No Sentiment is an Island

Sentiment Classification on Medical Forums

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

[1] Technical University of Moldova, Chişinău, Moldova
vika@rol.md
[2] Institute for Big Data Analytics, Halifax, Canada
sokolova@uottawa.ca
[3] University of Ottawa, Ottawa, Canada

Abstract. In this study we propose a new method to classify sentiments in messages posted on online forums. Traditionally, sentiment classification relies on analysis of emotionally-charged words and discourse units found in the classified text. In coherent online discussions, however, messages' non-lexical meta-information can be sufficient to achieve reliable classification results. Our empirical evidence is obtained through multi-class classification of messages posted on a medical forum.

1 Motivation

A rapid growth in the Internet access from 70 % of the population in 2010 to 81 % in 2014 has caused an increase in online networking from 38 % of the population in 2011 to 46 % in 2014[1]. European Commission's strategy on Big Data (July, 2014) highlights that "Data is at the centre of the future knowledge economy and society"… and that to seize the opportunities of the large and complex resulting datasets, and be able to process such 'big data', initiative must be supported e.g. in the health sector (personalized medicine). Health-care of the future will be based on community, collaboration, self-caring, co-creation and co-production using technologies delivered via the Web (Cambria et al. 2012)."

Online medical forums are platforms on which interested parties (e.g., patients, family members) collaborate for better health. The best forums promote empowerment of patients and improve quality of life for individuals facing health-related problems. An online survey of 340 participants of HIV/AIDS-related Online Support Groups revealed four most important factors that contribute to the patient empowerment: receiving social support, receiving useful information, finding positive meaning and helping others (Mo and Coulson 2012). On surveyed medical forums, personal testimonials attract attention of up to 49 % of the participants, whereas 25 % of the participants are motivated by

---

[1] http://ec.europa.eu/eurostat/.

scientific and practical content (Balicco and Paganelli 2011). In a survey of online infer-
tility support groups, empathy and shared personal experience constituted 45.5 % of
content, gratitude – 12.5 %, recognized friendship with other members – 9.9 %, whereas
the provision of information and advice and requests for information or advice took
15.9 % and 6.8 % respectfully (Malik and Coulson 2010). In many testimonials, infor-
mative content intervenes with emotions, e.g. *For a very long time I've had a problem
with feeling really awful when I try to get up in the morning* ties up the author's poor
feeling and her daily routine.

Restricted communication environment of online support groups can amplify rela-
tions between communication competence and emotional well-being, especially for
patients diagnosed with potentially life-threatening diseases (Shaw et al. 2008). A study
of 236 breast-cancer patient posting online showed that quality of life and psychological
concerns can be affected in both desired and undesired ways. Giving and receiving
emotional support has positive effects on emotional well-being for breast cancer patients
with higher emotional communication skills, while the same exchanges have detrimental
impacts on emotional well-being for those with a lower emotional communication
competence (Yoo et al. 2014). Challenges arise, however, when sentiments should be
analyzed in a large data set: traditional tools, e.g. general-purpose emotional lexicons,
are not efficient on medical forums, whereas domain-specific lexicons tend to over-fit
the data (Bobicev et al. 2015a, 2015b).

Our current work proposes that coherent online discussions allow classification
of sentiments by using information of the post's position in the discussion, senti-
ments of the neighboring posts and the author's activity level. Further, we test this
approach in multi-class sentiment classification of data gathered from an online
medical forum.

2   Related Work

Strong relationship exists between language of an individual and her health status
(Rhodewalt 1984). Language expressions of negative and undesirable events can be
predictors of cardio-vascular disease risks. This connection has led to the development
of Linguistic Inquiry and Word Count (LIWC) (Pennebaker and Francis 2001). A soft-
ware program calculates the degree to which people use different categories of words
across a wide array of texts, including emails, speeches, poems, or transcribed daily
speech. This tool automatically determines the degree to which any text uses positive
and negative emotions, self-references, cognitive and social words.

