CS671A Project: Affect In Tweets

Interim Project Report

Anshdha	**Debabrata Ghosh**	**Apurv Gupta**
13817128	13817226	14817124
Bhanu Yadav	**Piyush Bagad**	**Divyat Mahajan**
13817198	150487	14227(Auditor)

Abstract

We address the problem of Emotion Intensity detection in Tweets as a compilation of an array of tasks intended to automatically determine the intensity of emotions and intensity of sentiment or valence of the tweeters from their tweets. The individual tasks comprise of emotion and intensity regression and ordinal classification as well as multi-label classification of emotions. We also intend to study the effect of Emoticons and hash tags on the intensity of emotion/s in the tweets and plan to study the relation between different emotions, lexical qualities that sentences of a particular emotion may possess peculiarly and effects of using different features for different emotions.

1. Problem Description

The problem that we are working on is the *SemEval-2018 Task 1: Affect in Tweets (AIT-2018)*. It presents an array of tasks where systems have to automatically determine the intensity of emotions (E) and intensity of sentiment (aka valence V) of the tweeters from their tweets. We also include a multi-label emotion classification task for tweets. The details of each of the five individual tasks are described below.

Notations - Set of emotions: S={ anger, fear, joy, sadness }, E: emotion, EI: emotion intensity, V: valence or sentiment intensity, reg: regression, oc: ordinal classification, c:classification.

1. **Task EI-reg**: Detecting Emotion Intensity (regression)
   Given: a tweet, an emotion E $\in S$

   Task: determine the intensity of E that best represents the mental state of the tweeter- a real valued score between 0 and 1:

   - a score of 1: highest amount of E can be inferred
   - a score of 0: lowest amount of E can be inferred

   (Note that the absolute scores have no inherent meaning – they are used only as a means to convey that the instances with higher scores correspond to a greater degree of E than instances with lower scores.)

2. **Task EI-oc**: Detecting Emotion Intensity (ordinal classification)
   Given: a tweet, an emotion E $\in S$

Task: classify the tweet into one of four ordinal classes of intensity of E that best represents the mental state of the tweeter:

- 0: no E can be inferred
- 1: low amount of E can be inferred
- 2: moderate amount of E can be inferred
- 3: high amount of E can be inferred

3. **Task V-reg**: Detecting Valence or Sentiment Intensity (regression)
Given: a tweet

   Task: determine the intensity of sentiment or valence (V) that best represents the mental state of the tweeter- a real valued score between 0 and 1:

   - a score of 1: most positive mental state can be inferred
   - a score of 0: most negative mental state can be inferred

4. **Task V-oc**: Detecting Valence (ordinal classification)
Given: a tweet

   Task: classify the tweet into one of seven ordinal classes, corresponding to various levels of positive and negative sentiment intensity, that best represents the mental state of the tweeter:

   - 3: very positive mental state can be inferred
   - 2: moderately positive mental state can be inferred
   - 1: slightly positive mental state can be inferred
   - 0: neutral or mixed mental state can be inferred
   - -1: slightly negative mental state can be inferred
   - -2: moderately negative mental state can be inferred
   - -3: very negative mental state can be inferred

5. **Task E-c**: Detecting Emotions (multi-label classification)
Given: a tweet

   Task: classify the tweet as 'neutral or no emotion' or as one, or more, of eleven given emotions that best represent the mental state of the tweeter:

   - anger (also includes annoyance and rage) can be inferred
   - anticipation (also includes interest and vigilance) can be inferred
   - disgust (also includes disinterest, dislike and loathing) can be inferred
   - fear (also includes apprehension, anxiety, concern, and terror) can be inferred
   - joy (also includes serenity and ecstasy) can be inferred
   - love (also includes affection) can be inferred
   - optimism (also includes hopefulness and confidence) can be inferred
   - pessimism (also includes cynicism and lack of confidence) can be inferred
   - sadness (also includes pensiveness and grief) can be inferred
   - surprise (also includes distraction and amazement) can be inferred
   - trust (also includes acceptance, liking, and admiration) can be inferred

2. Data Collection

A major challenge for building an emotion intensity predictor is the lack of properly annotated data [MB17]. Most of the currently available data sets are categorical rather than having a measure of the degree of the concerned emotion. A technique of annotation, called as the *Best-Worst Scaling (BWS)* came up as a possible solution for a cardinal measure of emotion intensities (Louviere, 1991 [L91]; Louviere et al., 2015 [LTM15]; Kiritchenko and Mohammad, 2016 [KM16]).

