Confused and Thankful: Multi-label Sentiment Classification of Health Forums

Victoria Bobicev[1] and Marina Sokolova[2,3(✉)]

[1] Technical University of Moldova, Chisinau, Moldova
victoria.bobicev@ia.utm.md
[2] IBDA, Dalhousie University, Halifax, Canada
sokolova@uottawa.ca
[3] University of Ottawa, Ottawa, Canada

Abstract. Our current work studies sentiment representation in messages posted on health forums. We analyze 11 sentiment representations in a framework of multi-label learning. We use **Exact Match** and **F-score** to compare effectiveness of those representations in sentiment classification of a message. Our empirical results show that feature selection can significantly improve **Exact Match** of the multi-label sentiment classification (paired t-test, P = 0.0024).

Keywords: Sentiment classification · Multi-label learning · Medical forums

1 Motivation

Separation of sentiments is a major challenge in sentiment classification. Due to a *yes-no* approach which assigns a text with one label and one label only, single label learning algorithms thrive and succeed when sentiment classes are easily dichotomized. At the same time, even short texts can combine various sentiments and objective, factual information, e.g. *my oldest had his th bday today & he had the stomach flu it still was a nice day I even got to spend some special time whim & hubby*. Overlap in sentiments can hardly be resolved by single-label binary or multiclass classification. We hypothesize that annotating texts with ≥ 2 sentiment labels and applying multi-label classification can benefit our understanding of the text sentiments. Applied to online health forums, multi-label sentiment classification improves understanding of patients' needs and can be used in advancing patient-centered health care (Bobicev, 2016; Liu and Chen, 2015; Melzi et al. 2014).

Online health forums allow for studies of well-being and behavior patterns in uncontrolled environment (Aarts et al. 2015; Navindgi et al. 2016; Hidalgo et al. 2015). Giving and receiving emotional support has positive effects on emotional well-being for patients with higher emotional communication, while the same exchanges have detrimental impacts on emotional well-being for those with lower emotional communication competence (Yoo et al. 2014). It has been shown that positive emotions present more frequently in responding posts than in the posts initiating new discussions (Yu, 2011).

In this study, we analyze how 6 score-based, 4 multi-dimensional and 1 domain-based sentiment representations affect accuracy of multi-label sentiment classification of

M. Mouhoub and P. Langlais (Eds.): Canadian AI 2017, LNAI 10233, pp. 284–289, 2017.
DOI: 10.1007/978-3-319-57351-9_33

message posted on a health forum. Problem transformations (Binary Relevance and Bayesian Classification Chains) and classification algorithms (SVM, Naïve Bayes and Bayesian Nets) assess effectiveness of the sentiment representations. Our results show that feature selection can significantly improve **Exact Match** of sentiment classification (paired t-test, P = 0.0024).

2 Multi-label Data Annotation and Sentiment Representation

We have worked with 80 discussions, 10 – 20 posts each, obtained from the InVitroFertilization forum (www.ivf.ca); we had 1321 messages. The length of forum messages was 126 words on average. The target labels were *confusion*, *encouragement*, *gratitude* and *facts*; those labels were previously used in multi-class classification of the data (Sokolova and Bobicev, 2013). Three annotators independently worked with each post; each annotator assigned a post with one label. From 1321 posts, 658 posts had three identical labels; 605 posts had two identical labels, and 58 posts had three different labels. Note that multi-label learning algorithms automatically resolve difference in the number of assigned labels. When we account per classification category, 954 posts had the label *facts,* 642 posts – *encouragement*, *confusion* appeared in 285 posts, and *gratitude* appears in 161 posts.[1]We kept the assigned labels in classification experiments. Fleiss Kappa = 0.48 indicated a moderate agreement, comparable with three-label sentiment annotation of health messages (Melzi et al. 2014).

We used 11 sentiment lexicons to extract sentiment information from our texts: *SentiWordNet* **(SWN)**, *Bing Liu Sentiment Lexicon* (**BL**), *SentiStrength* **(SS),** ***AFINN*** *Hashtag Affirmative and Negated Context Sentiment Lexicon* **(HANCSL),** *Sentiment 140 Lexicon* **(140SL)** assign terms with polarity scores; ***MPQA*** *DepecheMood* **(DM)**, *Word-Emotion Association Lexicon* **(WEAL),** *General Inquirer* **(GI)** assign terms with multiple sentiment categories, and *HealthAffect* **(HA)** uses Point-wise Mutual Information to retrieve emotional scores (Sokolova and Bobicev, 2013). Among the emotional terms retrieved from the data, 6 terms appears in the 11 lexicons: *encouragement, horrible, negative, stupid, success, successful*, 2650 terms - in two lexicons, 928 terms - in three lexicons, and 3963 terms appear in one of the lexicons.

