Recognition of Sentiment Sequences in Online Discussions

Victoria Bobicev
Technical University of Moldova
vika@rol.md

Marina Sokolova
University of Ottawa, Institute for Big Data Analytics, Canada
sokolova@uottawa.ca

Michael Oakes
Research Group in Computational Linguistics, University of Wolverhampton, UK
Michael.Oakes@wlv.ac.uk

Abstract

Currently 19%-28% of Internet users participate in online health discussions. In this work, we study sentiments expressed on online medical forums. As well as considering the predominant sentiments expressed in individual posts, we analyze sequences of sentiments in online discussions. Individual posts are classified into one of the five categories *encouragement, gratitude, confusion*, *facts,* and *endorsement*. 1438 messages from 130 threads were annotated manually by two annotators with a strong inter-annotator agreement (Fleiss kappa = 0.737 and 0.763 for posts in sequence and separate posts respectively). The annotated posts were used to analyse sentiments in consecutive posts. In automated sentiment classification, we applied HealthAffect, a domain-specific lexicon of affective words.

1 Introduction

Development of effective health care policies relies on the understanding of opinions expressed by the general public on major health issues. Successful vaccination during pandemics and the incorporation of healthy choices in everyday life style are examples of policies that require such understanding. As online media becomes the main medium for the posting and exchange of information, analysis of this online data can contribute to studies of the general public's opinions on health-related matters. Currently 19%-28% of Internet users participate in online health discussions (Balicco and Paganelli, 2011). Analysis of the information posted online contributes to effectiveness of decisions on public health (Paul and Drezde, 2011; Chee et al., 2009).

Our interest concentrates on sequences of sentiments in the forum discourse. It has been shown that sentiments expressed by a forum participant affect sentiments in messages written by other participants posted on the same discussion thread (Zafarani et al., 2010). Shared online emotions can improve personal well-being and empower patients in their battle against an illness (Malik and Coulson, 2010). We aimed to identify the most common sentiment pairs and triads and to observe their interactions. We applied our analysis to data gathered from the In Vitro Fertilization (IVF) medical forum.[1]

Below is an example of four consecutive messages from an embryo transfer discussion:

Alice: Jane - whats going on??

Jane: We have our appt. Wednesday!! EEE!!!

Beth: Good luck on your transfer! Grow embies grow!!!!

Jane: The transfer went well - my RE did it himself which was comforting. 2 embies (grade 1 but slow in development) so I am not holding my breath for a positive. This really was my worst cycle yet!!

In automated recognition of sentiments, we use HealthAffect, a domain-specific affective lexicon.

The paper is organized as follows: Section 2 presents related work in sentiment analysis, Section 3 introduces the data set and the annotation results, Section 4 presents HealthAffect, Section 5 describes the automated sentiment recognition experiments, and Section 6 discusses the results.

[1] http://ivf.ca/forums

Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP), pages 44–49,
Dublin, Ireland, August 24 2014.

2 Related Work

The availability of emotion-rich text has helped to promote studies of sentiments from a boutique science into the mainstream of Text Data Mining (TDM). The "sentiment analysis" query on Google Scholar returns about 16,800 hits in scholarly publications appearing since 2010. Sentiment analysis often connects its subjects with specific online media (e.g., sentiments on consumer goods are studied on Amazon.com). Health-related emotions are studied on Twitter (Chew and Eysenbach, 2010; Bobicev et al, 2012) and online public forums (Malik and Coulson, 2010; Goeuriot et al, 2012).

Reliable annotation is essential for a thorough analysis of text. Multiple annotations of topic-specific opinions in blogs were evaluated in Osman et al. (2010). Sokolova and Bobicev (2013) evaluated annotation agreement achieved on messages gathered from a medical forum. Bobicev et al. (2012) used multiple annotators to categorize tweets into positive, negative and neutral tweets. Merits of reader-centric and author-centric annotation models were discussed in (Balahur, Steinberger, 2009). In this work, we apply the reader-centric annotation model. We use Fleiss Kappa (Nichols et al, 2010) to evaluate inter-annotator agreement.

An accurate sentiment classification relies on electronic sources of semantic information. In (Sokolova and Bobicev, 2013; Goeuriot et al, 2011), the authors showed that the sentiment categories of SentiWordNet[2], WordNetAffect[3] and the Subjectivity lexicon[4] are not fully representative of health-related emotions. In the current work, we use HealthAffect, a domain-specific lexicon, to automatically classify sentiments. The lexicon has been introduced in (Sokolova and Bobicev, 2013).

Although there is a correlation between emotions expressed in consecutive posts (Chmiel et al, 2011; Tan et al, 2011; Hassan et al, 2012), so far health-related sentiment classification has focused on individual messages. Our current work goes beyond individual messages and studies sequences of sentiments in consecutive posts.