The BWS techniques involves the following procedure for labeled data set generation (as described in [MB17]): Annotators are given n items (an n-tuple, where n $>$ 1 and commonly n = 4). They are asked the best (highest in terms of the property of inter- est) and which is the worst (lowest in terms of the property of interest). When working on 4-tuples, best-worst annotations are particularly efficient because each best and worst annotation will reveal the order of five of the six item pairs. For example, for a 4-tuple with items A, B, C, and D, if A is the best, and D is the worst, then A $>$ B, A $>$ C, A $>$ D, B $>$ D, and C $>$ D. BWS annotations for a set of 4-tuples can be easily converted into real-valued scores of association between the items and the property of interest (Flynn and Marley, 2014 ([FM14])).

We will be using the labeled data sets supplied for the competition *SemEval-2018: International Workshop on Semantic Evaluation - Affects in Tweets*. These datasets used for the SemEval task were adopted from the work of [MB17] on BWS technique.

For each of the tasks $1, 2$ we have 4 training sets and 4 test sets: one for each emotion E. For the tasks $3, 4$ and 5, we have 1 training set, 1 test set. All the data sets are available on the SemEvam-2018 website: $competitions.codalab.org/competitions/17751\#learn_the_details-datasets$. Another dataset can be found in the following online repository: $saifmohammad.com/WebPages/TweetEmotionIntensity-dataviz.html$.

3. Literature Survey

The heavy use of social media platforms for thought expression has attracted sentiment analysis techniques such as emotion detection and opinion mining ([PL08]). Emotion detection has additionally been applied to multiple domains such as novels ([Mo11]), e-mail ([MY11]), news headlines ([SM08]) etc. In the domain of micro-blogging sites, like Twitter, limiting the length of tweets is likely to have interesting effects on the way emotions and their intensities are expressed through micro blogs. Kim et al.(2011) ([KBO11]) and Danescu-Niculescu-Mizil et al. (2011) ([DGD11]) focus on emotion intensity analysis and studying changes in emotions in Twitter conversations.

For *NSEmo at EmoInt-2017*, Madisetty and Desarkar ([MD17]) use an ensemble approach for emotion intensity detection. They use an ensemble of three regression methods: Support Vector Regression, Neural Networks and the baseline method to give performance that is better than that obtained through these methods individually.

WASSA-2017 Shared Task on Emotion Intensity: A competetion held with the task of automatically detecting intensity of emotion felt by the speaker of a tweet. The techniques applied by the top three teams Prayas, IMS and SeerNet are described below (adopted from [MBr17]):

The best performing system, *Prayas*, used an ensemble of three different models: The first is a feed-forward neural network whose input vector is formed by concatenating the average word embedding vector with the lexicon features vector provided by the AffectiveTweets package from [MB17]. The second model treats the problem as a multi-task learning problem with the labeling of the four emotion intensities as the four sub-tasks. Authors use the same neural network architecture as in the first model, but the weights of the first two network layers are shared across the four subtasks. The final predictions are made by combining

Team Name	r avg.	r fear	r joy	r sadness	r anger
Prayas	0.747	0.732	0.762	0.732	0.765
IMS	0.722	0.705	0.726	0.690	0.767
SeerNet	0.708	0.676	0.698	0.715	0.745

Table 1: Official Competition Metric: Pearson correlations (r) obtained by the systems on the full test sets. The bottom-line competition metric, 'r avg.', is the average of Pearson correlations obtained for each of the four emotions.

the first two models with three variations of the third model into an ensemble. A weighted average of the individual predictions is calculated using cross-validated performances as the relative weights.

