Cogn Comput (2015) 7:609–621
DOI 10.1007/s12559-015-9327-y

What Goes Around Comes Around: Learning Sentiments in Online Medical Forums

Victoria Bobicev[1] · **Marina Sokolova**[2] · **Michael Oakes**[3]

Received: 12 July 2014 / Accepted: 16 March 2015 / Published online: 2 April 2015

Abstract It has been shown that online health-related discussions significantly influence the attitudes and behavioral intentions of the discussion participants. Although empirical evidence strongly supports the importance of emotions in health-related online discussions, there are few studies of the relationship between a subjective language and online discussions of personal health. In this work, we study sentiments expressed on online medical forums. Individual posts are classified into one of five categories. We identified three categories as sentimental (encouragement, gratitude, confusion) and two categories as neutral (facts, endorsement). A total of 1438 messages were annotated manually by two annotators with a strong inter-annotator agreement (Fleiss kappa = 0.737 when the posts were annotated in the context of discussion and Fleiss kappa = 0.763 when the posts were annotated as individual entities). Using machine learning multi-class classification approach, we assess the feasibility of automated recognition of the five sentiment categories. As well as considering the predominant sentiments expressed in individual posts, we analyze transitions between sentiments in online discussions.

Keywords Natural language processing · Sentiment analysis · Machine learning · Discourse analysis · Sentiment transitions

✉ Victoria Bobicev
vika@rol.md

Marina Sokolova
sokolova@uottawa.ca

Michael Oakes
Michael.Oakes@wlv.ac.uk

[1] Technical University of Moldova, Chişinău, Republic of Moldova

[2] Institute for Big Data Analytics, University of Ottawa, Ottawa, Canada

[3] Research Group in Computational Linguistics, University of Wolverhampton, Wolverhampton, UK

Motivation

User-friendly Web 2.0 technologies encourage the general public to actively participate in the creation of Web content. Blogs, social networks and message boards reach out to a global community of Web users. These online texts present personal experience and convey the sentiments and emotions of the authors. These emotion-rich posts are known to be important in setting interaction patterns in online discussions, as emotion-rich text has a strong influence on attitudes and behavioral intentions of the discussion participants [1]. Studies of online sentiments and opinions can help in the understanding of sentiments and opinions of the public at large. Such understanding is especially important for the development of public policies whose success greatly depends on public support, e.g., education, health care, housing and infrastructure. Study of affect and social aspects in online communication is preliminary steps for creation of affective dialogue system in which text-based system–user communication is used to model, generate and present different affective and social interaction scenarios [2].

Effective implementation of healthcare policies relies on the understanding of opinions expressed by the general public. Major healthcare initiatives such as vaccination during pandemics and the incorporation of healthy choices in everyday lifestyles are examples of policies that require

such understanding to be successfully implemented. 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. Users of online communities dedicated to special medical conditions can be exposed to materials where about 90 % of text is dedicated to patient experience [3]. Analysis of health information posted online contributes to identification the sources of information, its dissemination and possible impact on the general public [1, 3–5]. Although empirical evidence strongly supports the importance of emotions in health-related messages [6], there are few studies of the relationship between a subjective language and online discussions of personal health [7].

We focus on sentiments in the medical 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 [8, 9]. In this study, we aimed to identify the most common sentiments expressed in individual posts and the most common pairs and triads of sentiments appearing in the forum discussions. We applied our analysis to data collected from the in vitro fertilization (IVF) medical forum.[1] This forum is designed to bring together women who use IVF treatments in the hope of conceiving. As a result, women constitute 95 % of the forum participants and they post almost 99 % of the messages, although there are occasional messages posted by men. To give a glimpse of the emotionally charged data, we provide 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; it was the antagonist protocol which is supposed to be great when you are over 40 but not so much for me!!

In our sentiment analysis, we applied a twofold approach. First, our goal was to indentify a set of sentiment categories that represents the full spectrum of emotions appearing in the discussions and, at the same time, is compact and segregated enough to be used in a machine learning (ML) empirical study. Next, we compared the domain-specific lexicon HealthAffect and the general sentiment lexicons SentiWordNet, MPQA, SenticNet3, SentiStrength and DepecheMood according to their ability to represent messages in sentiment classification. In those experiments, we used ML multi-class classification technique to automatically recognize sentiment categories in four multi-class classification problems; the messages were represented by HealthAffect and the general sentiment lexicons.

The following results were obtained: We identified the dominant sentiments as *encouragement, gratitude, confusion, facts* and *endorsement*. A total of 1438 messages were annotated manually by two annotators with a strong inter-annotator agreement: Fleiss kappa = 0.737 when the posts were annotated in the context of discussion and Fleiss kappa = 0.763 when the posts were annotated as individual entities. Our empirical evidence shows that HealthAffect provides for more reliable sentiment classification than the other lexicons. Messages represented by HealthAffect were classified with up to 22 % improvement in the *F*-score over the benchmark classification obtained on SentiWordNet representation.

