# Thumbs Up *and* Down: Sentiment Analysis of Medical Online Forums

**Victoria Bobicev**
Technical University of Moldova
***victoria.bobicev@ia.utm.md***

**Marina Sokolova**
IBDA@Dalhousie University and
University of Ottawa
***sokolova@uottawa.ca***

## Abstract

In the current study, we apply multi-class and multi-label sentence classification to sentiment analysis of online medical forums. We aim to identify major health issues discussed in online social media and the types of sentiments those issues evoke. We use ontology of personal health information for Information Extraction and apply Machine Learning methods in automated recognition of the expressed sentiments.

## 1 Introduction

**Computational Health.** Online social media became an invaluable and ever growing source of Computational Health (Collier et al, 2017; Sarker et al, 2015). Personal health information, i.e. information about health that individuals share in clinical settings, had been found on Twitter, other social networks, in blogs and medical forums (Sokolova and Schramm, 2011). A diverse language and a subjective style of social media messages stipulate two principal components of Computational Health: i) automated recognition of medical concepts, ii) automated identification of sentiments. The former is essential for extraction of health information (Limsopatham and Collier, 2016); the latter enables to recognize personal attitude in discussion of one's health (Sokolova and Bobicev, 2013).

We apply multi-class and multi-labeled sentence classification in sentiment analysis of online medical forums. We aim to identify major health issues discussed in online social media and the types of sentiments those issues evoke. In order to do this, we adapt ontology of personal health information used in social media studies (Sokolova and Schramm, 2011). By using Machine Learning methods in multi-class classification, we significantly improve over the majority class baseline (paired t-test for all the eight labels: P = 0.0062) and over the look-up results (paired t-test over all the labels, P=0.0208).

## 2 Related Work.

Sentiment analysis of user-written content has been performed intensely for studies of goods and services reviews, tweets and blogs (Serrano-Guerrero et al., 2015). Khan et al (2016) have shown that a rule-based sentiment classification can be a viable method of sentence-based sentiment analysis. We differentiate between lexicon-based and aspect-based approaches in sentiment analysis studies. The lexicon-based analysis relies on retrieval of lexical expressions of sentiments (Taboada et al, 2011), whereas the aspect-based analysis focuses on sentiments and opinions related to specific features of the product or service (Liu, 2012).

Sentiment analysis of health information is an expanding research domain (Denecke and Deng, 2015). It had been shown that sentiments can be conclusively connected with health issues (Chen and Sokolova, 2018). Health-related texts often express complex sentiments, hence benefit from a multi-label approach in sentiment classification (Bobicev and Sokolova, 2017).

Navindgi et al. (2016) used syntactic features to compare document-level and sentence-level multi-class sentiment classification of online medical forums. They opine that adding social components can benefit the classification results.

Many health-related studies use Twitter data, a popular sphere of public communications (Grover et al, 2018). Tweets had been used in Information Extraction of personal health information (Sokolova et al, 2013), as well as in health studies of specific population groups (Bravo and Goetz, 2017) and in analysis of particular health-related issues (Abbasi et al, 2018). Sokolova et al (2013) had shown that personal pronouns and family relations significantly im-

*Proceedings of the 3rd Social Media Mining for Health Applications (SMM4H) Workshop & Shared Task*, pages 22–26
Brussels, Belgium, October 31, 2018. ©2018 Association for Computational Linguistics

proved accuracy of health information extraction from Twitter.

## 3 The Data Set Construction

We work with texts harvested from *in vitro fertilization* forums, namely, ivf.ca, with posts annotated by multiple sentiments.[1] The posts are comparatively informative, containing approx. 100-150 words each. Many posts express more than one sentiment and discuss more than one topic. The posts had been studied in a multi-label sentiment classification (Bobicev and Sokolova, 2017). In the said study, multi-label classification has been applied to a complete post, thus leaving aside a nuanced analysis of the expressed sentiments. In the current work, we use *sentences* as the units of the study to gain more detailed information about expressed sentiments.

**Sentiment categories.** We use two categories ***encouragement***, ***confusion***, and ***facts*** introduced in previous studies (Sokolova and Bobicev, 2013).

***Encouragement*** indicates sentiments expressed towards the interlocutors of the post author. The expressed sentiments aim to support and inspire other people reading the posts. At the same time, this support is expressed by describing details of treatment such as: clinics, doctors, procedures or medicines that could lead to the desired outcome.

***Confusion*** generalizes various nuances of negative sentiments: uncertainty, hopeless, frustration, complaint, etc. While analyzing the posts marked by ***confusion***, we aim to extract the cause of these negative sentiments; here we differentiate between health issues *per se* and issues of treatment.

