Multi-Emotion Detection in User-Generated Reviews

Lars Buitinck[1,2], Jesse van Amerongen[2], Ed Tan[2] and Maarten de Rijke[2]

[1] Netherlands eScience Center, Amsterdam, The Netherlands
l.buitinck@esciencecenter.nl
[2] University of Amsterdam, Amsterdam, The Netherlands
j.vanamerongen, e.tan, derijke@uva.nl

Abstract. Expressions of emotion abound in user-generated content, whether it be in blogs, reviews, or on social media. Much work has been devoted to detecting and classifying these emotions, but little of it has acknowledged the fact that emotionally charged text may express multiple emotions at the same time. We describe a new dataset of user-generated movie reviews annotated for emotional expressions, and experimentally validate two algorithms that can detect multiple emotions in each sentence of these reviews.

1 Introduction

The problem of emotion detection in written language has received much attention in recent years, as part of a larger trend toward "affective computing." Few researchers, though, seem to have tried to tackle the full problem of simultaneously detecting emotionally charged phrases and classifying them according to which emotion they mention or express. Instead, attention is usually focused on simple *valence* classification and opinion mining (positive vs. negative, sometimes with the addition of a neutral class) or the classification of utterances that are known to be emotionally charged a priori.

In the present work, we consider the combined problem of detecting and classifying expressions of emotion in the context of movie reviews. This work was borne of basic research into film, using reviews as reflections of the complexity of viewer emotions, but its results may find applications in product search and recommendation for films and other artistic products; e.g., clustering products by emotional charge.

We phrase the problem as *multi-label classification*: we label individual sentences from reviews with a subset (possibly empty) of a predetermined set of emotion labels. Our research question is how to tackle this problem in a supervised way. We contrast two methods that reduce multi-label learning to familiar binary and (disjoint) multi-class classification: one-vs.-rest and an ensemble method that learns from correlations between labels. Both methods use textual features only, where much other research into emotion detection has focused on facial expressions and pitch features in spoken language [4]. Our label set consists of the seven basic emotions identified in the hierarchical cluster analysis of Shaver et al. [12], with the emotion "interest" added [16].

We first survey the state of the art in emotion recognition in Section 2, then discuss a new purpose-built dataset in Section 3. Section 4 contains a description of our feature extraction and learning algorithms, with particular attention to parameter tuning in the multi-label setting. Experimental results are given in Section 5. Section 6 wraps up with conclusions and plans for future research.

2 Related Work

Affective computing has been the focus of much research in the past two decades; a survey of emotion/affect detection in writing, spoken language, and other modalities is given by Calvo and D'Mello [4]. Much of the initial work on written text (e.g., [9]) has focused on valence classification, also known as sentiment analysis or opinion mining, where the two allowed emotions are "positive" and "negative."

Aman and Szpakowicz [2], for example, perform binary classification of sentences in blog posts as emotional/non-emotional. Alm et al. [1] extend the scheme to a three-way classification of sentences as expressing positive, negative, or no emotions. Yang et al. [20] perform classification of blog posts into four categories, "happy," "joy," "sad" and "angry," apparently using the occurrence of certain emoticons as ground truth labels. Their work can be considered to be a finer-grained version of valence detection.

At SEMEVAL 2007 [14], various systems where benchmarked on the task of classifying news headlines according to a six-label annotation scheme, viz. anger, disgust, fear, joy, sadness and surprise. The focus was on unsupervised methods; Strapparava and Mihalcea [15] additionally tested a weakly supervised transfer learning approach.

Closer to our work is that of Danisman and Alpkocak [5], who perform supervised learning of emotion labels at the sentence/snippet level. They show that a simple nearest centroid classifier using bag-of-words features and tf–idf weighting can achieve an F_1 score of 32.22% in a five-way multiclass prediction problem using a set of 7,666 text snippets. D'Mello et al. [6] achieve higher scores, but in a problem that only involves three emotional states. Neither of these works takes into account a neutral state.