The article is organized as follows: Section "Related Work" presents relevant work in sentiment analysis, section "Data Set" introduces the data set, section "Data Annotation" describes the annotation scheme and its results, section "Correspondence Analysis for Sentiment Sequences" presents the correspondence analysis and results on sentiment sequences, section "Automated Sentiment Recognition" describes sentiment classification experiments, and section "Discussion and Future Work" discusses the results. Preliminary results of this work appeared in [10].

Related Work

Sentiment Analysis

The availability of emotion-rich text has helped to promote studies of sentiments from a boutique science into the mainstream of text data mining. Extraction and analysis of sentiments, opinions, attitudes, emotions, perceptions and intentions is one of the most asked-for types of text analysis, as was pointed out in Seth Grimes' Text Analytics Report 2014.[2] Sentiments and opinions are analyzed in texts of consumer-written product reviews [11], political discussions [12] and forums and blogs [13, 14]. Text analysis of user-written online messages has been motivated by both the demand for such studies and an easy access to the online data [15, 16].

In sentiment analysis, ML methods, affective lexicons and natural language processing (NLP) tools are used to classify text units (e.g., words, sentences, paragraphs) into

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

[2] http://altaplana.com/grimes.html.

sentiment categories [17]. The choice of text unit depends on the goal of the study. Our goal is the identification of sentiments in communication units. Hence, a message is the core text unit of forum communication [14], and we use it as our text unit.

Most sentiment analysis research concentrates on the polarity of discussions, e.g., positive and negative sentiments [13, 18]. A few studies have worked on the distinct universal emotions *anger, fear, enjoyment, sadness and disgust* [19] and dynamic, evolving sets of sentiments [20–22]. We analyzed sentiments that appeared in forum messages and created a set of sentiment labels that were most appropriate for health-related online discussions.

Reliable annotation is essential for a thorough analysis of text, although human errors and bias can be introduced during the annotation process [13]. Multiple annotations of topic-specific opinions in blogs were evaluated in [23]. The authors computed agreement among seven manual annotators for five classification categories, including positive, negative and mixed opinions and non-opinionated and non-relevant categories. Annotation agreement achieved on messages gathered from a medical forum was evaluated in [24]. Multiple annotations were used to categorize tweets into those positive, negative and neutral sentiments in [25]. Analysis of eight Twitter data sets released into the public domain was presented in [18]. This paper also presents an STS-Gold Twitter set of positive, neutral and negative tweets, where annotation agreement among three annotators had a Fleiss kappa score of 0.765. The merits of reader-centric and author-centric annotation models were discussed in [26]. In our current work, we apply the reader-centric annotation model and report the Fleiss kappa obtained after the evaluation of our inter-annotator agreement.

An accurate sentiment classification relies on lexical sources of semantic information. Sentiment research often uses lexicons where words are assigned into opinion, sentiment and emotion categories. However, in independent studies [24, 27], the authors showed that the sentiment categories of SentiWordNet,[3] WordNetAffect[4] and the subjectivity lexicon[5] are not fully representative of health-related emotions. As it is nearly impossible to create a lexicon for every domain, various techniques were proposed for lexicon adaptation, e.g., the feature ensemble model in order to learn a new labeling function which uses feature reweighting [28], contextualised sentiment lexicons for ambiguous terms to be identified and linked to their corresponding polarity [29], objective sentiment words from the SentiWordNet were reevaluated to improve the performance of word-of-mouth sentiment classification [30]. We use HealthAffect, a domain-specific lexicon, to automatically classify sentiments. A preliminary, much smaller version of the lexicon was introduced in [24]. In the current work, we repopulate the lexicon and use a manual filtering to prevent over-fitting the data.

Sentiment propagation is an emerging area in sentiment analysis. Although the relationship between consecutive sentiments is a popular subject of a fine-grained discourse analysis [31], it only recently started to make inroads into text mining. Subjective information posted by a user may affect subjectivity in posts written by other users [8]. Tsai et al. [32] used a two-step approach to evaluate sentiment propagation among related commonsense concepts. Correlations between emotions expressed in consecutive posts were studied in [16, 33, 34]. Until now, health-related sentiment classification has focused on individual messages. Our current work identifies the most common sentiment transitions in pairs and triads of consecutive posts. Studies of sentiment transitions are important if we want to better understand the emotional and cognitive processes of human interactions [35].

Concept-Level Sentiments

Our approach is reminiscent of concept-level sentiment analysis [36]. In the analysis of data, we retrieve and aggregate subjective information about different aspects of IVF treatment. Such information is directly linked with the basic IVF concepts and features and, thus, cannot be identified through a keyword search or the use of general lexical resources.