***Facts*** is used to label the objective discussions. In posts labeled by ***facts*** we seek to extract information related to health (e.g., treatment, procedures, prescribed medications).

**Health issue categories.** The health-related ontology introduced in (Sokolova and Schramm, 2011) was the main resource of Information Extraction procedures. The ontology has been created to study user-written online messages on health-related topics. It contained four main health issue categories: (1) 'Person' with subclasses 'Anatomical parts' and 'Physiological functions'; (2) 'Health-Related Problems' with subclasses 'Symptoms' and 'Diseases'; (3) 'Health Care System' with subclasses 'Health Care Providers', 'Health Care Setting' 'Health Care Procedures'; (4) Health-Related Environmental Factors.

We expanded the ontology with two new categories: Intakes and External Factors. Our initial version of the ontology listed the following categories: (1) Body: parts, organs, elements, functions; (2) Health conditions: symptoms, diseases; (3) Health care: providers, settings; (4) Health care actions: diagnostics, procedures; (5) Intakes: medicines, supplements, food; (6) External factors: family, work, finances.

However, a simple lookup resulted in high precision and low recall (Precision=0.97, Recall=0.23). The low Recall was due to various spelling of health related terms, especially multi-syllable medical terms (e.g. echocardiography') and specific abbreviations (e.g., ultrasound was written as US or U/S). Unlike in studies of Twitter data (Sokolova et al, 2013), adding personal pronouns and family relations did not improve accuracy of the health information retrieval. In our data, the authors used personal pronouns indiscriminately in description of health issues and other topics. When creating unigram models for posts with health information and without it, we observed that 'I' is the most frequent word in both. The next most frequent pronoun in health related text is 'my' and in non-health related texts - 'you'; family relationship mentioning is actually more frequent in non-health related texts.

The final set of the ontology term categories (i.e., health issues) was as following: (1) Body parts, organs; (2) Health conditions: symptoms, diseases; (3) Health care providers; (4) Actions: procedures; (5) Intakes.

**Sentence annotation.** We selected 160 posts for sentence annotation and further evaluation by machine learning methods. The selected posts i) had to have 2 or more sentiment labels, ii) had to be an average length (300 - 600 characters, or 50-100 words). Those posts had been split into sentences. Each sentence was manually annotated using two sets of labels: sentiments and health issues mentioned in this sentence.

It is important to note that sentences could have more than one label from the same category, e.g., ***encouragement*** and ***facts, providers*** and ***organs.*** Some sentences had multiple labels and some sen-

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

tences had zero labels. For example, "*So it's a matter of getting the balance right*." did not have assigned labels, whereas "*I just want to make it clear to anyone with DOR or LOR that there still is hope!*" has been assigned with ***encouragement*** and ***symptoms***.

The annotation resulted in 1087 sentences annotated with the total of 985 labels (Table 1).

| Labels | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Sentences | 490 | 297 | 226 | 61 | 12 | 1 |

Table 1: The statistics on the label distribution.

Further, we worked with the label distribution presented in Table 2.

| **Sentiments** | **Health Issues** |
|---|---|
| facts : 213<br>encouragement : 110<br>confusion : 70 | procedures : 234<br>symptoms : 127<br>providers : 86<br>organs : 84<br>intakes : 61 |

Table 2: Label distribution in the data set.

## 4 Empirical Studies

**Feature selection.** We tokenized each sentence and built the unigram model of the data. All the tokens have been used as features in the initial feature set.

To obtain the best set of features for each label we used Information Gain ($IG$) to calculate coefficients of the token importance for the current label: $IG\ (token, label) = H(label) - H(label/token)$.

For example, the highest coefficients for the topic 'organs' were: *eggs* - 0.079, *tubes*- 0.021, *egg*- 0.018, *ovary* 0.017, *ovaries* 0.014, and the lowest for the selected features were: *abdominal* - 0.0034, *sorta* - 0034, *like* - 0030. We calculated the coefficients for every word and selected words with the coefficients $> 0$.

**Multi-class classification.** We calculated the baseline F-measure (B) where all instances are attributed to the majority class. Thus, F-measure is quite high due to the data imbalance.

To assess difficulty of the multi-class classification, we used a straight-forward look-up to identify each label. The threshold for the label has been selected by balancing Precision (Pr) and Recall (R) of this label recognition. Table 3 shows Pr, R and F-measure (F) calculated for each label.

For the five health issue labels, the look-up non-significantly improved F-measure over the baseline (paired t-test for the five health issue labels: P=0.1308); classification improvement did not happen for the three sentiment labels, albeit F-measure decrease was not significant (paired t-test for the three sentiment labels: P=0.1060).

