IEST: WASSA-2018 Implicit Emotions Shared Task

Roman Klinger\textsuperscript{1}, Orphée De Clercq\textsuperscript{2}, Saif M. Mohammad\textsuperscript{3}, and Alexandra Balahur\textsuperscript{4}

\textsuperscript{1} Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany
klinger@ims.uni-stuttgart.de

\textsuperscript{2} LT3, Language and Translation Technology Team, Ghent University, Belgium
orphee.declercq@ugent.be

\textsuperscript{3} National Research Council Canada, Ottawa, Ontario, Canada
saif.mohammad@nrc-cnrc.gc.ca

\textsuperscript{4} Text and Data Mining Unit, European Commission Joint Research Centre, Ispra, Italy
alexandra.balahur@ec.europa.eu

Abstract

Past shared tasks on emotions use data with both overt expressions of emotions (I \textit{am so happy to see you!}) as well as subtle expressions where the emotions have to be inferred, for instance from event descriptions. Further, most datasets do not focus on the cause or the stimulus of the emotion. Here, for the first time, we propose a shared task where systems have to predict the emotions in a large automatically labeled dataset of tweets without access to words denoting emotions. Based on this intention, we call this the Implicit Emotion Shared Task (IEST) because the systems have to infer the emotion mostly from the context. Every tweet has an occurrence of an explicit emotion word that is masked. The tweets are collected in a manner such that they are likely to include a description of the cause of the emotion – the stimulus. Altogether, 30 teams submitted results which range from macro F\textsubscript{1} scores of 21\% to 71\%. The baseline (MaxEnt bag of words and bigrams) obtains an F\textsubscript{1} score of 60\% which was available to the participants during the development phase. A study with human annotators suggests that automatic methods outperform human predictions, possibly by honing into subtle textual clues not used by humans. Corpora, resources, and results are available at the shared task website at \url{http://implicitemotions.wassa2018.com}.

1 Introduction

The definition of emotion has long been debated. The main subjects of discussion are the origin of the emotion (physiological or cognitive), the components it has (cognition, feeling, behaviour) and the manner in which it can be measured (categorically or with continuous dimensions). The Implicit Emotion Shared Task (IEST) is based on Scherer (2005), who considers emotion as “an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems (information processing, support, executive, action, monitor) in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism”.

This definition suggests that emotion is triggered by the interpretation of a stimulus event (i.e., a situation) according to its meaning, the criteria of relevance to the personal goals, needs, values and the capacity to react. As such, while most situations will trigger the same emotional reaction in most people, there are situations that may trigger different affective responses in different people. This is explained more in detail by the psychological theories of emotion known as the “appraisal theories” (Scherer, 2005).

Emotion recognition from text is a research area in natural language processing (NLP) concerned with the classification of words, phrases, or documents into predefined emotion categories or dimensions. Most research focuses on discrete emotion recognition, which assigns categorical emotion labels (Ekman, 1992; Plutchik, 2001), e.g., \textit{Anger}, \textit{Anticipation}, \textit{Disgust}, \textit{Fear}, \textit{Joy}, \textit{Sadness}, \textit{Surprise} and \textit{Trust}.\footnote{Some shared tasks on fine emotion intensity include the SemEval-2007 Task 14, WASSA-2017 shared task EmoInt (Mohammad and Bravo-Marquez, 2017), and SemEval-2018 Task 1 (Mohammad et al., 2018).} Previous research developed statistical, dictionary, and rule-based models for...
several domains, including fairy tales (Alm et al., 2005), blogs (Aman and Szpakowicz, 2007) and microblogs (Dodds et al., 2011). Presumably, most models built on such datasets rely on emotion words (or their representations) whenever accessible and are therefore not forced to learn associations for more subtle descriptions. Such models might fail to predict the correct emotion when such overt words are not accessible. Consider the instance “when my child was born” from the ISEAR corpus, a resource in which people have been asked to report on events when they felt a specific predefined emotion. This example does not contain any emotion word itself, though one might argue that the words “child” and “born” have a positive prior connotation.

Balahur et al. (2012b) showed that the inference of affect from text often results from the interpretation of the situation presented therein. Therefore, specific approaches have to be designed to understand the emotion that is generally triggered by situations. Such approaches require common sense and world knowledge (Liu et al., 2003; Cambria et al., 2009). Gathering world knowledge to support NLP is challenging, although different resources have been built to this aim – e. g., Cyc and ConceptNet (Liu and Singh, 2004).

On a different research branch, the field of distant supervision and weak supervision addresses the challenge that manually annotating data is tedious and expensive. Distant supervision tackles this by making use of structured resources to automatically label data (Mintz et al., 2009; Riedel et al., 2010; Mohammad, 2012). This approach has been adapted in emotion analysis by using information assigned by authors to their own text, with the use of hashtags and emoticons (Wang et al., 2012).