3 The IVF Data and Annotation Results

We worked with online messages posted on a medical forum. The forum communication model promotes messages which disclose the emotional state of the authors. We gathered data from the In Vitro Fertilization (IVF) website dedicated to reproductive technologies, a hotly debated issue in the modern society. Among the IVF six sub-forums, we selected the IVF Ages 35+ sub-forum[5] as it contained a manageable number of topics and messages, i.e., 510 topics and 16388 messages, where the messages had 128 words on average[6]. All topics were initiated by the forum participants. Among those, 340 topics contained < 10 posts. These short topics often contained one initial request and a couple of replies and were deemed too short to form a good discussion. We also excluded topics containing > 20 posts. This exclusion left 80 topics with an average of 17 messages per topic for a manual analysis by two annotators. First, we used 292 random posts to verify whether the messages were self-evident for sentiment annotation or required an additional context. The annotators reported that posts were long enough to convey emotions and in most cases there was no need for a wider context. We applied an annotation scheme which was successfully applied in (Sokolova and Bobicev, 2013).

We started with 35 sentiment types found by annotators and generalized them into three groups:

- ***confusion***, which included worry, concern, doubt, impatience, uncertainty, sadness, anger, embarrassment, hopelessness, dissatisfaction, and dislike;
- ***encouragement***, which included cheering, support, hope, happiness, enthusiasm, excitement, optimism;
- ***gratitude***, which included thankfulness.

A special group of sentiments was presented by expressions of compassion, sorrow, and pity. According to the WordNetAffect classification, these sentiments should be considered negative. However,

[2] http://sentiwordnet.isti.cnr.it/
[3] http://wndomains.fbk.eu/wnaffect.html
[4] http://mpqa.cs.pitt.edu/#subj_lexicon
[5] http://ivf.ca/forums/forum/166-ivf-ages-35/
[6] We harvested the data in July 2012.

in the context of health discussions, these emotional expressions appeared in conjunction with moral support and encouragement. Hence, we treated them as a part of *encouragement.* Posts presenting only factual information were marked as *facts*. Some posts contained factual information and strong emotional expressions; those expressions almost always conveyed encouragement ("*hope, this helps*", "*I wish you all the best*", "*good luck*"). Such posts were labeled *endorsement.* Note that the final categories did not manifest negative sentiments. In lieu of negative sentiments, we considered *confusion* as a non-positive label. *Encouragement* and *gratitude* were considered positive labels, *facts* and *endorsement* - neutral. It should be mentioned that the posts were usually long enough to express several sentiments. However, annotators were requested to mark messages with one sentiment category.

The posts that both annotators labelled with the same label were assigned to this category; 1256 posts were assigned with a class label. The posts labelled with two different sentiment labels were marked as *ambiguous*; 182 posts were marked as *ambiguous.*

Despite the challenging data, we obtained Fleiss Kappa = 0.737 which indicated a strong agreement between annotators (Osman et al, 2010). This value was obtained on 80 annotated topics. Agreement for the randomly extracted posts was calculated separately in order to verify whether annotation of separate posts was no more difficult than annotation of the post sequences. Contrary to our expectations, the obtained Fleiss Kappa = 0.763 was slightly higher than on the posts in discussions. The final distribution of posts among sentiment classes is presented in Table 2.

Classification category	**Num of posts**	**Per-cent**
Facts	494	34.4%
Encouragement	333	23.2%
Endorsement	166	11.5%
Confusion	146	10.2%
Gratitude	131	9.1%
Ambiguous	168	11.7%
Total	1438	100%

Table 2: Class distribution of the IVF posts.

We computed the distribution of sentiment pairs and triads in consecutive posts. We found that the most frequent sequences consisted mostly of *facts* and/or *encouragement*: 39.5% in total. *Confusion* was far less frequent and was followed by *facts* and *encouragement* in 80% of cases. That sentiment transition shows a high level of support among the forum participants. Approximately 10% of sentiment pairs are *factual* and/or *encouragement* followed by *gratitude*. Other less frequent sequences appear when a new participant added her post in the flow. Tables 3 and 4 list the results.

Sentiment pairs	**Occurrence**	**Percent**
facts, facts	170	19.5%
encouragement, encouragement	119	13.7%
facts, encouragement	55	6.3%
endorsement, facts	53	6.1%
encouragement, facts	44	5.1%

Table 3: The most frequent sequences of two sentiments and their occurrence in the data.

Sentiment triads	**Occurrence**	**Percent**
factual, factual, factual	94	12.8%
encouragement, encouragement, encouragement	63	8.6%
encouragement, gratitude, encouragement	18	2.4%
factual, endorsement, factual	18	2.4%
confusion, factual, factual	17	2.3%

Table 4: The most frequent triads of sentiments and their occurrences in the data.