Another technique associated with concept-level analysis is correspondence analysis, a multivariate technique for analyzing matrices of data. Its implementation in the R programming language is described by Baayen [37]. The technique of correspondence analysis discovers whether groups of words tend to occur in the same messages as each other. Such groups are called "factors," and they are ordered according to their importance in terms of how much of the variation between the messages they explain. The idea for such a representation comes from work by Stanley and Meyer [38], who used another matrix analysis technique called Factor Analysis to plot students' ratings of their emotional states on various occasions on a two-dimensional graph. Stanley and Meyer call the discovered axes (and hence constructs) for representing affective experiences "affective space." We applied correspondence analysis in our study of sentiments.

The ConceptNet knowledge base represents information about contextual, pragmatic information expressed in texts as a graph with node concepts connected by twenty types of semantic relations [39]. The source was used in several text analysis studies [32, 40]. For example, important

[3] http://sentiwordnet.isti.cnr.it/.

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

[5] http://mpqa.cs.pitt.edu/#subj_lexicon.

concepts from ConceptNet were selected and redundant concepts were eliminated using the Minimum Redundancy and Maximum Relevance feature selection techniques [40]. At the same time, large semantic sources exhibit the "curse of dimensionality": The bigger the semantic network is, the more difficult it becomes to process and obtain the required knowledge from it. In our current study, we work with a domain-dependent sentiment lexicon, without building a semantic network.

Reproductive Technologies and Sentiments

Reproductive technologies are hotly debated in modern society. These highly spirited debates are in part due to a multitude of issues connected with the technologies. For example, the most popular reproductive technology—IVF—is linked to an uncertain chance of live birth and discussions of the health of the babies born, ongoing pregnancies, clinical pregnancies, miscarriages, multiple pregnancies, implantation rate, cryopreservation rate, embryo quality and fertilization rate [41], as well as age, obesity, a risk of breast cancer and overall financial costs to society [42]. The complexity of the problem causes the technology's recipients to seek information, advice and guidance not only from medical professionals, but also from peers. The peer connection is increasingly done online, through social media [43].

A meta-study of 19 studies on reproductive technologies published in 1999–2009 listed several reasons for the use of medical forums: (a) information searching—to learn about psychological, physical and social aspects of available treatments, evaluations of alternative treatments—(b) in seeking emotional support—anonymous communication, immediate and constant community access, easy contact with peers [43]. A survey of online infertility support groups showed that 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 up 15.9 and 6.8 %, respectively [9].

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 [25, 44] and online public forums [9, 17]. Sentic PROMs (patient reported outcome measures) analyze semi-structured texts and aggregate the input data [45–47]. This system complements the very structured tool used to monitor patient outcomes in the cases where patients express their opinions and feelings in free text. In our work, we continue studies of online forum data. In forum discussions, patients do not restrict themselves to giving only feedback about hospitals and health services but freely express their opinions, sentiments and attitudes and actively exchange them among each other. Our results can be applied for studies of patient opinions where differentiation between subjective (seeking opinions, emotions and other private states) and non-subjective (seeking factual information) messages is not a trivial task [14].

Data Set

Forums dedicated to specific medical conditions and health-related problems promote sharing of personal experience and disclosure of the emotional state of the forum participants [2]. We collected data from the IVF Web site dedicated to reproductive technologies. The Web site belongs to an infertility outreach resource community created by prospective, existing and past IVF patients. The IVF.ca Web site includes forums: *Cycle Friends*, *Expert Panel*, *Trying to Conceive*, *Socialize*, *In Our Hearts*, *Pregnancy*, *Parenting* and *Administration*.[6] 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 and Surrogacy Buddies* and *Adoption Buddies*. On every sub-forum, new topics are initiated by the forum participants. Depending on the interest among participants, a different number of messages is associated with each topic, e.g., *Human growth hormone and what to expect* has 120 messages posted from Oct 2012, while *Over 40 and pregnant or trying to be* has 3455 messages posted from May 2010.

We wanted the forum to represent a variety of discussions and contain a manageable number of topics and messages. The *IVF Ages* 35+ sub-forum[7] satisfied both requirements, i.e., it had 510 topics and 16,388 messages, where the messages had 128 words on average.[8] Figure 1 illustrates the distribution of posts among the forum topics.

Among those 510 topics, 340 topics contained less than ten 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.

The topics usually had the following structure:

(a) a participant started the theme with a post;

  (i) the initial post usually contained some information about the participant's problem,

[6] www.ivf.ca/forums.

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

[8] We harvested the data in July 2012.