With the Implicit Emotion Shared Task (IEST), we aim at combining these two research branches: On the one hand, we use distant supervision to compile a corpus of substantial size. On the other hand, we limit the corpus to those texts which are likely to contain descriptions of the cause of the emotion – the stimulus. Due to the ease of access and the variability and data richness on Twitter, we opt for compiling a corpus of microposts, from which we sample tweets that contain an emotion word followed by ‘that’, ‘when’, or ‘because’. We then mask the emotion word and ask systems to predict the emotion category associated with that word. The emotion category can be one of six classes: anger, disgust, fear, joy, sadness, and surprise. Examples from the data are:

(1) “It’s [#TARGETWORD#] when you feel like you are invisible to others.”
(2) “My step mom got so [#TARGETWORD#] when she came home from work and saw that the boys didn’t come to Austin with me.”

In Example 1, the inference is that feeling invisible typically makes us sad. In Example 2, the context is presumably that a person (mom) expected something else than what was expected. This in isolation might cause anger or sadness, however, since “the boys are home” the mother is likely happy. Note that such examples can be used as source of commonsense or world knowledge to detect emotions from contexts where the emotion is not explicitly implied.

The shared task was conducted between 15 March 2018 (publication of train and trial data) and the evaluation phase, which ran from 2 to 9 July. Submissions were managed on CodaLab. The best performing systems are all ensembles of deep learning approaches. Several systems make use of external additional resources such as pre-trained word vectors, affect lexicons, and language models fine-tuned to the task.

The rest of the paper is organized as follows: we first review related work (Section 2). Section 3 introduces the shared task, the data used, and the setup. The results are presented in Section 4, including the official results and a discussion of different submissions. The automatic system’s predictions are then compared to human performance in Section 5, where we report on a crowdsourcing study with the data used for the shared task. We conclude in Section 6.

2 Related Work

Related work is found in different directions of research on emotion detection in NLP: resource creation and emotion classification, as well as modeling the emotion itself.

Modeling the emotion computationally has been approached from the perspective of humans needs

\[^2\text{http://www.cyc.com}\]

\[^3\text{https://competitions.codalab.org/competitions/19214}\]
and desires with the goal of simulating human reactions. Dyer (1987) presents three models which take into account characters, arguments, emotion experiencers, and events. These aspects are modeled with first order logic in a procedural manner. Similarly, Subasic and Huettner (2001) use fuzzy logic for such modeling in order to consider gradual differences. A similar approach is followed by the OCC model (Ortony et al., 1990), for which Udochukwu and He (2015) show how to connect it to text in a rule-based manner for implicit emotion detection. Despite of this early work on holistic computational models of emotions, NLP focused mostly on a more coarse-grained level.

One of the first corpora annotated for emotions is that by Alm et al. (2005) who analyze sentences from fairy tales. Strapparava and Mihalcea (2007) annotate news headlines with emotions and valence, Mohammad et al. (2015) annotate tweets on elections, and Schuff et al. (2017) tweets of a stance dataset (Mohammad et al., 2017). The SemEval-2018 Task 1: Affect in Tweets (Mohammad et al., 2018) includes several subtasks on inferring the affectual state of a person from their tweet: emotion intensity regression, emotion intensity ordinal classification, valence (sentiment) regression, valence ordinal classification, and multi-label emotion classification. In all of these prior shared tasks and datasets, no distinction is made between implicit or explicit mentions of the emotions. We refer the reader to Bostan and Klinger (2018) for a more detailed overview of emotion classification datasets.

Few authors specifically analyze which phrase triggers the perception of an emotion. Aman and Szpakowicz (2007) focus on the annotation on document level but also mark emotion indicators. Mohammad et al. (2014) annotate electoral tweets for semantic roles such as emotion and stimulus (from FrameNet). Ghazi et al. (2015) annotate a subset of Aman and Szpakowicz (2007) with causes (inspired by the FrameNet structure). Kim and Klinger (2018) and Neviarouskaya and Aono (2013) similarly annotate emotion holders, targets, and causes as well as the trigger words.

One of the oldest resources nowadays used for emotion recognition is the ISEAR set (Scherer, 1997) which consists of self-reports of emotional events. As the task of participants in a psychological study was not to express an emotion but to report on an event in which they experienced a given emotion, this resource can be considered similar to our goal of focusing on implicit emotion expressions.

With the aim to extend the coverage of ISEAR, Balahur et al. (2011, 2012a) build EmotiNet, a knowledge base to store situations and the affective reactions they have the potential to trigger. They show how the knowledge stored can be expanded using lexical and semantic similarity, as well as through the use of Web-extracted knowledge (Balahur et al., 2013). The patterns used to populate the database are of the type ‘I feel [emotion] when [situation]’, which was also a starting point for our task.

Finally, several approaches take into consideration distant supervision (Mohammad and Kiritchenko, 2015; Abdul-Mageed and Ungar, 2017; De Choudhury et al., 2012; Liu et al., 2017, i. a.). This is motivated by the high availability of user-generated text and by the challenge that manual annotation is typically tedious or expensive. This contrasts with the current data demand of machine learning, and especially, deep learning approaches.

