Sentiment Analysis in Health Related Forums

Victoria BOBICEV
Technical University of Moldova
vika@rol.md

Marina SOKOLOVA
CHEO Research Institute
University of Ottawa
sokolova@uottawa.ca

***Abstract* — In this work, we have presented the sentiment analysis of messages posted on medical forums. We stated the sentiment analysis as a multi-class classification problem in which posts were classified into *encouragement, gratitude, confusion, facts, facts + encouragement* and *uncertain* categories. We applied the reader-centered manual annotation and achieved a strong agreement between the annotators: *Fleiss Kappa* = 0.73.**

We presented an ad-hoc method of the lexicon creation which is comparatively easy to implement. We have shown that the lexicon, which we call HealthAffect, provided the best accuracy in machine learning experiments. . We used two algorithms, NB and KNN, to solve a multi-class sentiment classification problem. The probability-based NB demonstrated a better performance than KNN. .

***Index Terms* — Computational linguistics, Natural Language Processing, sentiment analysis, social media analysis, Machine Learning.**

I. INTRODUCTION

The 'social web' that has evolved through the fast development of ICT technologies and improved access to the internet has in turn created an unprecedented digital resource of facts, opinions and views that has potential to influence considerably the development of policies and practice in health, patient engagement, economic efficiencies and co-creation in patient care. ICT will also be indispensable in contributing to key societal challenges and processes such as citizen behaviour and public governance[1].

Sentiment analysis offers a solution for responding to the challenge of how online data can be exploited for health sector and societal gain. Methods such as Text Data Mining (TDM) and Natural Language Processing (NLP) have already demonstrated their value in intensively analysing sentiments and opinions in consumer-written product reviews [1], financial blogs and political discussions [2]. Text analysis of user-written online messages has been stipulated by both the demand for such studies from the one hand and an easy access to the online data from the other [3,4]. Extraction and analysis of sentiments, opinions, attitudes, emotions, perceptions and intentions is one of the most requested types of text analysis, according to Seth Grimes Text Analytics Report 2014[2].

Although researchers studied sentiments and opinions in user-written Web texts of various types, there are few studies of the relationship between a subjective language and personal health information posted on social networking sites [5]. Amelia Burke-Garcia in her presentation at Sentiment Analysis Symposium in 2013, underlined that from the one hand "Family and social networks' role in personal health decisions is paramount" and from the other "Matching social conversation with other data can allow you to make data-driven decisions[3]."

She mentioned that 42% of companies have social listening as a top priority in 2013. Not only commercial companies are waking up to the use of novel technologies to listen to the 'wisdom of the patient'. Health-care of the future will be based on community, collaboration, self-caring, co-creation and co-production using technologies delivered via the Web [6]."

II. STATE OF THE ART

The large field of research called Sentiment Analysis or Opinion Mining includes in fact several related tasks. Researches in this field included opinion and attitude classification [7], mood summarization [8], subjectivity analysis [9], emotion and affect detection [10]. Their solutions highly depend on analysed texts, final scope and available resources.

Most of the investigated task used only two classes: positive and negative [11]. Only some researches operated with several so called "basic" emotions or with larger sets of sentiments. Several works operated with a large list of affects producing graphical representation of overall affective text characteristics [12].

One of the basic resources in this domain of study is corpus with affective annotation [13]. The simplest way to obtain sentiment information about texts was to find text labelled by their authors as, for example, customers reviews marked with zero to five stars, or simply "thumbs up – thumbs down".

However this annotation presents only two opposite classes of sentiments: positive and negative. However the spectrum of human sentiments is much more diverse.

The second lexical resource necessary for sentiment analysis is sentiment lexicon. Lists with positive and

[1] http://ec.europa.eu/programmes/horizon2020
[2] http://altaplana.com/grimes.html
[3] http://vimeo.com/67882832

negative emotion bearing words were collected by many researchers[4,5] and research groups[6,7]. Some lists contained opinionated words and opinion phrases and idioms.

Several sentiments were considered in DepecheMood lexicon [14]; WordNet-Affect lexicon represented sentiments from Princeton WordNet. It is obvious that affect-bearing words are not enough to detect sentiments and opinions. A word can be neutral, bear positive or negative connotation given its context. Emotion can be presented in text without any emotional word. In this case manual annotation of such texts is necessary. The other possible solution is to connect a sentiment lexicon with a lexico-semantic resource which represents pragmatic links between sentiments, senses and lexical constructions. For example, phrase "not enough money" means that a person cannot buy something (s)he wants and is associated with negative sentiment. However, attempt to collect all information in one semantic source bring the other type of problems: the bigger semantic network is, the more difficult is to process it and obtain knowledge necessary in each particular case.