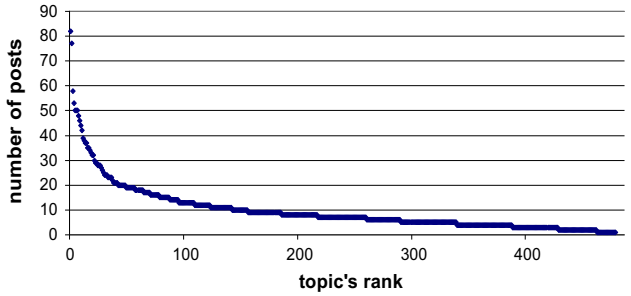

Fig. 1 Number of posts per topic in the IVF Ages 35+ sub-forum

expressed worry, concern, uncertainty and a request for help to the other forum participants.

(b) 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.

(c) the participant who started the topic often thanked other contributors and expressed appreciation for their help and support.

We wanted to identify what sentiments prevail in the forum messages. Our goal was to indentify a set of sentiment categories that represent the full spectrum of emotions appearing in the discussions and, at the same time, being compact and distinct enough to be used in a ML empirical study.

Data Annotation

Annotation of subjectivity can be centered either on the perception of a reader [20] or the author of a text [26]. In the current work, we opted for the reader perception model and asked annotators to analyze the topic's sentiment as it was addressed to the other forum participants. The data annotation was carried out by master students as their practical work for the course "semantic interpretation of text." The students had already completed courses on "computational linguistics" and "natural language processing." Most annotators already had experience in sentiment and opinion annotation. Each annotator independently annotated a set of topics. Each message was annotated by two annotators.

We used 292 randomly selected 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 [24]. In [9], the authors showed that most posts referred to sharing personal experiences, provision of information or advice, expressions of gratitude/friendship, chat, requests for information and expressions of universality (e.g., "we're all in this together"). Hypothesizing that binary sentiment categories (e.g., positive and negative polarity) would be too general and could not adequately cover emotions expressed in health-related messages, we intended to build a set of sentiments that

1. contains sentiment categories specific for posts from medical forums and
2. makes an automated sentiment detection feasible and reliable.

This was the first phase of the annotation process. We used the bottom-up approach to build that set. First, we asked annotators to read several topic discussions and describe the sentiments expressed by the forum participants and the sentiment propagation within these discussions.

We instructed 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 the other participants should be taken into consideration. For example, `I have had a few days now with heartburn/reflux---could be stress, a little achy tummy/pelvic and a tired aching back. More waiting, but getting more hopeful` is a description of symptoms and should not be annotated as subjective. In contrast, I `hope your visit with us infertilies is short and sweet and you get that baby soon!!!` exposes the author's sentiment toward another person.[9] 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.

After collecting the results of the initial annotation, we merged and summarized the annotations. That resulted in 35 sentiment types which we arranged 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; and
- ***gratitude***, which included thankfulness.

[9] All examples preserve original spelling and grammar.

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, 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.*

Not all posts had an emotional content. 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 include openly negative sentiments. We considered *confusion* as a non-positive label. *Encouragement* and *gratitude* were considered positive labels, and *facts* and *endorsement*—neutral.

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

We evaluated agreement between the annotators by using Fleiss kappa [48], a measure that evaluates agreement for a multi-class manual labeling.

$$\text{Flesiss kappa} = (P - P_{\text{class}})/(1 - P_{\text{class}})$$

where P is the average agreement per class observed and P_{class} is the average agreement per class which would be obtained by chance.

Despite the challenging data, we obtained Fleiss kappa $= 0.737$ which indicated a strong agreement between annotators [23]. 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 when the posts were annotated in the context of discussions. The final distribution of posts among sentiment classes is presented in Table 1.

Table 1 Class distribution of the IVF posts

Classification category	# of posts	%
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

Correspondence Analysis for Sentiment Sequences

We applied correspondence analysis [37] to recognize the affective groups of the most frequent words found in the data. We used the messages from the ART_over_35 topic, missing out only the very short ones. The messages are numbered in the order they appear in the discussion. As input, we produced a matrix where the columns corresponded to the 500 most frequent words in the ART_over_35 text collection and the rows each corresponded to one individual message. Since we were mainly interested in sentiment words, this original matrix was reduced by retaining only those columns corresponding to the 41 words conveying sentiments such as "best," "better" and "congratulations." From the list, 28 words were indicative of sentiment categories and appear in HealthAffect (e.g., able, against, interested, recommended, risk) and 13 words were not indicative of specific categories and thus do not appear in HealthAffect (e.g., "avoid," "luxury").