With our work in IEST, we combine the goal of the development of models which are able to recognize emotions from implicit descriptions without having access to explicit emotion words, with the paradigm of distant supervision.

3 Shared Task

3.1 Data

The aim of the Implicit Emotion Shared Task is to force models to infer emotions from the context of emotion words without having access to them. Specifically, the aim is that models infer the emotion through the causes mentioned in the text. Thus, we build the corpus of Twitter posts by polling the Twitter API\(^1\) for the expression ‘EMOTION-WORD (that|because|when)’, where EMOTION-WORD contains a synonym for one out of six emotions.\(^2\) The synonyms are shown in Table 1. The requirement of tweets to have either ‘that’, ‘because’, or ‘when’ immediately after the emotion word means that the tweet likely describes the cause of the emotion.

---
\(^1\)https://developer.twitter.com/en/docs.html
\(^2\)Note that we do not check that there is a white space before the emotion word, which leads to tweets containing ‘. . . unEMOTION-word. . . ’.
## Emotion Abbr. Synonyms

| Emotion | Train | Trial | Test |
|---------|-------|-------|------|
| Anger   | A     | 25562 | 1600 | 4794 |
| Fear    | F     | 25558 | 1597 | 4794 |
| Disgust | D     | 25575 | 1598 | 4791 |
| Joy     | J     | 27958 | 1736 | 5246 |
| Sadness | Sa    | 23165 | 1460 | 4340 |
| Surprise| Su    | 25565 | 1600 | 4792 |
| Sum     |       | 153383| 9591 | 28757|

Table 2: Distribution of IEST data.

The initially retrieved large dataset has a distribution of 25% surprise, 23% sadness, 18% joy, 16% fear, 10% anger, 8% disgust. We discard tweets with more than one emotion word, as well as exact duplicates, and mask usernames and URLs. From this set, we randomly sample 80% of the tweets to form the training set (153,600 instances), 5% as trial set (9,600 instances), and 15% as test set (28,800 instances). We perform stratified sampling to obtain a balanced dataset. While the shared task took place, two errors in the data preprocessing were discovered by participants (the use of the word unhappy as synonym for sadness, which lead to inconsistent preprocessing in the context of negated expressions, and the occurrence of instances without emotion words). To keep the change of the data at a minimum, the erroneous instances were only removed, which leads to a distribution of the data as shown in Table 2.

## 3.2 Task Setup

The shared task was announced through a dedicated website (http://implicitemotions.wassa2018.com/) and computational-linguistics-specific mailing lists. The organizers published an evaluation script which calculates precision, recall, and F1 measure for each emotion class as well as micro and macro average. Due to the nearly balanced dataset, the chosen official metric for ranking submitted systems is the macro-F1 measure.

In addition to the data, the participants were provided a list of resources they might want to use and they were allowed to use any other resources they have access to or create themselves. We also provided access to a baseline system. This baseline is a maximum entropy classifier with L2 regularization. Strings which match \[#a-zA-Z0-9_\=]+|[^ ]\] form tokens. As preprocessing, all symbols which are not alphanumeric or contain the # sign are removed. Based on that, unigrams and bigrams form the Boolean features as a set of words for the classifier.

## 4 Results

### 4.1 Baseline

The intention of the baseline implementation was to provide participants with an intuition of the difficulty of the task. It reaches 59.88% macro F1 on the test data, which is very similar to the trial data result (60.1% F1). The confusion matrix for the baseline is presented in Table 3; the confusion matrix for the best submitted system is shown in Table 4.

### 4.2 Submission Results

Table 5 shows the main results of the shared task. We received submissions through CodaLab from thirty participants. Twenty-six teams responded to a post-competition survey providing additional information regarding team members (56 people in total).
Table 5: Official results of IEST 2018. Participants who did not report on the system details did not get assigned a rank and are reported in gray. Column B provides the first row in the results table to which the respective row is significantly different (confidence level 0.99), tested with bootstrap resampling.