Qiu et al. (2011) studied dynamics among positive and negative sentiments
expressed on Cancer Survivors Network. They estimated that 75 %–85 % of the forum
participants change their sentiment in a positive direction through online interactions
with other community members.

The sheer volume of on-line messages commands the use of Sentiment Analysis to analyse emotions en masse. Taking advantage of Machine Learning technique, Sentiment Analysis has made considerable progress when applied population health (Chee et al. 2009) as well as on social networks (Zafarani et al. 2010). Empirical evidence shows a strong performance of Naive Bayes, K-Nearest Neighbor, Support Vector Machines, as well as scoring functions and sentiment-orientation methods that use Point-wise Mutual Information (Liu and Zhang 2012). Sentiment Analysis studies mostly identify text's sentiment through the text vocabulary (e.g., positive and negative adjective, positive and negative adverbs) and style (e.g., use of negations, modal verbs) (Taboada et al. 2011).

The In Vitro Fertilization (IVF) data set has been introduced in (Sokolova and Bobicev 2013). The data consists of 80 annotated discussions (1321 posts) gathered from the IVF Ages 35+ sub-forum[2]. Each post was annotated by two raters using three sentiment categories: 'confusion', 'encouragement' and 'gratitude', and one 'factual' category, a category transitional between 'factual' and 'encouragement' was named 'endorsement'. Each post was assigned with one of the labels: 'confusion' (117 posts), 'encouragement' (310 posts), 'gratitude' (124 posts), 'factual' (433), 'endorsement' (162 posts), and 176 'ambiguous' posts on which annotators disagreed. The annotators reached a strong agreement with Fleiss Kappa = 0.737. A detailed description of the manual annotation process can be found in (Sokolova and Bobicev 2013). Previously, sentiments transitions in this data had been studied by applying a domain-specific lexicon HealthAffect and a general-purpose emotional lexicon SentiWordNet (Bobicev et al. 2015a; Bobicev et al. 2015b).

In the current work, we, however, hypothesize that texts related in their content and context can be efficiently classified into sentiment categories without invoking vocabulary of these texts.

3   Problem Statement

We observed a certain pattern of sentiments transactions within discussions: in the first message, the author who started the discussion usually requested help with finding information or emotional help (*confusion* accounted for 56 % of the initial posts). The following posts were either with encouragement (24 %) or provided the factual information requested by the first author (30 %). In many cases, the discussion initiator either updated the interlocutors on the factual progress (39 %) or expressed gratitude for their helpful comments (33 %) (Bobicev et al. 2015b).

The following discussion exemplifies the discussion flow:

---

[2] http://ivf.ca/forums/forum/166-ivf-ages-35/.

post_id_140964 Hi Everyone,I am only five days past my ET and I'm tempted to do a pregnancy home test by next Wednesday. How many of you did home tests or were you patient enough to complete the 2WW without testing at home? I know most of you have mentioned how long the wait seemed or was. …

post_id_140968 I did do a HPT, first one was 6dp3dt and it was negative, then I did another one 8dp3dt and it was very faintly positive....and here we are with a happy healthy 2 month old! it's totally up to you whether you test, everyone has their opinions! good luck! let us know what you decide!

post_id_140971 I always test too! I just can't help it. I get too anxious and I would much rather know before beta day. …

post_id_140995 Its really hard to patient and not poas. I just wanted to mention that if you test too early and its negative it can really devasting! The wait is brutal but so is a negative. I tested early also and it was so hard to deal with. You have to decide what you can handle. Best of luck with your cycle.

post_id_141010 I have to admit that I was a chicken when it came to the poas. I never did. I was not sure I could handle it. I just waited it out. You will have to go with your gut. Good luck with whatever you decide.

post_id_141032 I believe that you should do what is right for you. I know some people like to wait and they have the patience of a saint! I know as long as you can keep yourself from being too bummed out if it is a BFN then go for it. There are many people who can and some who can't. I do know that no matter what you will eventually do what your mind wants to do.   It is just human nature. The urge is SO strong. If you do test early, GOOD LUCK!

post_id_141036 Wow! Thanks to all of you for the info and encouragement. Now I don't have to worry about that anymore and I'll certainly buy First Response and do my testing. Again, thanks and all the best to all of you too in every step of this fertility journey.