3 Empirical Evaluation

Multi-label classification allows an example to be simultaneously associated with >1 label (Trohidis, and Tsoumakas, 2007). In practice, multi-label classification can be transformed into ensemble of binary classification tasks. We applied two transformation methods: Binary Relevance (BR) and Bayesian Classifier Chains (BCC)[2]. We use **Exact Match** in performance evaluation (Sorower, 2010):

[1] The data set is available upon request at victoria.bobicev@ia.utm.md.

[2] http://meka.sourceforge.net/.

$$ExactMatch = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = Z_i) \quad (1)$$

Where n denotes the number of texts in the data set, Y_i, Z_i are sets of predicted and true labels for text i respectively. We compute a balanced **F-score** to evaluate classification of each label categories. We used the MEKA toolkit (Read, et al. 2016). SVM, Naïve Bayes and Bayesian Nets were the base classifiers; 10-fold cross-validation was used for model selection. To put our results in perspective, we compute the majority class baseline; text representation by concatenating the 11 lexicons provides the benchmark accuracy.

The 11 lexicons assessed sentiments through different schema; hence, we worked with 11 different sentiment representations. The highest **Exact Match** was obtained with 1131 terms extracted from *SentiStrength* (SS) (see Table 1). Although every **Exact Match** significantly beats the baseline, none of the lexicons provided for significantly better results. Similarly, non-significant improvement happens for the best *per category* **F-score** (Table 2).

Table 1. The best **Exact Match** on individual lexicons; the majority class ExactMatch = 0.270;

N	Features	Retrieved terms	Exact Match	Classifier
1	SWN	3 725	0.395	BR- NB Multinomial
2	SS	1 131	**0.410**	BCC-SVM
3	DM	4 467	0.407	BR-NB Multinomial
4	HA	1 190	0.403	BR-NB Multinomial
5	AFINN	793	0.399	BCC-SVM
6	GI	942	0.378	BCC-SVM
7	HANCSL	2 765	0.357	BCC-SVM
8	140SL	2 160	0.376	BCC-SVM
9	WEAL	1 368	0.335	BR-NB Multinomial
10	BL	1 103	0.362	BCC-SVM
11	MPQA	1 417	0.388	BCC-SVM
12	*All the 11 lexicons*	*9 086*	***0.450***	*BR-NB Multinomial*

Table 2. The best **F-score** obtained for each category; we use the majority class baseline;

Category	Baseline	F-score	Feature	Classifier
Confusion	0.784	0.802	SWN	BR NB Multinomial
Encouragement	0.486	0.731	HA	BR NB Multinomial
Gratitude	0.878	0.907	BL	BCC- SVM
Facts	0.722	0.805	DM	BR NB Multinomial

On the next step, we assessed whether reducing non-essential information can help in classification accuracy (Tables 3 and 4). To remove less contributing features, we applied three feature selection methods: CfsSubset (best subsets), Classifier SubsetEval,

Table 3. The best Exact Match obtained on the combinations of the selected feature sets.

N	Feature set	N of terms	Exact Match	Classifier
1	Selected attributes from all 11 lexicons (best subsets)	1009	**0.544**	BR-NB Multinomial
2	Selected attributes from all 11 lexicons (best subsets for SMV)	1446	0.521	BR-NB Multinomial
3	Selected attributes from all 11 lexicons (InfoGain)	2072	0.534	BR-NB Multinomial

Table 4. The best **F-score** obtained on combinations of the selected feature sets.

Category	F-score	Feature set	Classifier
Confusion	0.870	best subsets for SVM	BR-NB Multinomial
Encouragement	0.805	best subsets	BR-NB Multinomial
Gratitude	0.930	InfoGain	BR-NB Multinomial
Facts	0.833	InfoGain	BCC- SVM

and InfoGain. For each method, we applied feature selection to each lexicon and each label; then those 11 × 4 = 44 sets were concatenated; we removed all duplicate terms. We obtained the best **Exact Match = 0.544** on 1009 terms: 301 terms with positive scores, 200 - with negative scores, 249 - from HA; other 259 terms had multiple emotional indicators.

We computed a conservative *paired* t-test between three Exact Match results reported in Table 3 and the highest three Exact Match from Table 1, i.e., rows 2, 3 and 6. T-test's P = 0.0024 indicates that feature selection significantly increased examples with fully correctly identified labels. Although feature selection did not significantly improved **F-score** (*paired* t-test, P = 0.3245), it did improve classification for each category, esp. for *encouragement* where increase was >10%.