Our work differs from the work listed above in the following important ways. First, we do not make the simplifying assumption that emotional states are mutually exclusive. Second, while we use supervised learning and manual annotation, we use only a small labeled training set of a few hundred sentences, where earlier attempts have typically used thousands of training samples.

3 Dataset

We hand-labeled 44 movie reviews using the BRAT annotation interface [13], identifying emotionally charged phrases. The reviews were taken from IMDB and concern the films *American History X*, *The Bourne Identity*, *Earth* (2007), *The Godfather*, *Little Miss Sunshine*, *The Notebook*, *SAW*, and *Se7en*; all Hollywood productions, but of varying genres. Each film is covered by six reviews, except for *The Godfather* (two reviews, due to time constraints).[3]

We perform sentence splitting on each of the reviews, and turn the problem into a multi-label classification problem by assigning to each sentence the set of labels used to label any string of words within the sentence. Doing so yielded 629 sentences containing 13,409 tokens, distributed over the various films as shown in Table 1, with the label distribution given in Table 2.

Of the 629 sentences, 420 have at least one label, showing how prevalent the expression of emotions in film reviews is. The average number of labels per sentence is 0.887,

[3] `https://github.com/NLeSC/spudisc-emotion-classification`

while the maximum is five (the combination "Joy–Sadness–Love–Interest–Surprise," which occurs once). We reserve roughly 20% of our sentences as a test set, using the remainder for classifier training and tuning. Because the "Disgust/contempt" label has only six samples, we replace it with "Anger."

Table 1. Samples per film.

Title	Sent.	Emot.
American History X	77	63
The Bourne Identity	90	41
Earth	63	45
The Godfather	18	18
Little Miss Sunshine	95	51
The Notebook	107	73
SAW	65	54
Se7en	114	75

Table 2. Absolute label frequencies.

Label	All	Test
Anger	18 (24)	6
Disgust/contempt	6 (0)	–
Fear	37	11
Interest	69	20
Joy	47	9
Love	272	48
Sadness	35	10
Surprise	80	16

4 Classification Algorithms

We tested two algorithms for performing multi-label classification. Both use standard bag-of-words features with stop word removal and optional tf–idf weighting, and reduce the multi-label problem to either binary or multiclass learning, for which we use linear support vector machines. We implement these using scikit-learn [3, 10], which includes the linear SVM learner of Fan et al. [7].

4.1 Reduction to Binary Classifiers

The first algorithm we consider reduces the K-way multi-label classification problem to K independent binary SVMs that learn to distinguish one emotion from all others. This is variously called the *one-vs.-rest*, or *binary relevance* reduction [18]. While this problem reduction cannot take advantage of correlations between labels, it has the advantage that we can separately tune the settings of each SVM, so that we end up with an optimal model for each binary sub-problem.

I.e., for each label separately, we do a parameter sweep and select the parameter settings that result in the maximum F_1 score for that label according to five-fold stratified cross-validation on the training set. We try all parameter settings in the grid defined by $C \in \{.1, 1, 10, 100, 1000\}$, L_1 or L_2 regularization, linear or logarithmic tf, whether to use tf or tf–idf, and whether to oversample the minority class in each sub-problem. These settings were chosen based on experience with other text classification problems.

4.2 Learning from Label Dependencies

As an alternative to the one-vs.-rest reduction just sketched, we also benchmark the random k-labelsets (RAKEL) algorithm [17, 19]. To understand this method, we must

first introduce an alternative problem reduction strategy for multi-label classification, the *label powerset* method. A label powerset model is a regular classifier trained using all subsets of a multi-label problem's set of labels as its classes, so in our problem, the triple (Fear, Love, Surprise) would be one class. This method is very powerful in that it can learn dependencies between labels, but it requires solving an exponential-sized multiclass problem.

To prevent this combinatorial blowup, RAKEL builds an ensemble of label powerset classifiers, each trained on a subset of labels of fixed size k, chosen at random without replacement. Prediction proceeds by a voting scheme: for each randomly generated subset J of all labels, its associated classifier predicts a subset $f(x) \subseteq J$ to which sample x should belong. Each $j \in f(x)$ gets a positive vote; each $j \notin f(x)$ a negative vote. A positive tally for a label means a positive prediction in the full multi-label problem. We use the RAKEL algorithm with linear SVMs as its base learners.