| id | Team                | $F_1$ | Rank | B |
|----|---------------------|-------|------|---|
| 1  | Amobee              | 71.45 | 1    | (1) | 3 |
| 2  | IIDYT               | 71.05 | 2    |   | 3 |
| 3  | NTUA-SLP            | 70.29 | 3    |   | 4 |
| 4  | UBC-NLP             | 69.28 | 4    |   | 6 |
| 5  | Sentylic            | 69.20 | 5    |   | 7 |
| 6  | HUMIR               | 68.64 | 6    |   | 8 |
| 7  | nlp                 | 68.48 | 7    |   | 9 |
| 8  | DataSEARCH          | 68.04 | 8    |   | 10 |
| 9  | YNU1510             | 67.63 | 9    |   | 11 |
| 10 | EmotiKLUE           | 67.13 | 10   |   | 11 |
| 11 | wojtek.pierre       | 66.15 | 11   |   | 15 |
| 12 | hgsgnlp             | 65.80 | 12   |   | 15 |
| 13 | UWB                 | 65.70 | 13   |   | 15 |
| 14 | NL-FIT              | 65.52 | 14   |   | 15 |
| 15 | TubOslo             | 64.63 | 15   |   | 17 |
| 16 | YNU,lab             | 64.10 | 16   |   | 17 |
| 17 | Braint              | 62.61 | 17   |   | 19 |
| 18 | EmoNLP              | 62.11 | 18   |   | 19 |
| 19 | RW                  | 60.97 | 19   |   | 20 |
| 20 | Baseline            | 59.88 | 20   |   | 21 |
| 21 | UST-IR              | 58.37 | 21   |   | 22 |
| 22 | TH1LNGN             | 58.01 | 22   |   | 23 |
| 23 | SINAI               | 57.94 | 23   |   | 24 |
| 24 | UTPPR               | 56.92 | 24   |   | 26 |
| 25 | CNHZ2017            | 56.40 | 25   |   | 27 |
| 26 | by3b                | 55.87 | 26   |   | 27 |
| 27 | Adobe Research      | 53.08 | 27   |   | 28 |
| 28 | Anonymous           | 50.38 | 28   |   | 29 |
| 29 | dinel               | 49.99 | 29   |   | 30 |
| 30 | CHANDA              | 41.89 | 30   |   | 31 |
| 31 | NLP_LDW             | 31.03 | 31   |   | 32 |

Note that the macro $F_1$ evaluation took into account all classes which were either predicted or in the gold data. Two teams submitted results which contain labels not present in the gold data, which reduced the macro-$F_1$ dramatically. With an evaluation only taking into account 6 labels, id 22 would be on rank 9 and id 28 would be on rank 10.

4.3 Review of Methods

Table 6 shows that many participants use high-level libraries like Keras or NLTK. Tensorflow is only of medium popularity and Theano is only used by one participant. Table 7 shows a summary of machine learning methods used by the teams, as reported by themselves. Nearly every team uses embeddings and neural networks; many teams use an ensemble of architectures. Several teams use language models showing a current trend in NLP to fine-tune those to specific tasks (Howard and Ruder, 2018). Presumably, those are specifically helpful in our task due to its word-prediction aspect.

Finally, Table 8 summarizes the different kinds of information sources taken into account by the teams. Several teams use affect lexicons in addition to word information and emoji-specific information. The incorporation of statistical knowledge from unlabeled corpora is also popular.


4.4 Top 3 Submissions

In the following, we briefly summarize the approaches used by the top three teams: Amobee, IIIDYT, and NTUA-SLP. For more information on these approaches and those of the other teams, we refer the reader to the individual system description papers. The three best performing systems are all ensemble approaches. However, they make use of different underlying machine learning architectures and rely on different kinds of information.

4.4.1 Amobee

The top-ranking system, Amobee, is an ensemble approach of several models (Rozental et al., 2018). First, the team trains a Twitter-specific language model based on the transformer decoder architecture using 5B tweets as training data. This model is used to find the probabilities of potential missing words, conditional upon the missing word describing one of the six emotions. Next, the team applies transfer learning from the trained models they developed for SemEval 2018 Task 1: Affect in Tweets (Rozental and Fleischer, 2018). Finally, they directly train on the data provided in the shared task while incorporating outputs from DeepRankWordsLexiconsCharactersEmojiUnlabeled CorporaEmotion Emb.Sentence/DocumentSemEval

Table 7: Overview of methods employed by different teams (sorted by popularity from left to right).

Moji (Felbo et al., 2017) and “Universal Sentence Encoder” (Cer et al., 2018) as features.

4.4.2 IIIDYT

The second-ranking system, IIIDYT (Balazs et al., 2018), preprocesses the dataset by tokenizing the sentences (including emojis), and normalizing the USERNAME, NEWLINE, URL and TRIGGER-WORD indicators. Then, it feeds word-level representations returned by a pretrained ELMo layer into a Bi-LSTM with 1 layer of 2048 hidden units for each direction. The Bi-LSTM output word representations are max-pooled to generate sentence-level representations, followed by a single hidden layer of 512 units and output size of 6. The team trains six models with different random initializations, obtains the probability distributions for each example, and then averages these to obtain the final label prediction.

4.4.3 NTUA-SLP

The NTUA-SLP system (Chronopoulou et al., 2018) is an ensemble of three different generic models. For the first model, the team pretrains Twitter embeddings with the word2vec skip-gram
model using a large Twitter corpus. Then, these pretrained embeddings are fed to a neural classifier with 2 layers, each consisting of 400 bi-LSTM units with attention. For the second model, they use transfer learning of a pretrained classifier on a 3-class sentiment classification task (Semeval17 Task4A) and then apply fine-tuning to the IEST dataset. Finally, for the third model the team uses transfer learning of a pretrained language model, according to Howard and Ruder (2018). They first train 3 language models on 3 different Twitter corpora (2M, 3M, 5M) and then they fine-tune them to the IEST dataset with gradual unfreezing.

4.5 Error Analysis

Table 11 in the Appendix shows a subsample of instances which are predicted correctly by all teams (marked as +, including the baseline system and those who did not report on system details) and that were not predicted correctly by any team (marked as −), separated by correct emotion label.

