# UIUC at SemEval-2018 Task 1: Recognizing Affect with Ensemble Models

**Abhishek Narwekar**
Department of Computer Science
University of Illinois, Urbana Champaign
USA 61820
abhisheknkar@gmail.com

**Roxana Girju**
Linguistics Department,
Computer Science Department,
Beckman Research Institute,
University of Illinois, Urbana Champaign
USA 61820
girju@illinois.edu

## Abstract

Our submission to the SemEval-2018 Task1: *Affect in Tweets* shared task competition is a supervised learning model relying on standard lexicon features coupled with word embedding features. We used an ensemble of diverse models, including random forests, gradient boosted trees, and linear models, corrected for training-development set mismatch. We submitted the system's output for subtasks 1 (emotion intensity prediction), 2 (emotion ordinal classification), 3 (valence intensity regression) and 4 (valence ordinal classification), for English tweets. We placed $25^{th}, 19^{th}, 24^{th}$ and $15^{th}$ in the four subtasks respectively. The baseline considered was an SVM (Support Vector Machines) model with linear kernel on the lexicon and embedding based features. Our system's final performance measured in Pearson correlation scores outperformed the baseline by a margin of 2.2% to 14.6% across all tasks.

## 1 Introduction

Affective computing deals with the recognition, interpretation, processing, and simulation of human affects. It is a highly interdisciplinary field at the heart of a broad range of technological applications in health care, media & advertisement, automotive, and others.

Although emotions are a fundamental feature of human experience, they have been long ignored by technology development mainly due to their complex and subjective nature, as well as the lack of learning capabilities to detect them. Current affective computing systems focus mainly on facial expressions, body language, speech (tone of voice, rhythm, etc.), keystroke as well as physiological input (e.g., heart rate, body temperature) to capture and process changes in a user's emotional state. However, in environments such as social media and Internet forums, most often the only signal is written language. And since language per se is the smallest portion of human communication (Mehrabian, 1981), emotions are not easy to detect.

Although emotion detection is directly related to the more popular task of sentiment analysis, they differ in many respects. Sentiment Analysis aims to detect the positive, neutral, or negative orientation of the text, while emotion detection focuses on recognizing and classifying text snippets into a set of predefined, more or less universal emotions. Various such classification models have been proposed, two famous ones being Ekman's (Ekman, 1997) six basic emotions (anger, happiness, surprise, disgust, sadness, and fear) and Plutchik's wheel of eight emotions (Plutchik, 2001), where each primary emotion has a polar opposite (joy, trust, fear, surprise, sadness, anticipation, anger, and disgust).

To date, there are many freely available tools for sentiment polarity classification of input text, yet not so many exist for emotion detection. Major challenges are: (1) the difficulty in establishing ground truth for various emotions, (2) the high variability, vagueness, ambiguity, and implicitness of language that can make the detection very problematic, (3) the scarcity of non-verbal clues in written communication, as well as (4) the challenge of getting access to and being able to process the right type of context. This can be explained by the ”7% Rule” (Mehrabian, 1981): only 7% of human communication is verbal while over 90% is comprised of tone of voice (38%) and body language (55%).

This year, SemEval 2018 hosts Task1: *Affect in Tweets* (Mohammad et al., 2018) - a shared task competition aiming to predict emotions and sentiment in tweets. There are five sub-tasks (Table 1). The participating systems have to automati-

*Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pages 377–384
New Orleans, Louisiana, June 5–6, 2018. ©2018 Association for Computational Linguistics

cally determine the intensity of emotions (E) and intensity of sentiment (i.e., valence V) from a collection of tweets, as experienced by the authors of these tweets. The organizers also include a multilabel emotion classification task for tweets. For each task, separate training and test data sets for each language considered are provided to the participants.