4 HealthAffect

General affective lexicons were shown to be ineffective in sentiment classification of health related messages. To build a domain-specific lexicon, named HealthAffect, we adapted the Pointwise Mutual Information (PMI) approach (Turney, 2002). The initial candidates consisted of unigrams, bigrams and trigrams of words with frequency ≥ 5 appearing in unambiguously annotated posts (i.e., we omitted posts marked as uncertain). For each class and each candidate, we calculated PMI(*candidate, class*) as

$$\text{PMI}(\textit{candidate}, \textit{class}) = \log_2(\text{p}(\textit{candidate} \text{ in } \textit{class})/(\text{p}(\textit{candidate}) \text{p}(\textit{class}))).$$

Next, we calculated Semantic Orientation (SO) for each candidate and for each class as

$$\text{SO}(\textit{candidate, class}) = \text{PMI}(\textit{candidate}, \textit{class}) - \Sigma \text{PMI}(\textit{candidate}, \textit{other\_classes})$$

where *other_classes* include all the classes except the class that Semantic Orientation is calculated for. After all the possible SO were computed, each HealthAffect candidate was assigned with the class that corresponded to its maximum SO.

Domain-specific lexicons can be prone to data over-fitting (since, for example, they might contain personal and brand names). To avoid the over-fitting pitfall, we manually reviewed and filtered out non-relevant elements, such as personal and brand names, geolocations, dates, stop-words and their combinations (since_then, that_was_the, to_do_it, so_you). Table 5 presents the lexicon profile. Note that we do not report the *endorsement* profile as it combines *facts* and *encouragement.*

Class	unigrams	bigrams	trigrams	total	Examples
Facts	204	254	78	536	`round_of_ivf`, `hearbeat`, `a_protocol`
Encouragement	127	107	68	302	`congratulations`, `is_hard`, `only_have_one`
Confusion	63	143	34	240	`crying`, `away_from`, `any_of_you`
Gratitude	37	51	34	122	`appreciate`, `a_huge`, `thanks_for_your`

Table 5: Statistics of the HealthAffect lexicon.

5 Sentiment Recognition

Our task was to assess HealthAffect's ability to recognise sentiments of health-related messages. We used the sentiment categories described in Section 3. In the experiments, we represented the messages by the HealthAffect terms. There were 1200 distinct terms, and each term was assigned to one sentiment.

Our algorithm was straightforward: it calculated the number of HealthAffect terms from each category in the post and classified the post in the category for which the maximal number of terms was found. Table 5 demonstrates that the number of terms was quite different for each category. Hence, the algorithm tended to attribute posts to the classes with a larger numbers of terms. To overcome the bias, we normalised the number of the terms in the post by the total number of terms for each category. The algorithm's performance was evaluated through two multiclass classification results:

- 4-class classification where all 1269 unambiguous posts are classified into (*encouragement, gratitude, confusion*, and neutral, i.e., *facts* and *endorsement*), and
- 3-class classification (positive: *encouragement, gratitude;* negative: *confusion*, neutral: *facts* and *endorsement*).

We computed micro- and macro-average *Precision (Pr), Recall (R) and F-score (F)* (Table 6).

Metrics	4-class classification	3-class classification
microaverage F-score	0.633	0.672
macroaverage Precision	0.593	0.625
macroaverage Recall	0.686	0.679
macroaverage F-score	0.636	0.651

Table 6: Results of 4-class and 3-class classification.

For additional assessment of HealthAffect, we ran simple Machine Learning experiments using Naïve Bayes and representing the texts through the lexicon terms. The obtained results of F-score=0.44, Precision=0.49, Recall=0.47 supported our decision to use HealthAffect in the straight-forward manner as presented above. For each sentiment class, our results were as follows:

- The most accurate classification occurred for *gratitude*. It was correctly classified in 83.6% of its occurrences. It was most commonly misclassified as *encouragement* (9.7%). Posts classified as *gratitude* are mostly the shortest ones containing only some words of gratitude and appreciation of others' help. As they usually do not contain any more information than this, there were fewer chances for them to be misclassified.
- The second most accurate result was achieved for *encouragement.* It was correctly classified in 76.7% of cases. It was misclassified as neutral (9.8%) because the latter posts contained some encouraging with the purpose of cheering up the interlocutor.
- The least often correctly classified class was neutral (50.8%). One possible explanation is the presence of the sentiment bearing words in the description of facts in a post which is in general objective and which was marked as factual by the annotators.

Recall from Section 3, that we consider *encouragement* and *gratitude* to be positive sentiments and *confusion* to be a negative one. The reported results show that positive sentiments were most misclassified within the same group or with neutral, e.g., *encouragement* was misclassified more as neutral or *gratitude* than as *confusion*, *gratitude* - more as *encouragement* or neutral than as *confusion.* On the other hand, *confusion* and negative sentiments were most often misclassified as neutral.