The technique of correspondence analysis discovers whether groups of words tend to occur in the same messages as each other. Such groups are called "factors," and they are ordered according to their importance in terms of how much of the variation between the messages they explain. The graph below (Fig. 2) was produced by correspondence analysis and shows to what extent each word and each message is related to the two main factors. Only those words which are significantly associated with the factors ($p < 0.1$) are shown in the graph. The group of words making up the first factor explain 24.5 % of the

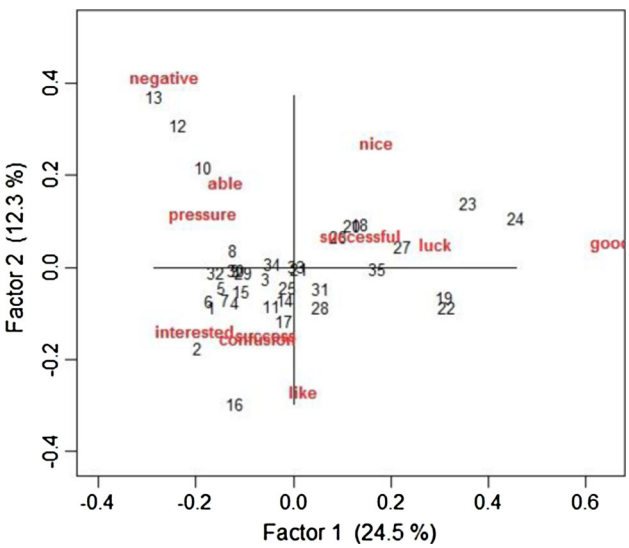

Fig. 2 Correspondence analysis of sentiments

variation between the posts, while those making up the second factor explain 12.3 % of this variation. The identified words occur together in three main groups: *concern, support and good will*, and *desire to know*. The groups form the affective *author-centric* space of the topic and can be representative of the affective space of the IVF discussion [38].

The graph shows that in the top left quadrant are words which occur together in messages expressing *concern* for the future, as in "I don't feel able to handle the negative pressure." In the top right quadrant are words which appear in messages of *support and good will*, such as "successful," "luck" and "good." Finally, in the lower left quadrant are words found in messages expressing a *desire to know*, "interested," "confusion," "success" and "like" (as in "I'd like to know the chances of success"). Most of the early messages (from 1 to 17) are in the topic-opening "desire to know" quadrant, apart from a short exchange of *anxious* messages (10, 12 and 13). There are then a series of encouraging messages (from 18 to 27), while the last few messages are more neutral, scoring about 0 on Factor 2, and slightly negative on Factor 1. Although this is not apparent from the graph, they correspond to messages where people looked back on their own experiences of IVF in a neutral, unemotional way.

Note that there are no significant words in the fourth quadrant. We show only the words which were significantly associated with the factors ($p < 0.1$), and there are none of these in the fourth quadrant. Although messages 19 and 22 are in the fourth quadrant, the most important thing is that they score highly on Factor 1 (i.e., over to the right-hand side of the graph). The last few messages (29–35) are not in the fourth quadrant, but appear about half way up (very close to the horizontal axis) mostly on the left. In our case, the poles of the affect space were positive–negative for Factor 1 and question–response for Factor 2.

To further identify sentiments that reinforced themselves and sentiments that were likely to trigger changes, we computed the distribution of sentiment pairs and triads in consecutive messages. We found that the most frequent sequences consisted mostly of facts and/or encouragement: 39.5 % in total. These two categories were most likely to propagate through next messages. The most frequent change was from endorsement to facts (6.1 % in total). Approximately 10 % of sentiment pairs were factual and/or encouragement followed by gratitude. Confusion was followed by facts and encouragement in 80 % of cases. The most frequent triad containing confusion was confusion, facts, facts. That sentiment transition showed a high level of support among the forum participants. Other less frequent sequences appeared when a new participant added her post in the flow. Tables 2 and 3 list the results. Figure 3 shows the most frequent pairs of sentiments. The node size corresponds to the proportion of the sentiment in the data, and the line weight to the proportion of the transaction.

Our next goal was to find a method that reliably identified sentiments in a large number of the forum texts. This method had to be general enough to accommodate the diversity of natural language expressions appearing in the forum data and exhaustive enough to recognize the opinions expressed toward the IVF treatment. We also wanted this method to be based on a compact set of features, thus avoiding the pitfalls of high dimensionality of feature space in text representation.

In this work, we concentrate on the classification of individual messages. Sentiment classification of pairs and triads of messages is left for future work.

Automated Sentiment Recognition

The first stage of our study identified that the forum messages belonged to five sentimental and neutral categories. For automated sentiment classification, we tested the multi-categorical SentiWordNet [49] and the MPQA subjectivity lexicon [50] which recognizes only positive and negative polarity of its terms. We also tested several lexicons with sentiment information which were announced recently: SentiStrength, sentiment analysis software [51], contains the list of English words that express emotions, SenticNet 3 [21] is a knowledge base that contains information about the semantics and sentics associated with multi-word expressions, and DepecheMood [22] that contains more than 37,500 terms that have been assigned numerical values representing degrees of eight sentiment categories: afraid, amused, angry, annoyed, dont_care, happy, inspired and sad. For every lexicon mentioned here, we created a set of features that will represent our data in ML experiments. We used different procedures to create the sets. The procedures were based on the characteristics of the lexicons:

1) SentiWordNet was created by assigning to each synset of WordNet three sentiment scores: positivity, negativity and objectivity. Every synset in SentiWordNet has all three scores simultaneously. Thus,

Table 2 Most frequent sequences of two sentiments and their occurrence in the data

Sentiment pairs	Occurrence	%
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 Most frequent triads of sentiments and their occurrences in the data

Sentiment triads	Occurrence	%
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

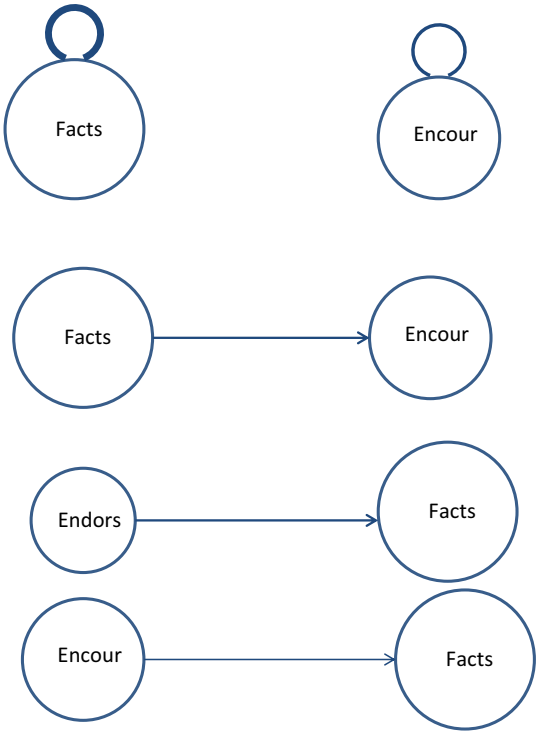

Fig. 3 Most frequent pairs of sentiments

there were positive terms with negative and objective scores equal to zero and positive score greater than zero; negative terms with positive and objective scores equal to zero and negative score greater than zero; objective terms with positive and negative scores equal to zero and objective score greater than zero; neutral terms with all scores equal to zero; and there were also ambiguous terms with several scores greater than zero. We selected only positive and negative synsets and searched for the presence of every term of these synsets in our texts. Only unambiguously positive or negative terms that were present in the texts were used as features in the experiments. Further, we used SentiWordNet as the benchmark representation for the comparison of empirical results.

2) MPQA lexicon assigns words with opinion clues, where the initial list of subjectivity clues from [52] was expanded [50] with positive and negative word lists from the General Inquirer.[10] Each clue has the following structure: `type=strongsubj len=1 word1=abuse pos1=verb stemmed1=y priorpolarity=negative` where `priorpolarity` values can be: positive, negative, both, neutral. As in the previous case, we selected only words with positive or negative polarity clues and compared them with the words which appeared in our texts.
3) SenticNet is a publicly available semantic resource for concept-level sentiment analysis. It associates polarity scores with ConceptNet concepts which are represented as words and multi-word expressions. Our downloaded version contained an XML file with more than 13,000 concepts. Each concept was associated with five characteristics: pleasantness, attention, sensitivity, aptitude and polarity. We used only the polarity attribute and extracted terms with nonzero polarity which were present in our texts. (1) SentiStrength is sentiment analysis (opinion mining) software. Its simplified version is free for academic research. The downloadable version contains Java code and lexicons in editable textual format. We used the lexicon with a polarity score associated with each word. Some terms in SentiStrength are stemmed; while comparing these terms with the words from our texts, we searched for all words that matched this stem. Thus, these stems could correspond to several words in our list of features.
4) DepecheMood is a high-coverage lexicon of approximately 37,500 terms annotated with emotion scores. This lexicon was crowdsourced from rappler.com news articles. Rappler's mood meter, a small interface, offers the readers the opportunity to click on the emotion that a given news article made them feel. Numerous votes have been collected, and document-by-emotion matrix was built which was transformed into a word–emotion matrix, e.g., concerned - 0.129322883 AFRAID, 0.100615215 AMUSED, 0.170474974 ANGRY, 0.161903853 ANNOYED, 0.120271172 DONT_CARE, 0.108064155 HAPPY, 0.098734566 INSPIRED and 0.110613182 SAD.
5) We also used the domain-specific lexicon HealthAffect introduced in [24]. To build the lexicon, we adapted the pointwise mutual information (PMI) approach [53]:

[10] http://www.wjh.harvard.edu/~inquirer/.