We used Machine Learning experiments to improve sentence-based sentiment classification. To find algorithms that can improve on the baseline, we applied applied Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, and Decision Tree classifiers from WEKA[2] toolkit.

We applied 10-fold cross validation on the set of the annotated sentences. Table 4 reports the best results for each label. SVM and KNN substantially outperformed other algorithms. The results show that the best results significantly improved over the baseline results: paired t-test for all the labels: P = 0.0062. Improvement over the look-up results is also statistically significant: paired t-test over all the labels, P=0.0208.

However, these experiments treated every label individually and did not reveal relationship among them. To seek relationship among the label categories and the individual labels, we involved multi-label classification.

| **Label** | **B** | **Pr** | **R** | **F** |
|---|---|---|---|---|
| facts | 0.717 | 0.735 | 0.653 | 0.692 |
| encourage | 0.851 | 0.971 | 0.600 | 0.742 |
| confusion | 0.904 | 0.891 | 0.700 | 0.784 |
| | | | | |
| procedures | 0.690 | 0.759 | 0.739 | 0.749 |
| symptoms | 0.828 | 0.958 | 0.888 | 0.922 |
| organs | 0.887 | 0.985 | 0.807 | 0.887 |
| providers | 0.883 | 0.925 | 0.860 | 0.892 |
| intakes | 0.917 | 0.919 | 0.934 | 0.927 |

Table 3: Multi-class labels' lookup results.

**Multi-label classification.** In multi-label classification (Sorower, 2010), we focused on joint detection of the sentiment and health issues labels assigned to a sentence. We had 667 sentences with at least one label. To convert from a multi-label to a uni-label problem, we used Binary Relevance (BR) problem transformation method. It creates k datasets, each for every single label, and trains the classifier on each of these data sets.

[2] https://www.cs.waikato.ac.nz/ml/weka/

Using the lookup method we obtained 292 sentences with Exact Match (EM) = 0.438.

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

where *n* denotes the number of sentences in the data set, $Y_i, Z_i$ are sets of predicted and true labels for sentence *i* respectively.

| Label | Alg. | B | Pr | R | F |
|---|---|---|---|---|---|
| facts | SVM | 0.717 | 0.845 | 0.854 | 0.831 |
| encourage. | KNN | 0.851 | 0.897 | 0.905 | 0.869 |
| confusion | KNN | 0.904 | 0.947 | 0.952 | 0.942 |
| | | | | | |
| procedures | SVM | 0.690 | 0.848 | 0.854 | 0.835 |
| symptoms | SVM | 0.828 | 0.906 | 0.908 | 0.884 |
| organs | KNN | 0.887 | 0.954 | 0.951 | 0.940 |
| providers | SVM | 0.883 | 0.922 | 0.930 | 0.907 |
| intakes | SVM | 0.917 | 0.967 | 0.966 | 0.959 |

Table 4: The best multi-class results of ML algorithms.

EM is the ultimate assessment of accuracy, as it counts only sentences with every label found and identified correctly. This means that the system detected correctly all the labels for more than 40% of sentences (443 labels in total).

The look-up classified 219 sentences with a partial match, where 294 labels were matched correctly, 145 labels were false negative and 115 labels were false positive. 'Match' indicates manually annotated a label found by the lookup; 'false positive' shows that a label was found by the lookup but not by the manual annotation; 'false negative' indicates an annotated label missed by the lookup.

Among 156 completely mismatched sentences, 103 labels were classified as false negative and 96 labels were classified as false positive.

We have applied multi-label Machine Learning algorithms from MEKA toolkit[3]. As in multi-class-classification, we used 10-fold cross-validation. In this task, SVM and Naïve Bayes outperformed the other algorithms. SVM obtained EM = 0.513, F (by label) = 0.438. Naïve Bayes obtained EM = 0.421, F (by label) = 0.406.

The best EM, obtained by SVM, is higher than EM = 0.450 reported for studies of the complete posts (Bobicev and Sokolova, 2017). In addition to classifying a bigger unit, the cited work analyzed only four sentiment labels, whereas we obtained a higher EM in a more complex classification of three sentiment labels and five health issue labels. However, our data set is considerably smaller that the data used in the previous study: 597 sentences vs 1321 posts.

**Error analysis.** We categorized reasons for errors as follows: (1) linguistic challenges: irony, misspellings, ambiguous sentence structure that requires application of specialized linguistic methods; (2) limitations of the knowledge source, i.e., deficiency of terms in the applied ontology; (3) system limitations, e.g., inability of our system to capture long distance relations of terms and sentiments.

