process

We developed a Human-Intelligence Task (HIT) on the AMT for annotation of a subset of the collected tweets with emotion-related information.

3.2.1 Task description

One HIT consisted of the annotation of one presented tweet. A worker was asked to read a tweet text and to fulfill the following subtasks:

Subtask 1 Decide on the dominant emotion the author of the tweet felt in the moment of its writing (***emotion label***) and how strong it was (***emotion strength***). Even though an emotion mixture could

Iteration		1	2 (B_{en})	2 (B_{all})	3	4	5	4+5
Polarity agreement		78.5	68	33.3	66.7	73.9	75.9	75.7
Emotion agreement		38.5	24.7	13.34	29.3	25.84	29.7	29.3
Average number of emotion	tweet	1.6	1.26	0.64	1.28	1.2	1.72	1.67
indicators per answer[a]	additional	-	0.25	0.36	1.41	1.3	2.05	1.99

Table 1: Basic statistics on the data collected over the annotation iterations.

[a] only among answers where non-neutral emotion label is assigned

be felt, a worker had to choose one emotion that prevailed all others. This kept him focused on one main emotion in the subtasks 2 and 3. To elicit this information we employed the Geneva Emotion Wheel (GEW) described in the Related Work with minor changes: we used 3 strength labels (low, medium and high) instead of 5 in initial version. The set of emotion categories remained unchanged: 20 GEW emotion categories plus 2 additional answer options: *No emotion* and *Other emotion*. We required workers to type the emotion name in latter case.

Subtask 2 In case an emotion was present, a worker was then asked to choose the excerpts of the tweet indicating its presence, the (***tweet emotion indicators***). She was asked to find all the expressions of the chosen emotion present in the tweet text. It could be one word, emoticon, or subsequence of the tweet words. We asked her to also include the words modifying the strength of emotion (e.g. to choose *so excited* instead of *excited*).

Subtask 3 Input ***additional emotion indicators*** of chosen emotion. Similarly to the previous subtask, a worker was asked to input the textual expressions of the chosen emotion. However, in this case the expressions had to be not from the tweet text, but generated based on personal experience. E.g. she could state that she uses *poor thing* to express *Pity*.

3.2.2 HIT Iterations

The design of annotation schema and corresponding instructions as well as search for the optimal HIT parameters took several iterations. Table 1 contains the statistics on inter-annotator agreements and on the number of provided emotion indicators for each iteration. Beside emotion agreement, we also consider polarity agreement. The ***polarity label*** of an answer is defined as the polarity of its emotion label. *No emotion* implies a *Neutral* polarity. For answers with *Other emotion* we manually detected their polarity based on provided emotion name if applicable, or set *Neutral* polarity otherwise.

Iteration 1 Firstly, we annotated 200 tweets (set $\mathcal{S}_1$), using respondents within our laboratory, into a set of 12 emotion categories ($SportEm$) which we considered first to be representative for the emotions incited by sport events: *Love, Pride, Excitement, Positive Surprise, Joy, Like, Other Positive, Anger/Hate, Shame, Anxiety, Shock, Sadness, Dislike, Other Negative*. For each tweet an annotator gave the *emotion label* and chose corresponding *tweet emotion indicators*. The tweets of $\mathcal{S}_1$ included both tweets with predefined emotional words and without. The details of selection process are omitted due to space limitations.

Iteration 2 We launched two batches of HITs on AMT: B_{all} and B_{en}. A HIT batch is defined by a set of tweets to label, with some parameters specific for AMT, such as the number of different workers for each tweet (we used 4 in all our experiments), the payment for one HIT, or specific worker requirements, (e.g. for B_{en} we also required that workers should be from the U.S.). We grouped 25 tweets from $\mathcal{S}_1$ with HIT payment of \$0.05 in B_{en}, whereas for B_{all} we included only 10 tweets with payment of \$0.03. The annotation schema used the emotions of $SportEm$. For each tweet an annotator gave the *emotion label* and provided *tweet* emotion indicators. The field for *additional* emotion indicators input was presented as optional.