Working on sentiment analysis *sans* vocabulary content, we pursued the following goals and connected them with feature sets representing the messages:

(A) Our first goal was to demonstrate that there are patterns of sentiments in forum's discussions and they mutually influence each other. Hence, we built a representation which reflected sentiment transitions in discussions. Having two annotation labels for each post we decided to use them both as features rather than merge them. This allowed us to disambiguate the *ambiguous* label, which appeared when two annotators selected different sentiment labels for the post. We then represented each post through the two labels assigned by each annotator to the previous post and two labels assigned by each annotator to the following post; posts lacking this information (e.g., the first post in discussion) were assigned a label "none" (Set I - 4 categorical features).

(B) We then concentrated on the position of the posts within the discussion, as its position can affect the expressed sentiments.

We built three binary features showing whether the previous, current and next messages are first, middle, or last ones. We used these features to enhance the previous representation (Set II – 4 categorical features + 3 binary features = 7 features).

(C) We were interested in the impact of the longer sequences of sentiment transitions on the post's sentiment. To assess this impact, we represented the post by four labels assigned by each annotator to the two previous messages and by four labels assigned by

each annotator to the two following messages (Set III - 8 categorical features). To investigate whether this information can be enhanced by the post's position, we expanded Set III with the three position features of the post (Set IV – 8 categorical + 3 binary features = 11 features).

(D) Next, we aimed to represent the influence of author's activity on the post sentiments. We built three features to present the post author's activity: a binary feature pr showing whether the author belongs to the most active authors of this forum (aka a prolific author); a binary feature i indicating whether the author of the post is the one who started this discussion; a binary feature f indicating whether the author posted in this discussion for the first time. Note that these features are independent and can simultaneously be true.

To investigate the mitigating impact of the author's activity, we enhanced the post representation through Set IV by all the three features (Set V – 11 features + 3 features = 14 features) and by each feature separately (Set VI – 11 features + pr = 12 features, Set VII – 11 features + f = 12 features, Set VIII – 11 features + n = 12 features).

Note that all the 12 feature sets omit references to the content of the post they represent.

For multi-class classification, we apply Support Vector Machines (SVM, the logistic model, normalized poly kernel, WEKA toolkit) and Conditional Random Fields (CRF, the default model, Mallet toolkit). SVM has shown a reliable performance in sentiment analysis of social networks. At the same time, we expect CRF to benefit from the feature sets that are sequences of mutually dependent random variables.

4   Empirical Evidence

We worked on four multi-class classification tasks:

6-class classification where 1322 posts are classified into confusion, encouragement, endorsement, gratitude, facts, ambiguous; the majority class F-score = 0.162;

5-class classification where the ambiguous class is removed and remaining 1146 posts are classified into the other 5 classes; the majority class F-score = 0.207;

4-class classification where 1322 posts are grouped as following: facts and endorsement classes make up a (factual) class, encouragement and gratitude classes become a positive class, and confusion and ambiguous classes remain; the majority class F-score = 0.280;

3-class classification where 176 ambiguous messages are removed and the remaining 1146 messages are classified in positive, confusion and factual as in 4-class classification; the majority class F-score = 0.355.

The best classifiers were found by 10-fold cross-validation. We calculated the macro-average F-score. Table 1 reports SVM's performance for each task, and Table 2 – on CRF. The feature sets are the same as in Sect. 3.

Analyzing the results of SVM, we notice that the best F-score is consistently obtained when the feature set conveys all the three aspects of the author's activity (Set V). The impact of the activity attributes is especially noticeable when we compare the results with those obtained on Set IV for 5-, 4-, and 3-class tasks: F-score = 0.448, F-score = 0.495, F-score = 0.594 respectively.