For the positive examples which are correctly predicted by all teams, specific patterns reoccur. For anger, the author of the first example encourages the reader not to be afraid – a prompt which might be less likely for other emotions. For several emotions, single words or phrases are presumably associated with such emotions, e.g., “hungry” with anger, “underwear”, “sweat”, “ewww” with disgust, “leaving”, “depression” for sadness, “why am i not” for surprise.

Several examples which are all correctly predicted by all teams for joy include the syllable “un” preceding the triggerword – a pattern more frequent for this emotion than for others. Another pattern is the phrase “fast and furious” (with furious for anger) which should be considered a mistake in the sampling procedure, as it refers to a movie instead of an emotion expression.

Negative examples appear to be reasonable when the emotion is given but may also be valid with other labels than the gold. For disgust, respective emotion synonyms are often used as a strong expression actually referring to other negative emotions. Especially for sadness, the negative examples include comparably long event descriptions.

5 Comparison to Human Performance

An interesting research question is how accurately native speakers of a language can predict the emotion class when the emotion word is removed from a tweet. Thus we conducted a crowdsourced study asking humans to perform the same task as proposed for automatic systems in this shared task.

We sampled 900 instances from the IEST data: 50 tweets for each of the six emotions in 18 pair-wise combinations with ‘because’, ‘that’, and ‘when’. The tweets and annotation questionnaires were uploaded on a crowdsourcing platform, Figure Eight (earlier called CrowdFlower). The questionnaire asked for the best guess for the emotion (Q1) as well as any other emotion that they think might apply (Q2).

About 5% of the tweets were annotated internally beforehand for Q1 (by one of the authors of this paper). These tweets are referred to as gold tweets. The gold tweets were interspersed with other tweets. If a crowd-worker got a gold tweet question wrong, they were immediately notified of the error. If the worker’s accuracy on the gold tweet questions fell below 70%, they were refused further annotation, and all of their annotations were discarded. This served as a mechanism to avoid malicious annotations.

Each tweet is annotated by at least three people. A total of 3,619 human judgments of emotion associated with the trigger word were obtained. Each judgment included the best guess for the emotion (response to Q1) as well as any other emotion that they think might apply (response to Q2). The answer to Q1 corresponds to the shared task setting. However, automatic systems were not given the option of providing additional emotions that might apply (Q2).

The macro $F_1$ for predicting the emotion is 45% (Q1, micro $F_1$ of 0.47). Observe that human performance is lower than what automatic systems reach in the shared task. The correct emotion was present in the top two guessed emotions in 57% of the cases. Perhaps, the automatic systems are honing in to some subtle systematic regularities in hope

| Gold Labels | A | D | F | J | Sa | Su |
|-------------|---|---|---|---|----|----|
| A           | 349 | 40 | 34 | 55 | 95 | 43 |
| D           | 195 | 92 | 30 | 84 | 157 | 69 |
| F           | 94 | 20 | 265 | 92 | 120 | 42 |
| J           | 39 | 6 | 22 | 398 | 36 | 13 |
| Sa          | 88 | 37 | 23 | 89 | 401 | 46 |
| Su          | 123 | 25 | 29 | 132 | 53 | 183 |

Table 9: Confusion Matrix Sample Annotated by Humans in Crowdsourcing

[9]https://www.figure-eight.com
that particular emotion words are used (for example, the function words in the immediate neighborhood of the target word). It should also be noted, however, that the data used for human annotations was only a subsample of the IEST data.

An analysis of subsets of Tweets containing the words because, that, and when after the emotion word shows that Tweets with “that” are more difficult (41% accuracy) than with “when” (49%) and “because” (51%). This relationship between performance and query string is not observed in the baseline system – here, accuracy on the test data (on the data used for human evaluation) for the “that” subset is 61% (60%), for “when” 62% (53%), and for “because” 55% (50%) — therefore, the automatic system is most challenged by “because”, while humans are more challenged by “that”. Please note that this comparison on the test data is somewhat unfair since for the human analysis, the data was sampled in a stratified manner, but not for the automatic prediction. The test data contains 5635 “because” tweets, 13649 with “that” and 9474 with “when”.

There are differences in the difficulty of the task for different emotions: The accuracy (F1) by emotion is 57% (46%) for anger, 15% (21%) for disgust, 42% (51%) for fear, 77% (58%) for joy, 59% (52%) for sadness and 34% (39%) for surprise. The confusion matrix is depicted in Table 9. Disgust is often confused with anger, followed by fear being confused with sadness. Surprise is often confused with anger and joy.

6 Conclusions & Future Work

With this paper and the Implicit Emotion Shared Task, we presented the first dataset and joint effort to focus on causal descriptions to infer emotions that are triggered by specific life situations on a large scale. A substantial number of participating systems presented the current state of the art in text classification in general and transferred it to the task of emotion classification.

