Advanced Issues on NLP - Emotion Classification

Anonymous ACL submission

1 Introduction

The analysis of discrete emotion categories has received increasing attention in last few years (Alm et al., 2005; Aman and Szpakowicz, 2007; Dodds et al., 2011). Even though comprehensive pychological models of emotion categorization have been proposed (Ekman, 1999; Plutchik, 2001), the task remains challenging from a machine learning perspective as noted in recent shared tasks (Mohammad et al., 2016). More recently, the field has openend to wider application scenarios such as muti-lingual setups (Barbieri et al., 2018) and irony detection (Van Hee et al., 2018).

In this report we tackle the task of discrete emotion recognition over social media text. We follow the categorization of emotions proposed by Ekman (1999) and Plutchik (2001), considering the following seven categories: anger, disgust, fear, joy, sadness, surprise, and trust.

We take a nominal, multi-label classification approach to the problem, i.e. a single instance is labeled with a set of emotion classes. We compare the performance of three kinds of classifiers over the SSEC dataset (Schuff et al., 2017).

2 Experimental Setup

2.1 Dataset

We use the version of the Stance Sentiment Emotion Corpus, SSEC (Schuff et al., 2017), extracted from the provided aggregated file (*unified-dataset.csv*). The dataset consists of 4,868 tweets annotated in a multi-label setup with 7 kinds of emotions: anger, disgust, fear, joy, sadness, surprise, and trust. We held out 800 samples for test and keep the rest as training set. Table 1 presents the distribution of emotion classes over the training and test set. The average number of labels annotated per sample is 3.23 in the training set and 3.24 in the test set.

Emotion	Train	Test
Anger	2,432	470
Disgust	1,834	349
Fear	1,526	314
Joy	1,732	335
Sadness	2,206	438
Surprise	934	174
Trust	2,245	455

Table 1: Emotion class distribution in the training and test set. Note that samples are annotated in a multi-label setup.

2.2 Feature Extraction and Preprocessing

For preprocessing, we lowercase and tokenize the text using NLTK library. [1] The vocabulary considered consists of all word types with frequency greater than one in the training set. All other word types in the training and test set are replaced with a special *unknown* token. Every document is represented as a bag of words with stopwords [2] and out-of-vocabulary words removed. Features were extracted calculating the document-term matrix with TF-IDF weights. Then, we apply Latent Semantic Analysis (LSA) in order to reduce the dimensionality of features to 500.

2.3 Models

We experiment with three models: Random Forest, k-Nearest-Neighbors, and a Multilayer Perceptron (MLP). We use available implementations from the Scikit-Learn library [3] for all experiments.

2.4 Tuning of hyper-parameters

We perform random search of hyper-parameters for 100 iterations. In each iteration, we evalu-

[1] http://www.nltk.org/
[2] Standard stopword list obtained from NLTK
[3] https://scikit-learn.org

ate a model's performance through 5-fold cross-validation over the training set. The MLP was trained using an Adam optimizer (Kingma and Ba, 2014) with L2-regularization. Table 2 presents the hyperparameters and their respective explored ranges, for all models.

3 Results and Discussion

Table 3 presents classification results for all models investigated. We observe that the MLP obtains significantly better performance than the other two models for 5 out of 7 emotion classes, according to micro and macro-averaged F1 score. When looking at the F1 scores per class, for classes in which MLP does better, we observe that the difference between MLP and the runner up model ranges between as litte as 0.41 points (MLP vs RF, *anger* class) up to 23.42 points (MLP vs KNN, *fear* class). For the *sadness* class, RF outperforms MLP by 1.65 points; wheres for the *surprise* class, KNN outperforms MLP by 3.39 points.

It is worth noting the low scores obtained for the class *surprise*, for which the best perfoming model barely obtaines 11.86 F1-score. This could be explained by the lesser amount of labeled samples for this class (just 174 in the test set, see Table 1). In turn, this situation could explain the low recall scores obtained thourghout Table 3.

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 579–586. Association for Computational Linguistics.

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

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. Semeval 2018 task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 24–33.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, 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):e26752.

Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion*, pages 45–60.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

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.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.

Model	Hyper-parameter	Range	Optimal value
Random Forest	Number of estimators	[10–100]	80
	Criterion	[gini, entropy]	entropy
	Maximum depth	[10–100]	38
KNN	Number of neighbors	[5–20]	6
	Weights	[uniform, distance]	distance
Muti-layer Perceptron	Batch size	[20–200]	158
	Learning rate	[1e-5 – 1e-1]	7.50e-05
	L2-regularization parameter (alpha)	[1e-5 – 1.0]	5.99e-05
	Activation function	[tanh, relu]	relu
	Hidden layer sizes	[(150),(100),(50),(10), (100,100),(50,50),(10,10), (100,50),(50,100), (50,50,50),(10,10,10)]	(100,50)

Table 2: Hyper-parameters tuned for each model, alongside the explored ranges and optimal values found for each one.

Emotion class	Random Forest			KNN			MLP		
	P	R	F1	P	R	F1	P	R	F1
Anger	64.84	92.98	76.40	64.03	75.74	69.4	72.99	81.06	**76.81**
Disgust	56.63	31.81	40.73	51.04	35.24	41.69	62.25	53.87	**57.76**
Fear	58.06	5.73	10.43	51.02	15.92	24.27	60.19	39.49	**47.69**
Joy	69.93	31.94	43.85	48.21	52.24	50.14	67.51	55.82	**61.11**
Sadness	63.62	76.26	**69.37**	57.14	41.10	47.81	66.96	68.49	67.72
Surprise	100.00	0.57	1.14	22.58	8.05	**11.86**	53.33	4.60	8.47
Trust	59.76	76.7	67.18	61.72	41.10	49.34	70.81	71.43	**71.12**
Micro avg	62.71	53.53	57.76	55.99	42.80	48.51	67.88	59.68	**63.52**
Macro avg	67.55	45.14	44.16	50.82	38.48	42.07	64.86	53.54	**55.81**

Table 3: Classification results per emotion label for all models investigated. Results are presented in terms of precision (P), recall (R), and F1 score. Best scores are presented in bold.