| ID | Task Label | Input | Output |
|---|---|---|---|
| 1 | El-reg | Tweet ($t$),<br>Emotion ($e$) | Intensity($e$, $t$) $\in$ (0, 1) |
| 2 | El-oc | Tweet ($t$),<br>Emotion ($e$) | 0 $\leq$ Intensity($e$, $t$) $\leq$ 3,<br>Intensity($e$, $t$) $\in \mathbf{N}$ |
| 3 | V-reg | Tweet ($t$),<br>Sentiment ($s$) | Intensity($s$, $t$) $\in$ (0, 1) |
| 4 | V-oc | Tweet ($t$),<br>Sentiment ($s$) | -3 $\leq$ Intensity($s$, $t$) $\leq$ 3,<br>Intensity($s$, $t$) $\in \mathbf{N}$ |
| 5 | E-c | Tweet ($t$),<br>Emotions ($s$) | Class: neutral/<br>no emotion/<br>multiple emotions |

Table 1: Description of the five sub-tasks of Task1: Affect in Tweets at SemEval 2018.

The contributions of the UIUC system are as follows: (1) In this competition, we demonstrate the use of a system that uses lexicon- and embedding-based features in an ensemble model of diverse approaches such as random forests, gradient boosted trees, and linear classifiers. We demonstrate how their combination in the final ensemble outperforms each of the individual methods. (2) We account for the train-development mismatch in the dataset by training a separate model to learn this mismatch. (3) We analyze the UIUC system and several variants of it, some of which improve on its performance. (4) We also perform an error analysis of "difficult" tweets, and explore areas for improvement of the model.

## 2 Related Work

**Word-Emotion Lexicons:** Word-emotion lexicons are a mapping between the words in the vocabulary to an emotion rating. Some lexicons map words to discrete emotions, such as General Inquirer (Stone et al., 1962), Wordnet Affect (Strapparava et al., 2004) and the NRC-10 Emotion Lexicon (Mohammad and Turney, 2013). Others, such as Affective Norms for English Words (ANEW) (Bradley and Lang, 1999) and WKB Corpus (Warriner et al., 2013), map them to dimensions such as valence, arousal and dominance.

**Sentence-Level Labeled Corpora:** Large scale corpora annotated with sentence-level emotion labels are uncommon in the literature. Affective Text (Strapparava and Mihalcea, 2007), created for SemEval 2007, contains emotion annotations headlines of news articles. Alm et al. annotated about 185 children's stories with the Ekman labels. Aman and Szpakowicz created annotated 5,000 sentences with additional labels for intensity and emotion bearing phrases. Preotiuc-Pietro et al. annotated 3,000 social media posts for valence and arousal, making this one of the few datasets that contains annotations based on the VAD model.

**Approaches:** *Rule-based approaches* incorporate domain knowledge. This can include term-based n-gram features, distance between certain terms or pre-specified POS patterns. Early work in this area focused mainly on linguistic heuristics (Hatzivassiloglou and McKeown, 1997). However, a major drawback of these rule-based approaches is that they are unable to detect novel expression of sentiment. *Keyword based approaches* classify text based on the detection of unambiguous words in language. They depend on large scale lexicons with affective labels for words, such NRC (Mohammad and Turney, 2013). *Knowledge-based approaches* use web ontologies or semantic networks. A major advantage of such systems is that they enable the system to use conceptual ideas derived from world knowledge (Cambria and Hussain, 2012). Recently, *distributed approaches* have been proposed that leverage word embeddings and train deep neural networks on the embedding space (Mohammad and Bravo-Marquez, 2017a).

**Shared evaluations** have encouraged the community to create benchmarks over shared tasks, and have been organized frequently. The Affective Text task at SemEval 2007 (Strapparava and Mihalcea, 2007) asked its participants to predict emotion labels for headlines of news articles. More recently, the Shared Task on Emotion Intensity (EmoInt) at WASSA 2017 (Mohammad and Bravo-Marquez, 2017a), had 22 participating teams who were given a corpus of 3,960 English tweets annotated with a continuous intensity score for each of four of Ekman's basic emotions: anger, fear, joy and sadness.