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

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). This was a list of candidates to be included in our HealthAffect lexicon. Next, for each class and each candidate, we calculated PMI(*candidate, class*) as

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

Next, we calculated semantic orientation (SO) for each candidate and for each class as

$$SO(\text{candidate}, \text{class}) = \text{PMI}(\text{candidate}, \text{class}) - \sum \text{PMI}(\text{candidate}, \text{other\_class})$$

where *other_classes* include all the classes except the class that SO is calculated for. After all, the possible SO was computed and 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. To avoid the over-fitting pitfall, we manually reviewed and filtered out conversation-specific terms (i.e., personal and brand names, geolocations, dates) and non-relevant elements, such as stop words and their combinations (since_then, that_was_the, to_do_it, so_you). Table 4 presents all the described lexicons.

Further in the ML experiments, the extracted terms are used as features to represent the messages. The classification's performance was evaluated through four multi-class classification results:

- 6-class classification where all 1438 posts were classified into 6 classes, including ambiguous.
- 5-class classification where 1269 unambiguous posts were classified into 5 classes.
- 4-class classification where all 1269 unambiguous posts were classified into *encouragement, gratitude, confusion* and neutral (i.e., *facts* and *endorsement*).
- 3-class classification of 1269 unambiguous posts into positive (*encouragement, gratitude*), negative (*confusion)* and neutral (*facts, endorsement*).

Table 4 Total number of terms and the number of extracted terms for the six lexicons

Semantic lexicon	Num of terms	Num of extracted terms
SentiWordNet	11,7659	3725
MPQA	8221	1418
SenticNet 3	13,741	1342
SentiStrength	2546	1131
DepecheMood	37,772	4467
HealthAffect	1190	1190

As is common in multi-class classification problems, the sentiment categories were unequally represented in the data, e.g., 34 % for the largest category versus approximately 10 % for the small categories in the 6-class and 5-class problems. We considered that this distribution was not skewed enough to invoke undersampling and oversampling techniques used on more imbalanced data [54]. Although in the 4-class and 3-class problems the imbalance had increased to 52 % for the largest category versus 10.3 % for the smallest category, we opted to keep the same learning setting for direct comparison of the learning results.

We applied Naïve Bayes (NB), NBText, NBMultinomial, SVM, Decision Trees and KNN from the WEKA toolkit. We considered that the number of individual posts was sufficient for tenfold cross-validation. To select the best classifier, we used standard metrics of text classification performance. We computed multi-class versions of *Precision (P), Recall (R)*, balanced *F-score (F)* and AreaUnderCurve (AUV):

$$\begin{aligned} \text{Precision} &= \text{tp}/(\text{tp}+\text{fp}) \\ \text{Recall} &= \text{tp}/(\text{tp}+\text{fn}) \\ F\text{score} &= 2\text{tp}/(2\text{tp}+\text{fn}+\text{fp}) \\ \text{AUC} &= (1/2)(\text{tp}/(\text{tp}+\text{fn}) + \text{tn}/(\text{tn}+\text{fp})) \end{aligned}$$

where tp = correctly recognized positive examples, tn = correctly recognized negative examples, fp = negative examples recognized as positives, and fn = positive examples recognized as negatives. Although Matthews' coefficient [55] can work well in multi-class optimization, we opted for the *F*-score and AUC as these are more commonly used measures in our discipline.

We applied NB, NBText (DMNBText), NBMultinomial, SVM, Decision Trees and KNN from the WEKA's toolkit. To select the best classifier, we used tenfold cross-validation and computed the measures listed above. We assessed classification based on *F*-score. WEKA computes the performance measures for each class individually and the weighted average of the measures for overall results; weights are assigned according to the number of instances with that particular class label. The best *F*-score and other corresponding measures for each class are reported in Tables 5, 6, 7 and 8.

The results reported above show considerable consistency: DMNBText and NBMultinomial algorithms

outperformed other algorithms in sentiment classification, with exception of *endorsement* classification in 6-class and 5-class problems where SVM was the best; among lexicons, HealthAffect and DepecheMood provided for the best classification of individual classes:

- The highest precision occurred for *gratitude/positive*, except for the 4-class problem where it was the second best. If misclassified, *gratitude* was commonly labeled as *encouragement*. Posts in the *gratitude* class tend to be the shortest and contain only 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 highest recall occurred for *facts/neutral* in all the four problems, the biggest class in the data. However, precision for this class was uneven and depended on the structure of other classes.

The best overall results appear in Tables 9, 10, 11 and 12. We report DMNBText and NBMultinomial, as they achieved the best and second best results. To put empirical evidence in perspective, we used the majority class baseline and designated SentiWordNet as the benchmark representation. SentiWordNet is commonly used in other studies, thus making comparison of the results feasible in future. The best results for each metric are in bold, the second best results are in bold italic, the benchmark are in italic.

The overall classification results improved when we decreased the number of sentiment categories; hence, uncertainty was reduced for the algorithms. The *F*-score obtained on the HealthAffect features was the best in all experiments. At the same time, the results provided by DepecheMood were better than the results provided by remaining lexicons. We hypothesize that the critical characteristic of DepecheMood was its ability to recognize several sentiments, not only positive and negative ones.