Based on the experiences during the organization and preparation of this shared task, we plan the following steps for a potential second iteration. The dataset was now constructed via distant supervision, which might be a cause for inconsistencies in the dataset. We plan to use crowdsourcing as applied for the estimation of human performance to improve preprocessing of the data. In addition, as one participant noted, the emotion words which were used to retrieve the data were removed, but, in a subset of the data, other emotion words were retained.

The next step, which we suggest to the participants and future researchers is introspection of the models – carefully analyse them to prove that the models actually learn to infer emotions from subtle descriptions of situations, instead of purely associating emotion words with emotion labels. Similarly, an open research question is how models developed on the IEST data perform on other data sets. Bostan and Klinger (2018) showed that transferring models from one corpus to another in emotion analysis leads to drops in performance. Therefore, an interesting option is to use transfer learning from established corpora (which do not distinguish explicit and implicit emotion statements) to the IEST data and compare the models to those directly trained on the IEST and vice versa.

Finally, another line of future research is the application of the knowledge inferred to other tasks, such as argument mining and sentiment analysis.

Acknowledgments

This work has been partially supported by the German Research Council (DFG), project SEAT (Structured Multi-Domain Emotion Analysis from Text, KL 2869/1-1). We thank Evgeny Kim, Laura Bostan, Jeremy Barnes, and Veronique Hoste for fruitful discussions.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

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

Alexandra Balahur, Jesus Hermida, and Andres Montoyo. 2012a. Building and exploiting emotinet, a knowledge base for emotion detection based on the
appraisal theory model. *IEEE Transactions on Affective Computing*, 3:88–101.

Alexandra Balahur, Jesús M. Hermida, Andrés Montoyo, and Rafael Muñoz. 2011. EmotiNet: A knowledge base for emotion detection in text built on the appraisal theories. In *Natural Language Processing and Information Systems*, pages 27–39, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alexandra Balahur, Jesús M. Hermida, and Hristo Tanev. 2013. Detecting implicit emotion expressions from text using ontological resources and lexical learning. In *New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems*, pages 235–255, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alexandra Balahur, Jess Hermida, and Andrés Montoyo. 2012b. Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*, 53(4):742753.

Jorge A. Balazs, Edison Marrese-Taylor, and Yutaka Matsuo. 2018. IIIIDYT at IEST 2018: Implicit Emotion Classification with Deep Contextualized Word Representations. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Brussels, Belgium. Association for Computational Linguistics.

Laura Ana Maria Bostan and Roman Klinger. 2018. A survey on annotated data sets for emotion classification in text. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*, Santa Fe, USA.

Erik Cambria, Amir Hussain, Catherine Havasi, and Chris Eckl. 2009. Affectivespace: Blending common sense and affective knowledge to perform emotive reasoning. In *WOMSA09: 1st Workshop on Opinion Mining and Sentiment Analysis. WOMSA'09: 1st Workshop on Opinion Mining and Sentiment Analysis*, pages 32–41, Seville, Spain.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhammi St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Alexandra Chronopoulou, Aikaterini Margatina, Christos Baziotis, and Alexandros Potamianos. 2018. NTUA-SLP at IEST 2018: Ensemble of neural transfer methods for implicit emotion classification. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Brussels, Belgium. Association for Computational Linguistics.

Munmun De Choudhury, Scott Counts, and Michael Gamon. 2012. Not all moods are created equal! exploring human emotional states in social media. In *Sixth international AAAI conference on weblogs and social media*, pages 66–73.

Peter S. Dodds, Kameron D. Harris, Isabel M. Kloumman, Catherine A. Bliss, and Christopher M. Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *Plos one*, 6(12).

Michael G. Dyer. 1987. Emotions and their computations: Three computer models. *Cognition and Emotion*, 1(3):323–347.

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

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *Computational Linguistics and Intelligent Text Processing*, pages 152–165, Cham. Springer International Publishing.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*, Santa Fe, USA.

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, IUI ’03*, pages 125–132, New York, NY, USA. ACM.

Hugo Liu and Push Singh. 2004. Conceptnet – a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Vicki Liu, Carmen Banea, and Rada Mihalcea. 2017. Grounded emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 477–483, San Antonio, Texas.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Conference of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*. 
Saif Mohammad. 2012. #emotional tweets. In *SEM 2012: 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 (SemEval 2012), pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Saif M. Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA), Copenhagen, Denmark.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in tweets. In Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (SemEval-2018), New Orleans, LA, USA.

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

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3):26:1–26:23.

Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2014. Semantic role labeling of emotions in tweets. In Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 32–41, Baltimore, Maryland. Association for Computational Linguistics.

Alena Neviarouskaya and Masaki Aono. 2013. Extracting causes of emotions from text. In Proceedings of the Sixth International Joint Conference on Natural Language Processing, pages 932–936.

Andrew Ortony, Gerald L. Clore, and Allan Collins. 1990. The cognitive structure of emotions. Cambridge University Press.