Dependent on these two main lexical resources methods used in this domain are classified on (1) lexicon-based and (2) based on machine-learning techniques [15]. Obviously, two these methods were combined in many studies [16]. The full overview of tasks, methods and approaches is given in the Bing Liu book [18].

The biggest advantage of machine learning methods is their independence of any lexical resource except of texts that are analysed. The biggest disadvantage has the same source: they rely only on texts that are analysed. Thus, for each new task and topic they need domain adaptation [1].

III. DATA

Our current research focuses on sentiment identification in messages posted on IVF forums. Such forums belong to an infertility outreach resource community created by prospective, existing and past IVF (In Vitro Fertilization) patients. The IVF.ca website includes forums: Cycle Friends, Expert Panel, Trying to Conceive, Socialize, In Our Hearts, Pregnancy, Parenting, and Administration.[8] Every forum hosts a few sub-forums, e.g. the Cycle Friends forum has six sub-forums: Introductions, IVF/FET/IUI Cycle Buddies, IVF Ages 35+, Waiting Lounge, Donor & Surrogacy Buddies, and Adoption Buddies. On every sub-forum, topics are initiated by the forum participants. Depending on the interest among participants, a different number of messages is associated with each topic.

We wanted the forum to represent many discussions, and so forums were selected to ensure a high number of topics and large number of posts. The IVF Ages 35+ sub-forum[9] satisfied both requirements.

In July 2012, it had 510 topics and 16388 messages. At this point, we discharged the largest four topics containing 7498, 2823, 1131 and 222 posts respectively; we discharged the shortest topics as well.

Among the remaining 506 topics, we looked for those where the forum participants discussed only one theme. A preliminary analysis showed that discussions with ≤ 20 posts satisfied this condition. Also, we wanted discussions be long enough to form a meaningful discourse. This condition was satisfied when discussion had ≥ 10 messages. As a result, for further analysis, we selected 74 topics with 10 - 20 posts, with an average 12.5 messages per topic. Most of the topics had a similar structure:

a) a participant started the theme with a post;
b) the initial post usually contained some information about the participant's problem, expressed worry, concern, uncertainty and a request for help to the other forum participants;
c) the following posts:
   i) provided the requested information by describing their similar stories, knowledge about treatment procedures, drugs, doctors and clinics, or
   ii) supplied moral support through compassion, encouragement, wishing all the best, good luck, etc.
d) the participant who started the topic often thanked other contributors and expressed appreciation for their help and support.

IV. ANNOTATION

We asked annotators to label the post with the dominant sentiment. Posts that combined factual information and sentiments usually expressed encouragement for specific participants, hence we suggested the label *"facts +encouragement"* for that category.

We wanted to know what types of sentiments were dominant in these forums and how these sentiments influence each other.

We intended to build a set of sentiments that

1. contains sentiment categories specific for posts from medical forums, and
2. makes feasible the use of machine learning methods for automate sentient detection.

To identify such a set, we asked annotators to read several topic discussions and describe sentiments expressed by the forum participants and the sentiment propagation within these discussions.

We asked annotators not to mark descriptions of symptoms and diseases as subjective; in many cases they appear in the post as objective information for other forum participants that have encountered similar issues. In such cases only the author's sentiments toward other participant should be taken into consideration.

The data annotation was carried on by Applied Informatics students as their practical work for the course "Semantic Interpretation of Text". Each annotator independently annotated a set of topics.

Based on the annotations, we built three groups of sentiments:

1. **confusion**, which included worry, concern, doubt, impatience, uncertainty, sadness, angriness, embarrassment, hopelessness, dissatisfaction, and dislike;

[4] http://www.cse.unt.edu/~rada/downloads.html#msa
[5] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
[6] http://www.affective-sciences.org/researchmaterial
[7] http://mpqa.cs.pitt.edu/
[8] www.ivf.ca/forums
[9] http://ivf.ca/forums/forum/166-ivf-ages-35/

2. **encouragement**, which included cheering, support, hope, happiness, enthusiasm, excitement, optimism;
3. **gratitude**, which included thankfulness.

A special case was presented by expressions of *compassion*, *sorrow*, and *pity* which did not appear individually but appeared in conjunction with encouragement; we treated them as a part of encouragement.

Also, we identified two types of posts with factual information: *facts* and *facts + encouragement*. Posts were marked as *facts* if they delivered factual information only. Posts were marked as *facts + encouragement* when they contained factual information supplemented by short emotional expressions; those expressions almost always conveyed encouragement ("*hope, this helps*", "*I wish you all the best*", "*good luck*").