Table 1. Classification results for SVM. For each task, the highest F-score is in **bold**, the lowest F-score – in *italics*.

Sets	Feat.	6-class			5-class			4-class			3-class		
		P	R	F	P	R	F	P	R	F	P	R	F
I	4	0.621	0.605	**0.613**	0.619	0.614	0.616	0.681	0.652	**0.665**	0.732	0.704	**0.717**
II	7	0.613	0.602	0.607	0.616	0.610	0.613	0.654	0.635	0.644	0.711	0.711	0.711
III	8	0.516	0.490	0.502	0.580	0.540	0.558	0.570	0.524	0.545	0.669	0.609	0.637
IV	11	0.516	0.487	0.500	0.566	0.535	0.548	0.568	0.528	0.546	0.671	0.612	0.640
V	14	0.438	0.436	0.436	0.507	0.500	*0.503*	0.489	0.480	*0.483*	0.640	0.625	0.631
VI	12	0.437	0.423	0.429	0.511	0.516	0.513	0.500	0.482	0.489	0.627	0.610	*0.617*
VII	12	0.569	0.544	**0.555**	0.637	0.624	**0.629**	0.559	0.524	0.541	0.657	0.610	0.631
VIII	12	0.489	0.467	*0.477*	0.561	0.522	0.540	0.536	0.506	0.520	0.661	0.611	0.634

The situation changes when we consider the classification results obtained by CRF. For 6-, 4-, 3- class classification, the most predictive feature set is the one that shows sentiment labels of the preceding and following posts, i.e. Set I. Enhancement of the four labels with indicators of the post position in the discussion outputs slightly lower results (Set II). However, these results are still higher than those obtained on other sets. 5-class classification is the only task where CRF benefited from a full spectrum of information available to it. Recall that in this task we removed the ambiguous posts, i.e., the ones labeled with two different labels, and kept original labels assigned by annotators.

Table 2. Classification results for CRF. For each task, the highest F-score is in **bold**, the lowest F-score – in *italics*.

Sets	Feat.	6-class			5-class			4-class			3-class		
		P	R	F	P	R	F	P	R	F	P	R	F
I	4	0.390	0.410	*0.355*	0.45	0.463	0.418	0.494	0.518	*0.487*	0.609	0.611	0.600
II	7	0.395	0.408	0.356	0.441	0.458	*0.414*	0.475	0.528	0.489	0.620	0.620	0.605
III	8	0.365	0.400	0.369	0.445	0.468	0.444	0.478	0.500	*0.484*	0.587	0.593	*0.587*
IV	11	0.371	0.407	0.373	0.449	0.471	0.448	0.493	0.513	0.495	0.595	0.601	0.594
V	14	0.431	0.460	**0.431**	0.524	0.528	**0.517**	0.507	0.538	**0.515**	0.632	0.630	**0.624**
VI	12	0.377	0.412	0.379	0.457	0.479	0.457	0.509	0.534	0.513	0.618	0.617	0.610
VII	12	0.404	0.437	0.408	0.506	0.514	0.501	0.498	0.527	0.504	0.620	0.618	0.612
VIII	12	0.387	0.416	0.387	0.470	0.479	0.463	0.503	0.530	0.508	0.631	0.629	0.623

If compared with the previous work on the same data (Bobicev et al. 2015a, 2015b). The best F-score = 0.613 for 6 class classification improved on the previously reported F-score = 0.491. Note, that we obtained this result based on the neighboring posts' sentiment labels, whereas the classification in (Bobicev et al. 2015a) was done on representing messages through emotional lexicons.