IMS applies a random forest regression model to a representation formed by concatenating three vectors:

1. a feature vector drawn from existing affect lexicons,
2. a feature vector drawn from expanded affect lexicons, and
3. the output of a neural network.

The first vector is obtained using the lexicons implemented in the AffectiveTweets package. The second is based on an extended lexicons built from feed-forward neural networks trained on word embeddings. The third vector is taken from the output of neural network that combines CNN and LSTM layers.

SeerNet creates an ensemble of various regression algorithms (e.g, SVR, AdaBoost, random forest, gradient boosting). Each regression model is trained on a representation formed by the affect lexicon features (including those provided by AffectiveTweets) and word embeddings. Authors also experiment with different word embeddings models: Glove, Word2Vec, and Emoji embeddings ([ERABR16]). Table 1 summarizes performances of the top three teams in terms of Pearson Correlation coefficients (correlations with the labelled data sets obtained by BWS technique):

4. Methods to be employed (Tentative Plan)

4.1 Text Preprocessing

We plan to perform a comparative study on different methods of generating feature vectors. Different methods we plan to compare are:

1. **Bag of Words:** Removing stop words form the tweets and then constructing each tweet into a vector of vocabulary size, with each component of vector representing the frequency of the word corresponding to that index.

2. **Word2Vec:** Word2vec is based on the skip-gram architecture, that the meaning of a word depends on the word's surrounding it or the words in its context. We would train the word2vec model on the Glove dataset, remove Stop Words and convert each word into a vector using word2vec model. Here, each tweet would be a vector of vectors as each word is a vector of fixed dimension, in contrast to Bag of Words where each tweet is a vector of vocabulary size.

Bag of Words does not include contextual information while learning embeddings, hence we expect word2vec to give better results.

4.2 Regression tasks

Given a tweet and an emotion X, we need to implement automatic systems to determine the intensity or degree of emotion X felt by the speaker which is a real valued score between 0 and 1 that basically comprises of basic regression tasks. We'll use L_2-regularized L_2-loss SVM regression model with the regularization parameter λ set to 1. For this we plan to use Weka regression models on transformed data implemented in *LIBLINEAR*. Other regression algorithms to be implemented to instigate comparison between their results are Neural Networks, Random Forest, Gradient Boosting, AdaBoost and Least Square Regression. The high-level neural networks API library Keras can be used. We will use the following features:

- *Word N-grams (WN)*: presence or absence of word n-grams from n = 1 to n = 4.
- *Character N-grams (CN)*: presence or absence of character n-grams from n = 3 to n = 5.
- *Word Embeddings (WE)*: an average of the word embeddings of all the words in a tweet.

4.3 Classification tasks

Given a tweet, we need to implement either multi-class or single class classification techniques to indicate the presence of emotions. We will use several state-of-the-art techniques such as Kernel SVMs, Random Forests and Multi-layered Perceptrons to achieve this.

5. Possible new explorations

5.1 Study of effect of Emoticons and Hash tags

It may be interesting to study the affect of a hash tag or an emoticon may have on the emotion and its intensity. Emoticons in particular can be extremely handy for crucial revelations about the state of mind of the tweeter. We plan to test intensity scores with the hash tagged words marked and without explicitly marking them as hash tags.

5.2 Different features for different Emotions (Feature selection)

Roberts et al. ([RRJGH12]) have described a binary classifier for seven emotions while using different features with a greedy additive feature selection process for different emotions. In a part of their work, for each emotion, they chose the best set of features with a greedy additive feature selection process. This greedy process iteratively adds the next-best feature to the feature set provided it increases the F_1 score on a development set. Their findings Interestingly, the best performing emotion was FEAR, which was also the least frequent. Furthermore, the FEAR classifier uses only two features (unigrams and topics). This suggests this emotion is highly lexicalized with less variation than the other emotions, as it has comparable recall but significantly higher precision. It may be interesting to use a different feature selection methods in the domain of cardinal intensities as against those used in simple binary classification. We plan to apply variants of feature selection for different emotions and try getting conclusions about the linguistic structure associated with sentences expressing a particular emotion.