As a result, our annotation schema was implemented as follows:

(a) annotation was performed on a level of individual posts; annotators were asked to select the most dominant sentiment in the whole post; descriptions of symptoms or diseases were omitted from the sentiment annotation;

(b) every post was marked with only one label; at this stage we did not aim to identify interrelations between sentiments; this task is delegated to the next stage of our study;

(d) finally, every post was labeled by two annotators.

We evaluated agreement between the annotators by using Fleiss Kappa [18], a measure that evaluates agreement for a multi-class manual labeling.

$$\textit{Fleiss Kappa} = (P - Pclass)/(1-Pclass)$$

where P is an average agreement per a post and Pclass is an average agreement per a class. For a five-class problem, the annotators achieved a high agreement: Fleiss Kappa = 0.73 which indicates a strong agreement.

Preparing our data for the machine learning experiments we assigned the five category labels only to posts that both annotators labeled with the same label, e.g., if a post was labeled *encouragement* by two annotators it was put into the *encouragement* category. We introduced a new class *disagreement* for the posts labeled with two different labels. The final number of posts per class was:

Encouragement – 206, *Gratitude* – 88, *Confusion* – 48, *Facts* – 187, *Facts + Encouragement* - 73, and *Uncertain*– 150; total – 752 posts.

V. HEALTHAFFECT

To the best of our knowledge, WordNet-Affect[10] is the only affective lexicon with a highly detailed hierarchy of sentiments [10]. However, comparison of the post vocabulary with WordNet-Affect words revealed that very few words appeared in both given post's texts and the lexicon.

As those matching result were unsatisfactory, we created a specific lexicon which we named HealthAffect. To build HealthAffect, we adapted the Pointwise Mutual Information (PMI) of *word*1 and *word*2 [19]:

$$PMI(word1, word2) = \log_2(p(word1 \,\&\, word2)/(p(word1)\, p(word2)))$$

First, we created a list of all words, bigrams and trigrams of words with frequency ≥ 5 from the unambiguously annotated posts (i.e., we omitted posts marked as *uncertain*). This was a list of candidates (aka *phrases*) to be included in our HealthAffect lexicon.

Next, for each class, we calculated PMI(*phrase*, *class*) as

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

Finally, we calculated Semantic Orientation (SO) for each phrase and for each class as

$$SO(phrase, class) = PMI(phrase, class) - \Sigma\, PMI(phrase, other\_classes)$$

where *other_classes* are all the classes except for the class that Semantic Orientation is calculated for.

After all the possible SOs were computed, each HealthAffect candidate was assigned with the class that corresponded to its maximum SO. Consequently, each candidate was considered an indicator of the class that provided it with the maximum SO. It should be noted that each class got different numbers of indicative candidates. From 459 trigrams with frequency ≥ 5, 46 had their maximum SO for *encouragement*, 40 - for *gratitude*, 139 - for *confusion*, 95 - for *facts* and 139 for *facts + encouragement*.

For each class, we sorted all potential N-grams in decreasing order of SO and selected the equal number of N-grams to represent each class in the lexicon. The number of N-grams was determined as ½ of the minimum *per class* number of N-grams; for example, we used only 20 (=40:2) top trigram indicators for each class. Similarly, we selected 50 bigrams and 25 unigrams and added them to the lexicon.

VI. MACHINE LEARNING EXPERIMENTS

We used personal pronouns, short words, the WordNetAffect terms and the HealthAffect lexicon in four data representations:

- all semantic features (AllSem),
- WordNetAffect and pronouns features (WNAP),
- WordNetAffect features (WNA).
- HealthAffect lexicon (HAL)

We used Naïve Bayes (NB) and K-nearest neighbor (KNN) to classify the messages into 6 classes.

We assessed the learning methods by computing multi-class *Precision (Pr), Recall (R), F-score (F)* and *Accuracy Under the Curve (AUC).*

We used 10-fold cross-validation to select the best classifier. Labeling all examples as the majority class gave the baseline for the performance evaluation: *Pr*= 0.075, *R* = 0.274, *F* = 0.118, *AUC* = 0.491. Table 1 and Table 2 report the empirical results.

TABLE 1: NB RESULTS IN 6-CLASS CLASSIFICATION.

NB results				
Features	*Pr*	*R*	*F*	*AUC*
AllSem	0.408	0.427	0.397	0.685
WNAP	0.324	0.395	0.333	0.661
WNA	0.322	0.350	0.303	0.605
HAL	0.527	0.541	0.518	0.799

[10] http://wndomains.fbk.eu/wnaffect.html

TABLE 2: KNN RESULTS IN 6-CLASS CLASSIFICATION.