5 Discussion

In this work, we proposed a method that eschews the use of a lexical content in sentiment classification of online discussions. Using a data set gathered from a medical forum, we have shown that sentiments can be reliably classified when posts are represented through sentiment labels of the previous and following posts, enhanced by information about the author activity and the post position in the discussion. We solved 6-,5-,4-, and 3-class classification problems. On the most difficult 6-class classification task, the best F-score = 0.613 improves on the previously obtained F-score = 0.491.

SVM's performance improved when we added information about the post's author (i.e., prolificness, the initiator of the discussion, the discussion's newcomer). CFR performance, however, demonstrated that relationship between sentiments in the consecutive posts provide for a higher classification F-score than longer sentiment sequences. Overall, CRF outperformed SVM due to its ability to gauge information from a sequence of elements.

In this work, we applied a supervised learning approach which relies on manually annotated data. To reduce dependency on manual annotation, we plan a transition to semi-supervised learning. We have shown that sentiment transitions help to predict the sentiment of the current post. A vast volume of messages posted on social media makes the use of fully annotated data unrealistic. Thus, we plan to combine a lexicon-based sentiment classification with the features discussion in this work. A possible approach would be to use Markov chains to disambiguate ambiguous sentiment labels.

References

Balicco, L., Paganelli, C.: Access to health information: going from professional to public practices. In: Information Systems and Economic Intelligence: 4th International Conference – SIIE 2011 (2011)

Bobicev, V., Sokolova, M., Oakes, M.: What goes around comes around: learning sentiments in online medical forums. J. Cogn. Comput. (2015a). http://link.springer.com/article/10.1007/s12559-015-9327-y

Bobicev, V., Sokolova, M., Oakes, M.: Sentiment and factual transitions in online medical forums. In: Barbosa, D., Milios, E. (eds.) Canadian AI 2015. LNCS, vol. 9091, pp. 204–211. Springer, Heidelberg (2015b)

Cambria, E., Benson, T., Eckl, C., Hussain, A.: Sentic PROMs: application of sentic computing to the development of a novel unified framework for measuring health-care quality. Expert Syst. Appl. **39**, 10533–10543 (2012)

Chee, B., Berlin, R., Schatz, B.: Measuring population health using personal health messages. In: Proceedings of AMIA Symposium, pp. 92–96 (2009)

Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Aggarwal, C.C., Zhai, C.X. (eds.) Mining Text Data, pp. 415–463. Springer, US, New York (2012)

Malik, S.H., Coulson, N.S.: Coping with infertility online: an examination of self-help mechanisms in an online infertility support group. Patient Educ. Couns. **81**(2), 315–318 (2010). Elsevier

Mo, P.K.H., Coulson, N.S.: Developing a model for online support group use, empowering processes and psychosocial outcomes for individuals living with HIV/AIDS. Psychol Health. **27**(4), 445–459 (2012)

Pennebaker, J.W., Francis, M., Booth, R.: Linguistic Inquiry and Word Count Manual (2001)

Qiu, B., Zhao, K., Mitra, P., Wu, D., Caragea, C., Yen, J., Greer, G., Portier, K.: Get online support, feel better–sentiment analysis and dynamics in an online cancer survivor community. In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), pp. 274–281 (2011)

Rhodewalt, F.: Self-involvement, self-attribution, and the type a coronary-prone behavior pattern. J. Pers. Soc. Psychol. **47**(3), 662–670 (1984)

Shaw, B., Han, J., Hawkins, R., McTavish, F., Gustafson, D.: Communicating about self and others within an online support group for women with breast cancer and subsequent outcomes. J. Health Psychol. **13**(7), 930–939 (2008)

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

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)

Yoo, W., Namkoong, K., Choi, M., Shah, D.V., Tsang, S., Hong, Y., Gustafson, D.H.: 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)

Zafarani, R., Cole, W.D., Liu, H.: Sentiment propagation in social networks: a case study in LiveJournal. In: Chai, S.-K., Salerno, J.J., Mabry, P.L. (eds.) SBP 2010. LNCS, vol. 6007, pp. 413–420. Springer, Heidelberg (2010)