## 3 Dataset

Tasks 1 and 2 share the same training and development data sets: a total of 7,500 sentences in training and about 1,600 sentences in development across the four emotions: anger, fear, joy, sadness. It is interesting to note that the training data sets for the emotions of fear, anger and sadness overlap significantly: all pairs have a Jaccard similarity of over 0.5. This means that over 67% of the data sets across these emotions contain the same tweets.

Tasks 3 and 4 share the same data sets as well, for a total of 1,200 tweets in training and 450 tweets in development across the four emotions.

Another interesting overlap is between the tweet collections for Tasks 5 and 1 (and therefore Task 2): The data set for Task 5 appears to be made up largely of the tweets for Task 1, for both the training and development sets. These overlaps of the training and development data sets across all emotions gave us the idea to tackle all tasks using a common set of features. For instance, Tasks 2 and 4 may be solved by simply transforming the output of Tasks 1 and 3, respectively. Task 5 involves a multi-label classification and thus, needs more thought.

In the test set, with the exception of the first 1,000 or so sentences, nearly 95% of the total sentences for Tasks 1A and 3A (i.e., for English) are the so called “mystery” sentences – meaning, essentially neutral sentences without any emotional content. The scores reported by the organizers are for the non-mystery sentences only (i.e., non-neutral).

## 4 The UIUC System

Our system takes as input features from affective lexicons and word embeddings trained on affective Twitter corpora. We then train an ensemble of diverse models over these features. Given that the training and development labels are not directly comparable, we also model the mismatch between the two sets. Moreover, we also describe additional models that we constructed after the competition deadline (section 4.4). We report results for tasks 1A, 2A, 3A and 4A (where ’A’ identifies the target language: English).

### 4.1 Feature Space

We have used the *AffectiveTweets* (Mohammad and Bravo-Marquez, 2017b), a package in Weka (Hall et al., 2009) for extracting certain features from a tweet. The features extracted are: MPQA (Wilson et al., 2005), BingLiu (Bauman et al., 2017), AFINN (Nielsen, 2011), Sentiment-140 Emoticon (Kiritchenko et al., 2014), NRC Hashtag Emotion Lexicon (Mohammad and Kiritchenko, 2015), NRC emotion lexicon (wordlevel) (Mohammad and Turney, 2013), SentiWordNet (Baccianella et al., 2010), NegatingWordList(Mohammad and Bravo-Marquez, 2017b). We call these *lexicon features*. In addition, we also extract *Embedding based features* (Twitter Edinburgh 100D / 400D corpus) using the AffectiveTweets package.

### 4.2 Models

The UIUC system contains an ensemble constructed using stacking of several base learners. A schematic of this ensemble is shown in figure 1. We obtained out-of-fold predictions for each of these three models using 5-fold cross validation on layer 1. These predictions were concatenated and provided as input to layer 2. Parameters of the models in this ensemble are detailed below:

**Layer 1**
**Random Forests**:
n_estimators=100, max_features=$\sqrt{F}$ (F=total features), max_depth=5, min_samples_leaf=2
**XGB**[1]: max_depth=5, min_child_weight=150, gamma=0, n_estimators=150, reg_alpha=0.01, reg_lambda=0.87, learning_rate=0.1
**SVM**:kernel=linear, C=0.1

**Layer 2**
**XGB**: max_depth=3, min_child_weight=1, gamma=0, n_estimators=100, reg_alpha=0.1, reg_lambda=1, learning_rate=0.1, random_state=0

### 4.3 Modeling the Mismatch Between the Training and Development Sets

According to the organizers, the training set for the task was created from an annotation effort in 2016 (Mohammad and Bravo-Marquez, 2017a). The development and test sets were created from a common 2017 annotation effort. As a result, the scores for tweets across the training and development sets or across the training and test sets are not directly comparable. We therefore devise a model that can predict and eliminate the mismatch between the two sets of labels. As a means to model the mismatch in the distributions of the two label sets, we train a linear model that, for the labels in the development set, learns a function between the predictions made for the development set and

[1] Note: XGB stands for the XGBoost implementation of gradient boosted decision trees. SVM was implmented using sklearn (http://scikit-learn.org/stable/).