KNN results				
Features	*Pr*	*R*	*F*	*AUC*
AllSem	0.330	0.342	0.310	0.598
WNAP	0.287	0.319	0.284	0.591
WNA	0.279	0.322	0.275	0.571
HAL	0.377	0.376	0.340	0.619

Empirical evidence shows that while solving the multi-class classification problem, we significantly improved over the baseline ($P < 0.01$, paired t-test). HealthAffect provided a more accurate classification of sentiments, and NB outperformed KNN on all the data representations. However, for NB, the difference between the best and the worst F-score was as high as 60%, whereas for KNN the difference was < 10%.

VII. CONCLUSION

In this work, we have presented the sentiment analysis of messages posted on medical forums. We stated the sentiment analysis as a multi-class classification problem in which posts were classified into *encouragement*, *gratitude*, *confusion*, *facts*, *facts + encouragement* and *uncertain* categories. We applied the reader-centered manual annotation and achieved a strong agreement between the annotators: *Fleiss Kappa* = 0.73.

Sentiment analysis of online medical discussions differs considerably from the traditional studies of sentiments in consumer-written product reviews, financial blogs and political discussions opinion detection. While in many cases positive and negative sentiment categories are enough, such dichotomies are not sufficient for medical forums. The same can be said about the existing sentiment and affective lexicons: their general terms and labels do not adequately serve for the analysis of medical posts. Thus, new lexical resources sensitive to this specific domain should be created. We presented an ad-hoc method of the lexicon creation which is comparatively easy to implement. We have shown that the lexicon, which we call HealthAffect, provided the best accuracy in machine learning experiments. However, as many other lexical resources, the lexicon requires manual review and filtering.

We used two algorithms, NB and KNN, to solve a multi-class sentiment classification problem. The probability-based NB demonstrated a better performance than KNN. The best F-score was achieved when posts were represented through HealthAffect.

REFERENCES

[1] Federica Bisio, Paolo Gastaldo, Chiara Peretti, Rodolfo Zunino, Erik Cambria: Data intensive review mining for sentiment classification across heterogeneous domains. ASONAM 2013.

[2] Kim, S.-M., E. Hovy. Crystal: Analyzing predictive opinions on the web. EMNLP-CoNLL, 2007.

[3] Dodds, P., K. Harris, I. Kloumann, C. Bliss, C. Danforth. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. PLoS ONE, 6, e26752, 2011.

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

[5] Smith, C. Consumer language, patient language, and thesauri: A review of the literature. *Journal of the Medical Library Association*, 99(2), 2011.

[6] Erik Cambria, Tim Benson, Chris Eckl, Amir Hussain. Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality. Expert Systems with Applications 39, 2012.

[7] Bhuiyan, Touhid and Xu, Yue and Josang, Audun. State-of-the-Art Review on Opinion Mining from Online Customers' Feedback. 9th Asia-Pacific Complex Systems Conference, 2009.

[8] Thelwall, M., & Buckley, K. Topic-based sentiment analysis for the Social Web: The role of mood and issue-related words. Journal of the American Society for Information Science and Technology, 64(8), 2013.

[9] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, Siddharth Patwardhan: OpinionFinder: A System for Subjectivity Analysis. HLT/EMNLP 2005.

[10] Carlo Strapparava, Rada Mihalcea: Learning to identify emotions in text. SAC 2008.

[11] Peter D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002.

[12] P. Subasic, A. Huettner. Affect analysis of text using fuzzy semantic typing. IEEE Transactions on Fuzzy Systems, 2001.

[13] Banea, C., R. Mihalcea, and J. Wiebe. Multilingual sentiment and subjectivity analysis. Book chapter in Multilingual Natural Language Applications: From Theory to Practice. Editors D. M. Bikel and I. Zitouni. Prentice-Hall. 2012.

[14] Staiano, J.,and Guerini, M. 2014. DepecheMood: a Lexicon for Emotion Analysis from Crowd-Annotated News. Proceedings of ACL-2014.

[15] Boiy, Erik; Moens, Marie-Francine. A machine learning approach to sentiment analysis in multlingual Web texts, Information Retrieval, volume 12, issue 5, 2009.

[16] Alexander Osherenko. Opinion Mining and Lexical Affect Sensing. Computer-aided analysis of opinions and emotions in texts. Südwestdeutscher Verlag für Hochschulschriften. 2011.

[17] Bing Liu. Sentiment Analysis and Opinion Mining. Morgan & Claypool, 2012.

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

[19] Turney, P.D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of ACL'02, 2002.