6 Discussion and Future Work

We have presented results of sentiment recognition in messages posted on a medical forum. Sentiment analysis of online medical discussions differs considerably from polarity studies of consumer-written product reviews, financial blogs and political discussions. While in many cases positive and negative sentiment categories are powerful enough, such a dichotomy is not sufficient for medical forums. We formulate our medical sentiment analysis as a multi-class classification problem in which posts were classified into *encouragement, gratitude, confusion, facts* and *endorsement*.

In spite of sentiment annotation being highly subjective, we obtained a strong inter-annotator agreement between two independent annotators (i.e., Fleiss Kappa = 0.73 for posts in discussions and Fleiss Kappa = 0.76 for separate posts). The Kappa values demonstrated an adequate selection of classes of sentiments and appropriate annotation guidelines. However, many posts contained more than one sentiment in most cases mixed with some factual information. The possible solutions in this case would be (a) to allow multiple annotations for each post; (b) to annotate every sentence of the posts.

A specific set of sentiments on the IVF forum did not support the use of general affective lexicons in automated sentiment recognition. Instead we applied the PMI approach to build a domain-specific lexicon HealthAffect and then manually reviewed and generalized it.

In our current work we went beyond analysis of individual messages: we analyzed their sequences in order to reveal patterns of sentiment interaction. Manual analysis of a sample of data showed that topics contained a coherent discourse. Some unexpected shifts in the discourse flow were introduced by a new participant joining the discussion. In future work, we may include the post's author information in the sentiment interaction analysis. The information is also important for analysis of influence, when one participant is answering directly to another one citing in many cases the post which she answered to.

We plan to use the results obtained in this study for analysis of discussions related to other highly debated health care policies. One future possibility is to construct a Markov model for the sentiment sequences. However, in any online discussion there are random shifts and alternations in discourse which complicate application of the Markov model.

In the future, we aim to annotate more text, enhance and refine HealthAffect, and use it to achieve reliable automated sentiment recognition across a spectrum of health-related issues.

References

Balicco, L., C. Paganelli. 2011. *Access to health information: going from professional to public practices*, Information Systems and Economic Intelligence: 4th International Conference - SIIE'2011.

Bobicev, V., M, Sokolova, Y. Jaffer, D. Schramm. 2012. *Learning Sentiments from Tweets with Personal Health Information.* Proceedings of Canadian AI 2012, p.p. 37–48, Springer.

Chee, B., R. Berlin, B. Schatz. 2009. *Measuring Population Health Using Personal Health Messages*. Proceedings of AMIA Symposium, 92 - 96.

Chew, C. and G. Eysenbach. 2010. *Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak*. PLoS One, 5(11).

Chmiel, A., J. Sienkiewicz, M. Thelwall, G. Paltoglou, K. Buckley, A. Kappas, J. Holyst. 2011. *Collective Emotions Online and Their Influence on Community Life*. PLoS one.

Goeuriot, L., J. Na, W. Kyaing, C. Khoo,Y. Chang, Y. Theng and J. Kim. 2012. *Sentiment lexicons for health-related opinion mining*. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, p.p. 219 – 225, ACM.

Hassan, A., A. Abu-Jbara, D. Radev. 2012. *Detecting subgroups in online discussions by modeling positive and negative relations among participants*. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 59-70).

Malik S. and N. Coulson. 2010. *Coping with infertility online: an examination of self-help mechanisms in an online infertility support group.* Patient Educ Couns, vol. 81, no. 2, pp. 315–318

Nichols, T., P. Wisner, G. Cripe, and L. Gulabchand. 2010. *Putting the Kappa Statistic to Use.* Qual Assur Journal, 13, p.p. 57-61.

Osman, D., J. Yearwood, P. Vamplew. 2010. *Automated opinion detection: Implications of the level of agreement between human raters.* Information Processing and Management, 46, 331-342.

Paul, M. and M. Dredze. 2011. *You Are What You Tweet: Analyzing Twitter for Public Health*. Proceedings of ICWSM.

Sokolova, M. and V. Bobicev. 2013. *What Sentiments Can Be Found in Medical Forums*? Recent Advances in Natural Language Processing, 633-639

Tan, C., L. Lee , J. Tang , L. Jiang , M. Zhou, P. Li, 2011. *User-level sentiment analysis incorporating social networks*, Proceedings of the 17th ACM SIGKDD international conference on KDDM.

Turney, P.D. 2002. *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*. Proceedings of ACL'02, Philadelphia, Pennsylvania, pp. 417-424.

Zafarani, R., W. Cole, and H. Liu. 2010. *Sentiment Propagation in Social Networks: A Case Study in LiveJournal*. Advances in Social Computing (SBP 2010), pp. 413–420, Springer.