We discovered that the answers in B_{all} had an unacceptable quality, with a low agreement and many impossible labels. This can be explained either by lower understanding of English or less reliability of workers from all around the world compared to the U.S. workers. Consequently, all our next iterations had the requirement on workers to be from the U.S.

Iteration 3 We launched a new HIT batch to an-

notate the full $\mathcal{S}_1$ with emotions from $SportEm$. Starting with this iteration, the payment was set to $\$0.04$. The *additional emotion indicators* field was shown as compulsory. The experiment showed that AMT workers generally followed the instructions achieving emotion agreement only slightly worse than ours.

Iteration 4 We decided to use the more fine-grained and well researched GEW emotion categories. Thus, we launched another HIT batch to annotate $\mathcal{S}_1$ again, in terms of GEW emotion categories (with a schema given in Task Description). Even though a new task contained more answer options emotion agreement stayed in the same range between 0.25 and 0.3.

Iteration 5 We launched a final batch with the described GEW schema to annotate more tweets. We selected Olympics related tweets that had a high likelihood of being emotional. We first selected tweets using the emotion indicators obtained during the previous iterations and found more than 5 times in the collected corpus (418 terms). For each keyword in this list we extracted up to 3 tweets containing this term (1244 tweets). In addition, we added the tweets without keywords from the list, but posted by the users who used these emotional keywords in their other tweets, supposing that these users are more likely to express their emotions. Overall, 1800 tweets were selected, but 13 were excluded because they were not written in English.

The resulting corpus contains the data gathered during the iterations 4 and 5. It consists of 1987 tweets annotated each by 4 workers with emotion label, emotion strength, and related emotion indicators. The Fleiss Kappa (Fleiss, 1971) for emotion labels is 0.24 which is considered to be fair by Landis and Koch (1977), but quite low compared to usual kappa values in other tasks (e.g. polarity annotation usually has Kappa in a range of 0.7–0.8). We conclude that the annotation in terms of multi-category emotions is highly subjective and ambiguous task, confirming our assumptions on existence of emotion mixtures.

3.3 Quality Control

The results of crowdsourcing usually require additional refinement. The workers who give malicious answers intentionally or due to lack of understanding worsen the data quality. We detect such workers automatically using the following 2 criteria:

Average Polarity Conformity A worker's answer has a *polarity conformity* of 1 if at least one worker indicated the same polarity for the same tweet (0 otherwise). A worker's average polarity conformity is computed from all his answers. This criterion aims to detect the workers who repeatedly disagree with other workers.

Dominant Emotion Frequency The dominant emotion of a worker is the one which appears most frequently in his answers. The dominant emotion frequency, among the worker's answers, is the criterion value. This criterion aims to detect workers biased towards specific emotion.

A worker who has the average polarity conformity below a predefined threshold or the dominant emotion frequency above a threshold is considered to have an insufficient quality and all his answers are excluded from the corpus. The threshold for each criterion is computed as a percentile of an approximated normal distribution of workers criterion values for probability limit of 0.01.

To increase the confidence in the computed criteria values, we establish a minimum number of tweets T_{min} any worker should annotate to be subjected to the criteria. To establish this number for each criterion, we use the following algorithm:

Let $X_n(w)$ be the criterion value computed using only first n answers of worker w in order of their submission. For each worker we detect $N_{min}(w)$ – the minimum number of answers after which the criterion value stops varying greatly:

$$|X_n(w) - X_{n-1}(w)| \leq 0.05,\ \forall n \geq N_{min}(w) \quad (1)$$

We then compute T_{min} as the ceiling of the average value of of $N_{min}(w)$ among workers who annotated at least 20 tweets.

The described procedure on detection of bad workers allowed the analysis of 83% of the answers. Using it, we excluded 8 workers, with their corresponding 260 answers.

In addition to removing these workers, we also excluded malicious answers: 736 answers that had a polarity conformity of 0. This additional filter was applied to all the remaining answers from the previous method. We also excluded the 121 answers with

Other emotion and the answers for 12 tweets, that were left with only 1 answer by this stage.

As a result of quality control, there were excluded 14.2% of initial answers. Overall, 1957 tweets with corresponding 6819 annotations remained (3.48 answers per tweet in average). These answers compose the final Sport-Related Emotion Corpus (SREC).