A problem with RAKEL is that it is not clear how to tune its parameters with a small amount of training data. We might like to apply the same tuning as for the one-vs.-rest strategy, i.e., optimize each base learner separately before combining them; but this is infeasible, because the label powerset classifiers must solve overly sparse sub-problems. Some label subsets, such as (Interest, Love, Sadness), occur only once in the training set, making proper stratification impossible. Fitting multiple RAKEL ensembles in a stratified CV setting may be possible with the multi-label stratification strategy of Sechidis et al. [11], but time constraints prevented us from implementing it. We therefore use tf–idf weighting with logarithmic tf, automatic oversampling, and a fixed regularization parameter $C = 1$ for all SVMs.

We let $k = 3$ be the size of the label subsets in RAKEL, which has the effect of undoing the randomization: only 35 size-k subsets of our label set occur in the training set, so we can simply fit a classifier to each of them.

5 Results

Our main research question is to find out how a relatively simple but carefully tuned one-vs.-rest baseline compares against a more advanced multi-label classification method on the task of emotion classification. To answer this question, we empirically evaluate the algorithms from the previous section on the dataset described in Sect. 3.

We report accuracy and F_1 scores per class and averaged over all classes. We compute the overall accuracy score as defined by Godbole and Sarawagi [8], i.e., one minus the Hamming loss. Since accuracy has the problem of overestimating performance in highly-unbalanced classification problems, we consider F_1 score to be our main evaluation metric. All scores are averaged over ten runs of each training algorithm to account for the randomization in both; in the case of RAKEL, the results of all runs achieved the exact same scores despite randomization in the SVM learner [7].

Our main results are shown in Table 3. We see that RAKEL achieves slightly, but significantly, better overall F_1 score. Because its parameters are fixed, it also achieves this result noticeably faster than OvR: the expensive tuning of OvR takes many minutes of computing time, whereas RAKEL finishes in mere seconds.

Table 3. Sentence-level accuracy and F_1 score for one-vs.-rest (OvR) and RAKEL. Differences in F_1 score between the two algorithms were tested using Welch's one-sided t-test. $\triangle$: significantly better at $\alpha = .05$, ▲: significantly better at $\alpha = .001$, or consistently better with zero variance.

	Algorithm/performance metric			
	OvR accuracy	OvR F_1 score	RAKEL acc.	RAKEL F_1
Anger	$.940 \pm .018$	$.105 \pm .129^{\triangle}$	.937	.000
Fear	$.910 \pm .015$	$.267 \pm .039$	.921	$.546^{\blacktriangle}$
Interest	$.802 \pm .000$	$.359 \pm .000^{\blacktriangle}$	.818	.343
Joy	$.939 \pm .005$	$.494 \pm .056$	.929	.471
Love	$.706 \pm .000$	$.626 \pm .000^{\blacktriangle}$	.675	.586
Sadness	$.849 \pm .000$	$.296 \pm .000$	.905	.400
Surprise	$.740 \pm .051$	$.231 \pm .029$	.794	$.278^{\blacktriangle}$
Overall	$.841 \pm .007$	$.432 \pm .008$	.854	$.456^{\blacktriangle}$

However, RAKEL is not superior on all labels, and in particular does not learn to predict the "Anger" label at all. The OvR learner similarly shows difficulty with this label, achieving $F_1 \geq .25$ in four runs, but zero in the remaining six. Inspection of the dataset indicates that the problem with the "Anger" label is that it is often used to mark disappointment or criticism, and reviews tend to express this disappointment in a subtle and indirect way. Words like "frustrated" or "contrived" are rare, and reviewers may express their disappointment by praising a movie that they preferred over the one being reviewed, using a positive register of expression.