## 5 Conclusions and Future Work

We present a preliminary sentence-level sentiment analysis of posts gathered from a medical forum. The posts were informative enough to express several sentiments and cover several health issues. As a result, we analyzed a multi-labeled data set, where some labels revealed sentiments and other labels indicated underlying health issues.

We adapted ontology that was previously used in personal health information extraction from a heterogeneous social media data to identify health issues in the data set. Respectively, we added Intake terms and populated the ontology with domain specific terms of In Vitro Fertilization and their slang spellings used by the online forum participants. By using Machine Learning methods in multi-class classification, we have obtained significant improvement over the majority class baseline (paired t-test for all the eight labels: P = 0.0062) and significant improvement over the look-up results (paired t-test over all the labels, P=0.0208). The obtained results on multi-label classification are less conclusive, in part, because a small data set.

Hence, we want to expand the data set through annotation of more posts on the sentence level. This will allow us to use syntactic structures of sentences in order to better capture their semantics.

At the same time, more work should be done for development of an automated and robust system that can reliably classify sentiments and related to them health issues on social media. To improve on Information Extraction, we plan to augment the current ontology.

Finally, we want to test the same approach on posts collected from other medical forums.

[3] http://waikato.github.io/meka/

## Acknowledgements

We thank the SMM4H anonymous reviewers for thorough and helpful comments.

## References

Abbasi, Rabeeh Ayaz, Onaiza Maqbool, Mubashar Mushtaq, Naif R. Aljohani, Ali Daud, Jalal S. Alowibdi, and Basit Shahzad. 2018. Saving lives using social media: Analysis of the role of twitter for personal blood donation requests and dissemination. *Telematics and Informatics* 35(4), pp. 892-912.

Bobicev, Victoria, and Marina Sokolova. 2017. Confused and Thankful: multi-label sentiment classification of health forums. *Proceeding of Canadian Conference on Artificial Intelligence 2017,* pp 284-289.

Bravo, Caroline, and Laurie Hoffman-Goetz. 2017. Social media and men's health: a content analysis of Twitter conversations during the 2013 Movember campaigns in the United States, Canada, and the United Kingdom. *American journal of men's health* 11 (6), pp. 1627-1641.

Chen, Qufei, and Marina Sokolova. 2018. Word2Vec and Doc2Vec in Unsupervised Sentiment Analysis of Clinical Discharge Summaries. *arXiv preprint* arXiv:1805.00352.

Collier, Nigel, Nut Limsopatham, Aron Culotta, Mike Conway, Ingemar J. Cox and Vasileios Lampos. 2017. WSDM 2017 *Workshop on Mining Online Health Reports.*

Denecke, Kerstin, and Yihan Deng. 2015. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine*, 64(1):17–27.

Grover, Purva, Arpan Kumar Kar, and Gareth Davies. 2018. "Technology enabled Health"–Insights from twitter analytics with a socio-technical perspective." *International Journal of Information Management* ,43, pp. 85-97.

Khan, Jawad, Byeong Soo Jeong, Young-Koo Lee, and Aftab Alam. 2016. "Sentiment analysis at sentence level for heterogeneous datasets." In *Proceedings of the Sixth International Conference on Emerging Databases: Technologies, Applications, and Theory*, pp. 159-163.

Limsopatham, Nut, and Nigel Collier. 2016. Normalizing medical concepts in social media texts by learning semantic representation. *In Meeting of the Association for Computational Linguistics pp.* 1014-1023.

Liu, Bing. 2012. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Navindgi, Amit, Caroline Brun, Cecile Boulard, Scott Nowson. 2016. Steps Toward Automatic Understanding of the Function of Affective Language in Support Groups. *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pp. 26-33.

Sarker, Abeed, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: a review. *Journal of biomedical informatics* 54: 202-212.

Serrano-Guerrero, Jesus, Jose A. Olivas, Francisco P. Romero, and Enrique Herrera-Viedma. 2015. Sentiment analysis: A review and comparative analysis of web services. *Information Sciences* 311: pp. 18-38.

Sokolova, Marina, and David Schramm. 2011. Building a Patient-based Ontology for User-written Web Messages. *RANLP 2011*.

Sokolova, Marina, and Victoria Bobicev. 2013. What Sentiments Can Be Found in Medical Forums? *RANLP 2013*, pp. 633–639.

Sokolova, Marina, Stan Matwin, Yasser Jafer, David Schramm. 2013. How Joe and Jane Tweet about Their Health: Mining for Personal Health Information on Twitter. *RANLP 2013*.

Sorower, Mohammad S. 2010. A literature survey on algorithms for multi-label learning. *Technical report, Oregon State University, Corvallis*.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37 (2), pp. 267-307