Robert Plutchik. 2001. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist, 89*(4):344–350.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling Relations and Their Mentions without Labeled Text. In Proceedings of the European Conference on Machine Learning and Principles and Practice in Knowledge Discovery from Databases.
## Results by emotion class

Table 10 shows breakdown of the results by emotion class.

| Team                | Joy   | Sadness | Disgust | Anger | Surprise | Fear  |
|---------------------|-------|---------|---------|-------|----------|-------|
|                     | P | R | F₁ | P | R | F₁ | P | R | F₁ | P | R | F₁ | P | R | F₁ |
| Amobee              | 82 | 82 | 82 | 70 | 68 | 69 | 73 | 70 | 72 | 62 | 66 | 64 | 66 | 70 | 68 |
| IIIDYT              | 79 | 81 | 80 | 71 | 67 | 69 | 70 | 71 | 71 | 66 | 63 | 64 | 66 | 71 | 68 |
| NTUA-SLP            | 81 | 77 | 79 | 71 | 66 | 69 | 72 | 70 | 71 | 63 | 64 | 63 | 62 | 71 | 67 |
| UBC-NLP             | 79 | 79 | 79 | 67 | 67 | 67 | 69 | 68 | 69 | 62 | 63 | 62 | 65 | 67 | 66 |
| SentiLyic           | 80 | 77 | 79 | 68 | 66 | 67 | 69 | 69 | 69 | 63 | 61 | 62 | 63 | 69 | 66 |
| HUMIR               | 77 | 78 | 78 | 70 | 64 | 66 | 70 | 68 | 69 | 61 | 63 | 62 | 61 | 69 | 65 |
| nlp                 | 77 | 78 | 78 | 68 | 62 | 65 | 70 | 67 | 69 | 62 | 63 | 62 | 62 | 68 | 65 |
| DataSEARCH          | 77 | 77 | 77 | 66 | 64 | 65 | 69 | 68 | 69 | 61 | 62 | 62 | 65 | 64 | 65 |
| YNU1510             | 78 | 75 | 76 | 64 | 64 | 64 | 68 | 68 | 68 | 60 | 63 | 62 | 64 | 65 | 64 |
| EmotiKLUE           | 77 | 78 | 77 | 69 | 59 | 64 | 67 | 67 | 67 | 60 | 61 | 60 | 60 | 68 | 64 |
| wojtek.pierre       | 77 | 75 | 76 | 67 | 61 | 64 | 66 | 68 | 67 | 57 | 60 | 58 | 62 | 63 | 62 |
| hgsgnlp             | 75 | 75 | 75 | 66 | 59 | 62 | 67 | 66 | 67 | 59 | 59 | 59 | 59 | 67 | 63 |
| UWB                 | 74 | 77 | 77 | 61 | 68 | 64 | 74 | 59 | 65 | 57 | 63 | 60 | 66 | 56 | 61 |
| NL-FIT              | 76 | 74 | 75 | 62 | 64 | 63 | 69 | 63 | 66 | 61 | 57 | 59 | 58 | 65 | 61 |
| TubOslo             | 82 | 67 | 74 | 62 | 63 | 62 | 62 | 68 | 65 | 59 | 56 | 58 | 57 | 66 | 62 |
| YNU_Lab             | 74 | 74 | 74 | 66 | 56 | 61 | 63 | 67 | 65 | 55 | 61 | 58 | 63 | 56 | 60 |
| BrainIt             | 77 | 70 | 73 | 61 | 60 | 60 | 60 | 68 | 64 | 56 | 55 | 55 | 60 | 57 | 59 |
| EmoNLP              | 73 | 72 | 73 | 62 | 57 | 60 | 63 | 62 | 63 | 55 | 56 | 56 | 56 | 61 | 58 |
| RW                  | 71 | 72 | 72 | 60 | 57 | 59 | 62 | 63 | 62 | 55 | 52 | 53 | 56 | 60 | 58 |
| Baseline            | 69 | 71 | 70 | 58 | 54 | 56 | 62 | 62 | 62 | 54 | 51 | 52 | 55 | 59 | 57 |
| USI-IR              | 71 | 69 | 70 | 58 | 51 | 54 | 59 | 59 | 59 | 49 | 58 | 53 | 57 | 50 | 53 |
| THU_NGN             | 77 | 78 | 77 | 69 | 63 | 66 | 68 | 68 | 68 | 60 | 63 | 62 | 61 | 66 | 64 |
| SINAI               | 68 | 68 | 68 | 52 | 52 | 52 | 59 | 60 | 59 | 52 | 51 | 52 | 56 | 55 | 55 |
| UTFPR               | 64 | 53 | 58 | 54 | 60 | 57 | 59 | 58 | 58 | 50 | 53 | 52 | 51 | 62 | 56 |
| CNH2017             | 65 | 70 | 67 | 58 | 47 | 52 | 58 | 59 | 59 | 51 | 48 | 50 | 49 | 58 | 53 |
| ly3b3               | 72 | 64 | 68 | 58 | 46 | 52 | 55 | 62 | 58 | 46 | 53 | 50 | 47 | 50 | 49 |
| AdobeResearch       | 62 | 65 | 63 | 52 | 52 | 52 | 52 | 51 | 52 | 48 | 45 | 46 | 49 | 52 | 50 |
| Anonymous           | 76 | 77 | 76 | 64 | 67 | 65 | 70 | 64 | 67 | 62 | 59 | 60 | 59 | 69 | 64 |
| dinel               | 61 | 61 | 61 | 52 | 37 | 43 | 52 | 49 | 50 | 44 | 50 | 47 | 44 | 54 | 48 |
| CHANDA              | 46 | 64 | 54 | 39 | 36 | 38 | 54 | 42 | 47 | 38 | 37 | 37 | 51 | 20 | 29 |
| NLP_LDW             | 33 | 38 | 36 | 18 | 12 | 14 | 20 | 31 | 25 | 22 | 26 | 24 | 18 | 7 | 10 |