6 Conclusion

We have shown how the problem of emotion detection and classification at the sentence level can viably be tackled as one of supervised classification, even with relatively small labeled datasets, using standard bag-of-words features and while allowing for multiple emotion labels per sentence. We have shown that careful tuning of a baseline method can make it almost as strong as the more advanced RAKEL algorithm; tuning of RAKEL is an interesting problem that requires further attention.

In future work, we intend to further classify emotionally charged utterances according to the trigger of the emotion: either a film regarded as an artifact, or the content (storyline) of the film. E.g., we intend to automatically determine whether anger is caused by a bad performance on the part of actors or directors, or by a good performance that evokes genuine anger at the "bad guy" in the plot. This should decouple emotion from opinion, and provide further insight into the emotional response that films evoke.

Acknowledgements. This research was partially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 312827 (VOX-Pol), the Netherlands Organisation for Scientific Research (NWO) under project nrs 727.011.005, 612.001.116, HOR-11-10, 640.006.013, 612.066.930, the Center for Creation, Content and Technology (CCCT), the Dutch national program COM-

MIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), the Netherlands eScience Center under project number 027.012.105, the Yahoo! Faculty Research and Engagement Program, the Microsoft Research PhD program, and the HPC Fund.

References

[1] C. O. Alm, D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proc. HLT–EMNLP*, pages 579–586, 2005.

[2] S. Aman and S. Szpakowicz. Identifying expressions of emotion in text. In *Text, Speech and Dialogue*, volume 4629 of *LNCS*, pages 196–205. Springer, 2007.

[3] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Müller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. Vanderplas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop on Languages for Machine Learning*, 2013.

[4] R. A. Calvo and S. K. D'Mello. Affect detection: an interdisciplinary review of models, methods, and their applications. *IEEE Trans. on Affective Computing*, 1(1):18–37, 2010.

[5] T. Danisman and A. Alpkocak. Feeler: emotion classification of text using vector space model. In *Proc. AISB Convention*, 2008.

[6] S. K. D'Mello, S. D. Craig, J. Sullins, and A. C. Graesser. Predicting affective states expressed through an emote-aloud procedure from AutoTutor's mixed-initiative dialogue. *Int'l J. AI in Education*, 16:3–28, 2006.

[7] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

[8] S. Godbole and S. Sarawagi. Discriminative methods for multi-labeled classification. In *Proc. Pacific-Asia Conf. on KDD*, 2004.

[9] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity. In *Proc. ACL*, 2004.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *JMLR*, 12, 2011.

[11] K. Sechidis, G. Tsoumakas, and I. Vlahavas. On the stratification of multi-label data. In *Proc. ECML PKDD*, pages 145–158, 2011.

[12] P. Shaver, J. Schwartz, D. Kirson, and C. O'Connor. Emotion knowledge: further exploration of a prototype approach. *J. Personality and Social Psychology*, 52(6), 1987.

[13] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. BRAT: a web-based tool for NLP-assisted text annotation. In *Demos at 13th Conf. EACL*, pages 102–107, 2012.

[14] C. Strapparava and R. Mihalcea. SemEval-2007 task 14: Affective text. In *Proc. 4th Int'l Workshop on Semantic Evaluations*, pages 70–74, 2007.

[15] C. Strapparava and R. Mihalcea. Learning to identify emotions in text. In *Proc. ACM Symp. Applied Computing*, pages 1556–1560, 2008.

[16] E. Tan. *Emotion and the structure of narrative film*. Erlbaum, Mahwah (NJ), 1996.

[17] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas. Multi-label classification of music into emotions. In *Proc. Int'l Conf. on Music IR*, pages 325–330, 2008.

[18] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *Int'l J. Data Warehousing and Mining*, 3(3):1–13, 2007.

[19] G. Tsoumakas and I. Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. In *ECML*, pages 406–417. Springer, 2007.

[20] C. Yang, K. H.-Y. Lin, and H.-H. Chen. Emotion classification using web blog corpora. In *IEEE/WIC/ACM Int'l Conf. on Web Intelligence*, pages 275–278, 2007.