4 Discussion of Sentiment Representations

As expected, emotionally charged adjectives are frequent among the selected features, e.g., *amazing, awful, bad, desperate, excited.* At the same time, polarity of the selected terms has a nuanced relationship with the expressed sentiments. For every category, selected features contain words with positive and negative connotation: the best **F-score** for *confusion* was obtained with representation containing 425 terms with positive scores and 333 terms - with negative scores, for *gratitude* - on representation containing 583 terms with positive scores and 323 terms with negative scores, for *encouragement* - on representation containing 301 with positive scores and200 terms with negative scores, and for *facts* - on representation containing 583 terms with positive scores and323 terms with negative scores. This can be attributed to a sentiment

flow typical to health forum posts: empathy (positive polarity), followed by reference to interlocutors' problems (negative polarity), followed by good wishes (positive polarity).

Many selected terms appear in several lexicons, e.g., *lovely, progress, exciting, fearful, hopeless, luck, worse* appeared in 8–10 lexicons of the discussed 11; *lovely, hopeless* appeared in all the lexicons but HA; *progress* - in all the lexicons but SS; *worse*- in all the lexicons but HA and HANCSL. Also, no sentiment representation was left behind: for each category, selected terms represented almost every lexicon. Some terms were repeatedly selected for several categories. For example, *luck* was selected for *encouragement* and for *gratitude*; *good* was selected for *facts* and *confusion*.

5 Conclusions and Future Work

In this work we have studied effects of sentiment representation on sentiment classification of a message posted on a health forum. We used a framework of Multi-label Learning as many messages convey >1 sentiment. We have analyzed 11 sentiment representations: 6 score-based, 4 multi-dimensional and 1 domain-based. We applied Exact Match to evaluate usefulness of the sentiment representations. Counting only examples with fully correctly identified labels (i.e., examples with partially identified labels were discarded), we found that redundancy reduction through feature selection significantly improves classification (paired t-test, $P = 0.0024$).

Using **F-score** to find the most effective sentiment representations of each category, we observed that both positive and negative polarity *within* message text play an important role in correct identification of the message sentiment. Those results hold for *encouragement, gratitude* (aka positive sentiments), *confusion* (a substitute for the negative sentiment), and *facts*. For the label *facts*, which we considered a non-sentimental category, the highest **F-score** appeared on representation containing terms with high polarity scores. Co-occurrence of opposite polarities shows complexity of sentiment conveyance and supports multi-label sentiment classification.

In future, we plan to work with finer grained sentiment representations. One venue would be to explore relations between polarity strength and accuracy of sentiment classification in a message. Another promising venue is to apply discourse analysis to investigate the use of sentiment-bearing words in factual messages.

References

Aarts, J., Faber, M., Cohlen, B., van Oers, A., Nelen, W., Kremer, J.: Lessons learned from the implementation of an online infertility community into an IVF clinic's daily practice. Hum. Fertil. **18**(4), 238–247 (2015)

Bobicev, V.: Text classification: the case of multiple labels. In: 2016 International Conference on Communications (COMM), pp. 39–42 (2016)

Liu, S.M., Chen, J.H.: A multi-label classification based approach for sentiment classification. Expert Syst. Appl. **42**(3), 1083–1093 (2015)

Melzi, S., Abdaoui, A., Aze, J., Bringay, S., Poncelet, P., et al.: Patient's rationale: patient knowledge retrieval from health forums. In: 6th International Conference on eTELEMED: eHealth, Telemedicine, and Social Medicine (2014)

Navindgi, A., Brun, C., Boulard, S., Nowson, S.: Steps Toward Automatic Understanding of the Function Of Affective Language in Support Groups. NLP for Social Media (2016)

Read, J., Reutemann, P., Pfahringer, B., Holmes, G.: MEKA: a multi-label/multi-target extension to Weka. J. Mach. Learn. Res. **21**(17), 1–5 (2016)

Rodríguez Hidalgo, C.T., Tan, E.S.H., Verlegh, P.W.J.: The social sharing of emotion (SSE) in online social networks: a case study in live journal. Comput. Hum. Behav. **52**, 364–372 (2015)

Sokolova, M., Bobicev, V.: What sentiments can be found in medical forums? In: Proceedings of RANLP 2013, pp. 633–639 (2013)

Sorower, M.S.: A literature survey on algorithms for multi-label learning. Technical report, Oregon State University, Corvallis (2010)

Trohidis, K., Tsoumakas, G.: Multilabel classification: an overview. Int. J. Data Warehouse. Min. **3**, 1–13 (2007)

Yoo, W., Namkoong, K., Choi, M., et al.: Giving and receiving emotional support online: communication competence as a moderator of psychosocial benefits for women with breast cancer. Comput. Hum. Behav. **30**, 13–22 (2014)

Yu, B.: The emotional world of health online communities. In: Proceedings of the 2011 iConference, pp. 806–807, New York, USA (2011)