Table 10: Results by emotion class. Note that this table is limited to the six emotion labels of interest in the data set. However, other labels predicted than these six were taken into account for calculation of the final macro F₁ score. Therefore, the macro F₁ calculated from this table is different from the results in Table 5 in two cases (THU_NGN and Anonymous, who would be on rank 9 and rank 10, when predictions for classes outside the labels were ignored.).
### B Examples

Table 11 shows examples which have been correctly or wrongly predicted by all instances. They are discussed in Section 4.5.

| Emotion | Instance |
|---------|----------|
| **Anger** | + You can’t spend your whole life holding the door open for people and then being TRIGGER when they don’t thank you. Nobody asked you to do it. |
| | + I get impatient and TRIGGER when I’m hungry |
| | + Anyone have the first fast and TRIGGER that I can borrow? |
| | – I’m kinda TRIGGER that I have to work on Father’s Day |
| | – @USERNAME she’ll become TRIGGER that I live close by and she will find me and punch me |
| | – This has been such a miserable day and I’m TRIGGER because I wish I could’ve enjoyed myself more |
| | + I find it TRIGGER when I can see your underwear through your leggings |
| | + @USERNAME ew ew ewww your weird I can’t I would feel so TRIGGER when people touch my hair |
| | + nyc smells TRIGGER when it’s wet. |
| **Disgust** | + I wanted a cup of coffee for the train ride. Got ignored twice. I left TRIGGER because I can’t afford to miss my train. #needcoffee :(
| | – So this thing where other black people ask where you’re “really” from then act TRIGGER when you reply with some US state. STAH |
| | – I’m so TRIGGER that I have to go to the post office to get my jacket that I ordered because delivering it was obviously rocket science |
| | + @USERNAME & explain how much the boys mean to me but I’m too TRIGGER that they’ll just laugh at me bc my dad laughed after he |
| | + I threw up in a parking lot last night. I’m TRIGGER that’s becoming my thing. #illbutmostlymentally |
| | + When you holding back your emotions and you’re TRIGGER that when someone tries to comfort you they’ll come spilling out http://url.removed |
| **Fear** | – It’s so funny how people come up to me at work speaking Portuguese and they get TRIGGER when I respond in Portuguese |
| | – @USERNAME it seems so fun but i haven’t got to try it yet. my mom and sis are always TRIGGER when i try do something new with food |
| | – @USERNAME It’s hard to be TRIGGER when your giggle is so cute |
| | + maybe im so unTRIGGER because i never see the sunlight? |
| | + @USERNAME you’re so welcome !! i’m super TRIGGER that i’ve discovered ur work ! cant wait to see more !! |
| | + @USERNAME Im so TRIGGER that you guys had fun love you |
| **Joy** | – @USERNAME Not TRIGGER that your show is a rerun. It seems every week one or more your segments is a rerun. |
| | – I am actually TRIGGER when not invited to certain things. I don’t have the time and patience to pretend. |
| | – This has been such a miserable day and I’m TRIGGER because I wish I could’ve enjoyed myself more |
| | + this award honestly made me so TRIGGER because my teacher is leaving http://url.removed |
| | + It is very TRIGGER that people think depression actually does work like that... http://url.removed |
| | + @USERNAME @USERNAME @USERNAME It’s also TRIGGER that you so hurl about it :(
| | – Some bitch stole my seat then I had to steal the seat next to me. The boy looked TRIGGER when he saw me, and he was smart! #wasponnapass |
| | – I was so TRIGGER because I was having fun lol then i slipped cus I wasn’t wearing shoes |
| | – @USERNAME I wipe at my eyes next, then swim a bit. “I’m sorry.” I repeat, TRIGGER that I made him worry. |
| **Sadness** | + why am i not TRIGGER that cal said that |
| | + @USERNAME why am I not TRIGGER that you’re the founder |
| | + @USERNAME I’m still TRIGGER when students know my name. I’m usually just “that guy who wears bow ties” =) (and there are a few at WC) |
| | – It’s TRIGGER when I see people that have the same phone as me no has htcs |
| | – There is a little boy in here who is TRIGGER that he has to pay for things and that we won’t just give him things |
| | – totally TRIGGER that my fans celebrating easter today because my sister goes back to uni sunday |

Table 11: Subsample of Tweets that were correctly predicted by all teams and of Tweets that were not correctly predicted by any team.