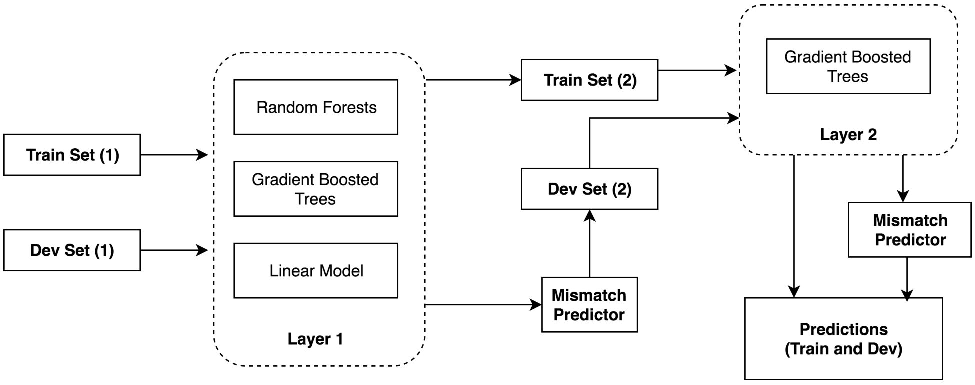

Figure 1: Ensemble used in the UIUC system.

the ground truth. This learner does not affect the training in any way, but is a way to transform the predictions made for the development set so that they are comparable to the ground truth labels.

### 4.4 Additional Models

After the competition deadline, we built and evaluated additional models. The overall model was an ensemble with the same structure as the official submission. Additions include implementation of neural models of computation. In particular, we implement feed-forward neural networks (using the average word embeddings as input), LSTM-CNN (using individual word embeddings) and character level LSTM (using the character stream). The neural networks were implemented in Keras (Chollet et al., 2015) with the Tensorflow (Abadi et al., 2016) backend. Details about these additional models are shown below.

**Layer 1:**
**SVM (layer 1)**: C=0.1, kernel=RBF
**XGB**: max_depth=5, reg_lambda=0.87, min_child_weight=150, n_estimators=150
**FFNN (feed forward neural network)**: Dense (256, sigmoid), Dropout (0.2, sigmoid), Dense (64, sigmoid), Dense (32, sigmoid), Dense (1, relu)
**LSTM-CNN**: Conv1D (300, 3, relu), Dropout (0.2), LSTM (150), Dropout (0.2), Dense (32, sigmoid), Dense (1, relu)
**Character level language model (Char)**: LSTM (150), Dropout (0.2), Dense (64, sigmoid), Dense (1, relu)

**Layer 2:**
**SVM**: C=1, kernel=RBF

## 5 Results

In this section, we describe the results of our official submission to SemEval 2018 (subsection 5.1) as well as the results of experiments on additional models constructed after the competition deadline (section 5.2).

### 5.1 Performance of the UIUC system

Tables 2 and 3 show the performance of our model for Tasks 1A, 2A, 3A and 4A, respectively. We have shown the results by comparing our model against the baseline, which has been trained using an SVM with linear kernel on the lexicon and embedding based features. Our submission outperforms the baseline in nearly all the task-emotion pairs.

In particular, we observe that the results for the prediction of data points in the 0.5 – 1 range are poorer than in the overall range. The reason for this is that the finer prediction is a harder task than the overall prediction, and exactly predicting the emotion intensity given that it is high has significant variance. Scores for Task 2A are worse than those for Task 1A in spite of the similarity of the tasks. This is because in Task 2A, we essentially discretize the output, thereby either increasing or decreasing the absolute error between the intensity predicted and the actual intensity, depending on whether the discretized output is correct or not. On the whole, evidently, the correlation drops as the effect of the latter case (increase in the absolute error) dominates over the former.