3.4 Emotion distribution in SREC

To provide a glimpse of the data we present the distribution of emotion categories among all answers in the figure 1. The most frequently answered emotion category was *Pride*, followed by *Involvement*. These emotions are natural in the context of sport events, however course-grained emotion models could not distinguish them. It highlights the advantage of fine-grained GEW emotion set to express the subtleties of the domain.

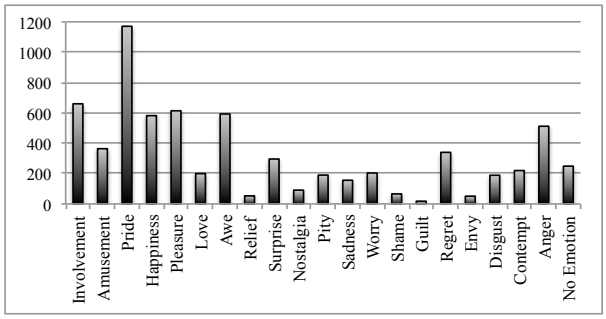

Figure 1: Distribution of emotion labels in worker's answers (after application of quality control)

4 Emotion Recognition Model

The output of our emotion recognition method is the distribution of emotions within a text, in terms of GEW emotion categories. It is represented as a tuple in the probability space

$$\mathbb{P} = \left\{ \bar{p} = (p_1, \ldots, p_{21}),\ \sum_{i=1}^{21} p_i = 1 \right\} \qquad (2)$$

where p_i represents the percentage of ith emotion in felt emotion mixture. The emotion set contains 20 GEW categories and *No Emotion* as 21st category.

We use a lexicon of **emotion indicators**, which are words or word sequences indicative of emotion presence. Each indicator $term_t$ has attached emotion distribution tuple $\bar{p}_t \in \mathbb{P}$. To compute the result tuple $\bar{p}$ for a text d we sum up all the tuples of emotion indicators found within this text with the number of times they were found:

$$\bar{p}(d) = \sum_{term_t \in d} n_t(d)\ \bar{p}_t \qquad (3)$$

If no indicators are present in the text, a full weight is given to *No emotion* category ($p_{21} = 1$). We also neglect all negated indicators occurrences detected by the negation words (*no*, *not*, **n't*, *never*) placed ahead of an indicator.

Lexicon Construction We construct the lexicon by selecting the emotion indicators and computing their emotion distributions. We use a training corpus that has a format described in the previous section. The training process consists of the following steps:

Among all *tweet* and *additional* emotion indicators provided by workers, we select those that were suggested more than once.

For each tweet we have several emotion labels from the data. We determine the emotion distribution of the tweet by computing the frequency of each emotion label over all the answers corresponding to that tweet.

For each answer we construct a link between each term suggested in the *additional* emotion indicators field and the answer's emotion label. This link is represented as a tuple $\bar{p} \in \mathbb{P}$ with weight 1 for linked emotion category. Then, for each detected emotion indicator we compute its emotion distribution by averaging all the emotion distributions it appeared in. This includes the emotion distributions of the tweets where this indicator occurred without a negation and the emotion distributions of the corresponding indicator-emotion links.

We define an indicator to be ambiguous if its dominant polarity (polarity having the highest sum of the weights for corresponding emotions) has summary weight smaller than 0.75. All such terms are removed from the result lexicon.

Result Lexicon Description Following the specified process over the full SREC data, we computed an emotion lexicon, $OlympLex$, that contains 3193 terms. The ratio of positive terms to negative ones is 7:3 (term polarity is defined as dominant polarity of term emotion distribution). Unigrams compose 37.5% of the lexicon, bigrams – 30.5%, all other terms are ngrams of a higher order (up to 5).

5 Experimental Evaluation

We evaluated our lexicon on the SREC corpus as a classifier, using ten-fold cross-validation to avoid possible overfitting. The precompiled universal lexicons were used for benchmarking. As no training is required, we tested them over the full data.

5.1 Polarity Classification

We considered the basic polarity classification task with 3 classes (*Positive*, *Negative* and *Neutral*). We used only 1826 tweets that have one dominant polarity based on workers' answers. This dominant polarity was taken as a true polarity label of a tweet.