5.3 Skip-Thoughts for learning vector embedings

Skip-Thoughts train an encoder-decoder model where each input sentence would be mapped to a vector embedding through the Encoder and the Decoder generates the sentences in its context. It also has skip-gram idea, sentences in the context are used to learn embeddings for sentences. Here, we would consider the tweets dataset as a paragraph made up of sentences, each tweet as a sentence. Hence, it would learn a vector embedding for each tweet utilizing each the tweets present in its context.

Since, different tweets in the context of a tweet from different users might not share similar semantic properties (in contrast to normal paragraphs where sentences in context hope to share similar properties). However, it would be interesting to see how skip-thoughts compare with word2vec. If we get comparable performance, this will lead to interesting suggestions like different tweets corresponding to different aspects might still share some similar semantic information, and using contextual tweets might help in learning better embeddings.

References

[MBSK18] 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-2018)*, New Orleans, LA, USA, June 2018.

[MK16] Saif M. Mohammad, Mohammad Salameh and Svetlana Kiritchenko, "How Translation Alters Sentiment." *Journal of Artificial Intelligence Research*, 2016. Volume 55, pages 95-130

[MK18] Saif M. Mohammad and Svetlana Kiritchenko, "Understanding Emotions: A Dataset of Tweets to Study Interactions between Affect Categories." *In Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, May 2018, Miyazaki, Japan.

[MBr17] Saif M. Mohammad and Felipe Bravo-Marquez, "WASSA-2017 Shared Task on Emotion Intensity." *In Proceedings of the EMNLP 2017 Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media (WASSA)*, September 2017, Copenhagen, Denmark.

[MB17] Saif M. Mohammad and Felipe Bravo-Marquez, "Emotion Intensities in Tweets" *In Proceedings of the Sixth Joint Conference on Lexical and Computational Semantics (*Sem)*, August 2017, Vancouver, Canada.

[PL08] Bo Pang and Lillian Lee, "Opinion Mining and Sentiment Analysis" *Foundations and Trends in Information Retrieval archive Volume 2 Issue 1-2,* , January 2008 Pages 1-135.

[Mo11] Saif M. Mohammad, "From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales." *In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage,* , 2011, pages 105–114.

[MY11] Saif M. Mohammad and Tony Yang, "Sentiment in Mail: How Genders Differ on Emotional Axes." *In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)* , 2011, pages 70–79.

[SM08] Carlo Strapparava and Rada Mihalcea, " Learning to Identify Emotions in Text" *In Proceedings of the ACM Conference on Applied Computing.* , 2008.

[KBO11] Suin Kim, JinYeong Bak, and Alice Oh, " Do You Feel What I Feel? Social Aspects of Emotions in Twitter Conversations." , 2011.

[DGD11] Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais, " Mark my words! Linguistic style accommodation in social media." *In World Wide Web.*, 2011.

[MD17] Sreekanth Madisetty and Maunendra Sankar Desarkar, " NSEmo at EmoInt-2017: An Ensemble to Predict Emotion Intensity in Tweets" *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Copenhagen, Denmark, September 7–11, 2017, pages 219–224.

[ERABR16] Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bosnjak and Sebastian Riedel, "emoji2vec: Learning emoji representations from their description." *In Proceedings of The Fourth International Workshop on Natural Language Pro- cessing for Social Media. Association for Computational Linguistics,* 2016. Austin, TX, USA, pages 48–54.

[RRJGH12] Kirk Roberts, Michael A Roach, Joseph Johnson, Josh Guthrie and Sanda M Harabagiu, "EmpaTweet: Annotating and Detecting Emotions on Twitter." *Proc. Language Resources and Evaluation Conf.,* 2012.

[FM14] Terry Flynn, and A.A. Marley, "Best-worst scaling: Theory and methods." *178-201. 10.4337/9781781003152.00014.* 2014.

[LTM15] Jordan Louviere, Terry Flynn, and A.A. Marley, "Best-worst scaling: Theory, methods and applications." *10.1017/CBO9781107337855.* 2015.

[LTM15] Jordan Louviere, "Experimental choice analysis: introduction and overview" *Journal of Business Research* 1991. Volume 23 Issue 4 Pages 291-297.