Tasks 3 and 4 follow similar trends as Tasks 1 and 2 respectively, but we see a higher correlation for these tasks as compared to Tasks 1 and 2, respectively. This leads to the conclusion that predicting the sentiment is an easier task than pre-

| Subtask | Submission | Pearson (all instances) | | | | | Alternate evaluation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | macro-avg | anger | fear | joy | sadness | macro-avg | anger | fear | joy | sadness |
| 1 | UIUC System | 0.647 | 0.663 | 0.646 | 0.649 | 0.628 | 0.463 | 0.514 | 0.431 | 0.422 | 0.485 |
| | Baseline | 0.630 | 0.652 | 0.625 | 0.632 | 0.610 | 0.462 | 0.523 | 0.418 | 0.424 | 0.481 |
| 2 | UIUC System | 0.518 | 0.514 | 0.449 | 0.576 | 0.533 | 0.463 | 0.414 | 0.392 | 0.562 | 0.484 |
| | Baseline | 0.448 | 0.512 | 0.258 | 0.527 | 0.493 | 0.334 | 0.335 | 0.219 | 0.415 | 0.366 |

Table 2: Results of the UIUC system for subtasks 1a (emotion intensity regression) and 2a (emotion ordinal classification) and comparison with baseline. The alternate evaluation is Pearson correlation for tweets with scores between 0.5 and 1 for subtask 1 and the Cohen Kappa (Cohen, 1960) for subtask 2.

| Subtask | Model | Pearson | Alternate evaluation |
|---|---|---|---|
| 3 | UIUC system | 0.762 | 0.582 |
| | Baseline | 0.746 | 0.565 |
| 4 | UIUC system | 0.724 | 0.694 |
| | Baseline | 0.688 | 0.673 |

Table 3: Results of the UIUC system for tasks 3a (valence intensity regression) and 4a (valence ordinal classification) and comparison with baseline. The alternate evaluation was Pearson correlation for tweets with score 0.5-1 for subtask 3a and Cohen's Kappa for subtask 4a.

dicting the intensity of a given emotion.

### 5.2 Performance of Additional Models

**Ablation Study for Task 1:** Given the multiple subsections of data, it is difficult to optimize the architecture and parameters for all emotions for all subtasks. Therefore, we focus on optimizing the architecture and parameters for only the first subtask (emotion intensity prediction) for the emotion *anger*. Given the many models developed and presented here, it is interesting to see how they perform individually on this subtask. Table 4 shows the performance of various feature-model combinations. Note that **L** and **E** in the *Features* column indicate lexicon-based and embedding-based features respectively.

| Feature | Model | CV | Dev | Test |
|---|---|---|---|---|
| L | SVM | 0.646 | 0.616 | 0.654 |
| | XGB | 0.648 | 0.646 | 0.634 |
| | FFNN | 0.699 | 0.674 | 0.664 |
| | SVM+XGB | 0.662 | 0.651 | 0.663 |
| | SVM+XGB+FFNN [M1] | 0.695 | 0.674 | 0.673 |
| E | SVM | 0.564 | 0.553 | 0.555 |
| | LSTM | 0.640 | 0.635 | 0.633 |
| | LSTM-CNN | 0.641 | 0.639 | 0.635 |
| | LSTM-CNN (Att) [M2] | 0.651 | 0.642 | 0.644 |
| L+E | M1+M2 | 0.733 | **0.713** | 0.701 |
| | M1+M2+Char | **0.735** | 0.711 | **0.704** |

Table 4: An ablation study of various features and models for subtask 1: emotion intensity prediction for the specific case of the emotion *anger*.

We use the SVM trained on lexical features as the baseline. We can see that the SVM+XGB+FFNN (referred to as M1) performs better than the SVM alone. LSTM-CNN with attention (referred to as M2) performs similarly to the SVM baseline. However, when combined together, the model M1+M2+Char performs better than each of the individual models on the test set. This means that the different models capture complementary information about the input, and work better in unison, thus demonstrating the efficacy of the idea of ensembling.