The output polarity label of our classifier is dominant polarity of found emotion distribution: a polarity having the highest sum of the weights for corresponding emotions. The output of prior sentiment lexicons is computed analogously: we sum up the number of found lexicon terms in the tweet text for each emotion or polarity category (depending on which categorization is provided by the lexicon) and output the polarity having the highest sum value. If two polarities have the same sum weight, the output polarity is *Neutral*.

We used standard classification evaluation measures: accuracy, precision, recall and F1-score. We considered only non-neutral classes (*Positive* and *Negative*) for precision and recall. Table 2 shows the results of our classifier, compared with other known sentiment lexicons. The proposed lexicon outperforms every other one, both in terms of accuracy and F1-score. As it was the only lexicon fitted to the Olympic gymnastics data, its superiority reveals the advantage of domain-targeted lexicon construction.

Lexicon	P	R	F1	A
*OlympLex**	**81.7**	**73.2**	**77.2**	**72.5**
BingLiu	80.4	52.9	63.8	53.6
OpinionFinder	66.0	46.6	54.6	46.6
GeneralInquirer	69.8	44.4	54.3	44.5
NRC*	60.6	39.7	48.0	40.4
WnAffect*	78.6	28.1	41.4	30.1
GALC*	81.6	25.6	39.0	27.9

Table 2: The results of polarity classification evaluation. P=precision, R=recall, F1 = F1-score, A=accuracy
*A lexicon employing several emotion categories

5.2 Emotion Classification

We evaluated emotion recognition results in the setting of a multi-label classification problem. The output is a set of labels instead of a standard single label answer. In this case, the output of the classifier (O_C) was defined as a set of dominant emotions in the found emotion distribution $\bar{p}$. This set contained the emotions having the highest weights p_i. The set of emotion labels given for this tweet by workers formed a true output – a set of true labels (O_T) of emotion classification. As a baseline for multi-category emotion classification we considered the GALC lexicon (Scherer, 2005).

Multi-label Evaluation We used the standard evaluation metrics adapted for multi-label output (Tsoumakas and Katakis, 2007). For each tweet, we first computed the precision $P = \frac{|O_C \cap O_T|}{|O_C|}$, which shows how many of emotions outputted by the classifier were correct. Then the recall $R = \frac{|O_C \cap O_T|}{|O_T|}$, which shows how many of true labels were found by classifier, and the accuracy $A = \frac{|O_C \cap O_T|}{|O_C \cup O_T|}$, which shows how close the sets of classifier and true labels were. These values were averaged among all applicable tweets. For precision and recall we used only the tweets with non-neutral answers in O_C and O_T correspondingly (meaning that *No emotion* label was not present in a set).

Table 3 shows the comparative results of our and GALC lexicons. Compared to the GALC baseline, our classifier has both higher precision and recall. Higher recall is explained by the fact that our lexicon is larger and contains also ngram terms. In addition, it includes not only explicit emotion expressions (e.g. *sad* or *proud*), but also implicit ones (e.g. *yes* or *mistakes*).

Per-Category Evaluation Another way to evaluate the output of multi-label classifier is to evaluate it for each emotion category separately. For each category we computed precision, recall and F1-score.

Lexicon	P	R	F1	A
GALC	49.0	10.2	16.8	12.5
OlympLex	**53.5**	**24.9**	**34.0**	**25.4**

Table 3: Results of multi-label evaluation. P=precision, R=recall, F1 = F1-score, A=accuracy

	GALC			*OlympLex*				GALC			*OlympLex*		
Negative	P	R	F1	P	R	F1	Positive	P	R	F1	P	R	F1
Anger	48.4	10.8	17.7	53.3	26	**35**	Involvement	52.4	2.4	4.6	49.4	17.6	**26**
Contempt	-	0	-	42.1	4.7	**8.5**	Amusement	51	11.6	18.9	55	24.6	**34**
Disgust	50	1.4	2.8	39.4	9.4	**15.2**	Pride	89.6	6.7	12.5	60.8	59.4	**60.1**
Envy	100	11.1	20	55.6	13.9	**22.2**	Happiness	46.3	8.8	14.8	45.1	9.8	**16.1**
Regret	53.3	3.4	6.4	36.3	12.4	**18.5**	Pleasure	44.8	5.9	10.4	48.8	17.9	**26.2**
Guilt	25	5.6	**9.1**	0	0	-	Love	38.1	27.4	**31.9**	48.0	8.2	14
Shame	18.5	9.8	**12.8**	25	3.9	6.8	Awe	42.9	6.7	11.5	54.2	23.7	**33**
Worry	54.8	21.5	**30.9**	43.2	15	22.2	Relief	100	17.1	**29.2**	50	4.9	8.9
Sadness	52.5	19.6	**28.6**	41.7	9.3	15.3	Surprise	38.3	9	**14.6**	33.3	6	10.2
Pity	75	2.5	4.9	57.8	31.4	**40.7**	Nostalgia	20.5	14.5	**17**	28.6	3.2	5.8

Table 4: Evaluation results at per-category level. P=precision, R=recall, F1 = F1-score

The results of this evaluation in comparison with benchmark GALC lexicon are presented in the table 4. Overall, our lexicon performs better on most of the categories (12 out of 20) in terms of F1-score. The highest F1-score is achieved for such Olympic related emotion as *Pride*.

5.3 Discussion

The fact that the terms from the GALC lexicon are found in 31% of tweets indicates that people do express their emotions explicitly with emotional terms. However, a list of currently available explicit emotional terms is not extensive. For instance, it does not cover slang terms. Moreover, people do not limit themselves to only explicit emotional terms. Our lexicon constructed based on the answers provided by non-expert humans achieves a significantly higher recall. This highlights the importance of employing the human common knowledge in the process of extraction of emotion bearing features.

6 Conclusion

We presented a context-aware human computation method for emotion labeling and feature extraction. We showed that inexpert annotators, using their common sense, can successfully attach emotion labels to tweets, and also extract relevant emotional features. Using their answers, we carefully constructed a linguistic resource for emotion classification. The suggested method can be reused to construct additional lexicons for different domains.

An important aspect that differentiates our work is the emotion granularity. To the best of our knowledge, this was the first attempt to create lexical resources for emotion classification based on the Geneva Emotion Wheel (GEW), which has as many as 20 emotion categories. This level of granularity enabled us to capture the subtleties of the emotional responses in the target domain, tweets regarding the 2012 summer Olympics in London. In this dataset, we found that the prevalent emotion is *Pride*, a detail which is unattainable using previous methods.

Another differentiator is that, unlike most previous approaches, we relied on human computation for both labeling and feature extraction tasks. We showed that human generated features can be successfully used in emotional classification, outperforming various existing methods. A further difference from prior lexicons is the fact that ours was built with a context-sensitive method. This led to a higher accuracy on the target domain, compared to the general purpose lexicon.

We benchmarked the cross-validated version of created $OlympLex$ lexicon with the existing universal-domain lexicons for both polarity and multi-emotion problems. In suggested settings we showed that it can outperform general purpose lexicons in the binary classification due to its domain specificity. We also obtained significant improvements over the baseline GALC lexicon, which was the only preexisting one compatible with the GEW.

However, high domain specificity of the created lexicon and restricted variety of data used in its construction implies possible limitations of its usage for other types of data. Its porting and generalization to other domains is one of the future directions.

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*, pages 579–586. Association for Computational Linguistics.

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

Margaret M Bradley and Peter J Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

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

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

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

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A Calvo. 2010. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 62–70. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Hugo Liu, Henry Lieberman, and Ted Selker. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 125–132. ACM.

Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292.

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34.

Saif M Mohammad and Peter D Turney. 2012. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*.

Saif M Mohammad. 2012. # emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics.

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):95.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, 1(3):3–33.

Klaus R Scherer. 2005. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics.

Phil J Stone, Dexter C Dunphy, Marshall S Smith, and DM Ogilvie. 1968. The general inquirer: A computer approach to content analysis. *Journal of Regional Science*, 8(1).

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet-affect: an affective extension of wordnet. In *Proceedings of LREC*, volume 4, pages 1083–1086.

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

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433.