Henceforth, we use **M1** to refer to the SVM + XGBoost + Feedforward Neural Network architecture trained on lexical features, **M2** to refer to the LSTM-CNN architecture with attention trained on the embedding features and **Char** to refer to the character level LSTM model trained on the individual characters.

| Task | Features | Model | Pearson Correlation Coefficient | | | |
|---|---|---|---|---|---|---|
| | | | Anger | Fear | Joy | Sadness |
| 1 | L | SVM | 0.654 | 0.646 | 0.649 | 0.628 |
| | L | M1 | 0.673 | 0.668 | 0.698 | 0.642 |
| | E | M2 | 0.644 | 0.659 | 0.685 | 0.644 |
| | L+E | M1+M2+Char | **0.704** | **0.688** | **0.713** | **0.652** |
| 2 | L | SVM | 0.514 | 0.449 | 0.576 | 0.533 |
| | L | M1 | 0.549 | **0.462** | 0.58 | 0.557 |
| | E | M2 | 0.544 | 0.455 | 0.571 | 0.542 |
| | L+E | M1+M2+Char | **0.558** | 0.461 | **0.601** | **0.566** |

Table 5: Evaluation for subtask 1 (emotion intensity prediction) and subtask 2 (emotion ordinal classification) for all emotions with various features and models.

**Tasks 1 and 2 with with Additional Models:** Table 5 shows the performance of the models described above to the first two subtasks: emotion intensity prediction and emotion ordinal classification. We have shown the results for all the four emotions. As we can see, here too, the model combination M1+M2+Char combination performs the best for all emotions in subtask 1. The performance of the model is the best for the emotion *joy*, and the worst for the emotion *fear*.

**Tasks 3 and 4 with with Additional Models:** Coming to subtasks 3 and 4 (valence intensity prediction and valence ordinal classification respectively), Table 6 shows the performance of the var-

| Task | Features | Model | Alternate Evaluation |
|---|---|---|---|
| 3 | L | SVM | 0.762 |
|  | L | M1 | 0.78 |
|  | E | M2 | 0.764 |
|  | L+E | M1+M2+Char | **0.784** |
| 4 | L | SVM | 0.724 |
|  | L | M1 | 0.733 |
|  | E | M2 | 0.745 |
|  | L+E | M1+M2+Char | **0.75** |

Table 6: Evaluation for subtask 3 (emotion valence regression) and subtask 4 (valence ordinal classification) for various features and models. The alternate evaluation is the Pearson correlation for tweets with scores 0.5 - 1 for subtask 3 and the Cohen's Kappa for subtask 4.

ious models on these tasks. Consistent with the results of subtasks 1 and 2, the combined model M1+M2+Char performs the best for both tasks.

In general, we note that the correlation is significantly higher on valence prediction tasks as compared to the emotion intensity tasks. This is likely because the emotion intensity prediction is a fine grained task, requiring the model to observe patterns specific to an emotion. Valence is more of an "aggregated" effect of all the emotions.

Had the best model in additional experiments for all subtasks been submitted to SemEval with all other factors constant, its rank based on the macro-average for the first four subtasks would have been $15^{th}$, $15^{th}$, $18^{th}$ and $13^{th}$ respectively.

## 6 Discussion

In order to identify areas where the model can improve, it is necessary to study cases where it performs poorly. To do so, we select 5 sentences where the baseline SVM model performs very poorly while predicting anger intensity (based on absolute error) and 1 sentence where it performs well. We have restricted the number of sentences to 6 for brevity. In particular, for sentences 1 and 2, the model significantly overestimates the intensity, for sentence 3, the model predicts the intensity accurately. For sentences 4, 5 and 6, the model significantly underestimates the intensity. Table 7 shows the sentences considered and the true value of emotion intensity for the emotion anger.

| # | Tweet | Intensity |
|---|---|---|
| 1 | never had a dull moment with u guys | 0.078 |
| 2 | Fast and furious marathon soon! | 0.118 |
| 3 | They cancelled Chewing Gum. #devastated | 0.625 |
| 4 | Its taking apart my lawn! GET OFF MY LAWN! | 0.797 |
| 5 | I need a beer #irritated | 0.806 |
| 6 | Working with alergies is the most miserable shit in the world #miserable #alergies | 0.856 |

Table 7: Test Examples for Error Analysis with intensity annotations for anger.

We then compare the absolute error between the true value and model prediction for various models. This comparison is shown in Table 8. Given that 5 of the 6 sentences are "difficult" for the models, we observe that there is no clear winner over these sentences. However, we observe that for sentences 1 and 2, the model M1 performs relatively well. For sentences 4, 5 and 6, the models involving M2 perform relatively well. This suggests that M1 is better at predicting the lower intensities, while M2 is better at the higher intensities. This may explain why though the overall scores for the two models was similar, the ensembled model outperformed the individual models. Another interesting observation is that for sentence 4, the presence of the capital letters is the reason for the high intensity. The model M1+M2+Char is able to identify this well, and contributes to reducing the error significantly as compared to all the other models.

| Features | Model | Sentence-wise error | | | | | |
|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | **5** | **6** |
| L | SVM | 0.310 | 0.308 | **0.004** | -0.377 | -0.326 | -0.327 |
| L | M1 | **0.305** | **0.287** | -0.067 | -0.391 | -0.286 | -0.245 |
| E | M2 | 0.344 | 0.366 | 0.051 | -0.265 | **-0.241** | **-0.199** |
| L+E | M1+M2+Char | 0.339 | 0.373 | 0.071 | **-0.213** | -0.242 | -0.203 |

Table 8: Absolute error values for various features and models for subtask 1: emotion intensity prediction for emotion *anger*.

## 7 Conclusion

In this paper we presented the UIUC system that performs regression and ordinal classification of the emotion and sentiment present in English tweets. Our system comprised an ensemble trained on lexicon based and embedding based features. We also provided an account for the training and development mismatch in a given data set by training an adaptive model between the model predictions and the final test predictions. We finally perform an error analysis over the various models to identify potential sources of improvement to the model.

## References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283.

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 *International Conference on Text, Speech and Dialogue*, pages 196–205. Springer.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*. European Language Resources Association.

Konstantin Bauman, Bing Liu, and Alexander Tuzhilin. 2017. Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews. In *Proceedings of the 23rd ACM SIGKDD*, pages 717–725. ACM.

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.

Erik Cambria and Amir Hussain. 2012. *Sentic computing: Techniques, tools, and applications*, volume 2. Springer Science & Business Media.

François Chollet et al. 2015. Keras.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Journal of Educational and Psychological Measurement*, 20(1):37.

Paul Ekman. 1997. *Basic Emotions*. Handbook of Cognition and Emotion, John Wiley & SOns, Sussex, UK.

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

Vasileios Hatzivassiloglou and Kathleen R McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pages 174–181. Association for Computational Linguistics.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Albert Mehrabian. 1981. *Silent messages: Implicit communication of emotions and attitudes (2nd ed)*. Wadsworth Pub. Co., Belmont, California.

Saif Mohammad and Felipe Bravo-Marquez. 2017a. Wassa shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, pages 34–49.

Saif M Mohammad and Felipe Bravo-Marquez. 2017b. Emotion intensities in tweets. *arXiv preprint arXiv:1708.03696*.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval)*, 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 and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Robert Plutchik. 2001. The nature of emotions. *American Scientist*, 89.

Daniel Preotiuc-Pietro, H Andrew Schwartz, Gregory Park, Johannes C Eichstaedt, Margaret Kern, Lyle Ungar, and Elizabeth P Shulman. 2016. Modelling valence and arousal in facebook posts. In *Proceedings of NAACL-HLT*, pages 9–15.

Philip J Stone, Robert F Bales, J Zvi Namenwirth, and Daniel M Ogilvie. 1962. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7(4):484–498.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics.

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

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT/EMNLP*, pages 347–354. Association for Computational Linguistics.