Discussion and Future Work

We have presented the 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*. We have run four multi-class sentiment classification problems on which we compared the performance of ML algorithms and the ability of sentiment lexicons to represent the data. We have shown that Naïve Bayes Text and Naïve Bayes Multinomial provide reliable sentiment classification for each class individually and for overall classification. In the four problems, the domain-based HealthAffect provided for a higher *F*-score than DepecheMood, SentiWordNet, MPQA, SenticNet3 and SentiStrength. DepecheMood provided for a higher *F*-score than the other general sentiment lexicons.

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 annotated in the context of discussions and Fleiss kappa $= 0.76$ for posts annotated as separate instances). 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 and (b) to annotate every sentence of the posts.

In the current work, we identified message 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. Identifying sentiment propagation among related semantic concepts is another venue of the future work.

We plan to use the results obtained in this study for the analysis of discussions related to other highly debated healthcare policies. One future possibility is to construct a

Table 5 Best *F*-score and corresponding precision, recall and AUC of each class for the 6-class problem

Class	#	Precision	Recall	*F*-score	AUC	Lexicon	Algorithm
Facts	494	0.512	0.720	0.599	0.737	HealthAffect	DMNBText
Encouragement	333	0.603	0.678	0.638	0.861	HealthAffect	NBMulti
Endorsement	166	0.313	0.301	0.307	0.741	HealthAffect	SVM
Confusion	146	0.475	0.462	0.469	0.824	HealthAffect	NBMulti
Gratitude	131	0.632	0.423	0.507	0.858	HealthAffect	DMNBText
Ambiguous	168	0.271	0.128	0.174	0.567	HealthAffect	DMNBText

Table 6 Best *F*-score and corresponding precision, recall and AUC of each class for the 5-class problem

Class	#	Precision	Recall	*F*-score	AUC	Lexicon	Algorithm
Facts	494	0.580	0.757	0.657	0.760	HealthAffect	DMNBText
Encouragement	333	0.646	0.699	0.671	0.869	HealthAffect	NBMulti
Endorsement	166	0.358	0.325	0.341	0.751	HealthAffect	SVM
Confusion	146	0.572	0.572	0.572	0.858	HealthAffect	NBMulti
Gratitude	131	0.735	0.469	0.573	0.878	HealthAffect	DMNBText

Table 7 Best *F*-score and corresponding precision, recall and AUC of each class for the 4-class problem

Class	#	Precision	Recall	*F*-score	AUC	Lexicon	Algorithm
Neutral	660	0.671	0.853	0.751	0.774	DepecheMood	DMNBText
Encouragement	333	0.650	0.666	0.658	0.866	HealthAffect	NBMulti
Confusion	146	0.574	0.559	0.566	0.855	HealthAffect	NBMulti
Gratitude	131	0.624	0.562	0.591	0.870	HealthAffect	NBMulti

Table 8 Best *F*-score and corresponding precision, recall and AUC of each class for the 3-class problem

Class	#	Precision	Recall	*F*-score	AUC	Lexicon	Algorithm
Neutral	660	0.713	0.811	0.759	0.788	DepecheMood	DMNBText
Positive	464	0.715	0.742	0.728	0.846	DepecheMood	DMNBText
Negative	146	0.517	0.524	0.521	0.842	HealthAffect	NBMulti

Table 9 Classification results for 6 classes, the baseline *F*-score = 0.171

	DMNBText				NBMultinomial			
	P	*R*	*F*	AUC	*P*	*R*	*F*	AUC
SentiWordNet	*0.431*	*0.469*	*0.424*	*0.711*	*0.419*	*0.446*	*0.385*	*0.678*
MPQA	0.403	0.449	0.388	0.701	0.424	0.449	0.394	0.684
SenticNet 3	0.436	0.457	0.399	0.713	0.411	0.446	0.393	0.688
SentiStrength	0.444	0.474	0.407	0.719	0.417	0.446	0.394	0.699
DepecheMood	0.437	0.471	0.432	0.723	0.418	0.449	0.384	0.680
HealthAffect	**0.490**	**0.509**	***0.484***	**0.756**	***0.483***	***0.502***	**0.491**	**0.756**

The best *F*-score is **0.491**, the second best—***0.484***

Table 10 Classification results for 5 classes, the baseline *F*-score = 0.215

	DMNBText				NBMultinomial			
	P	*R*	*F*	AUC	*P*	*R*	*F*	AUC
SentiWordNet	*0.531*	*0.554*	*0.518*	*0.734*	*0.501*	*0.507*	*0.453*	*0.699*
MPQA	0.506	0.528	0.476	0.725	0.489	0.513	0.463	0.711
SenticNet 3	0.493	0.514	0.461	0.726	0.482	0.503	0.459	0.702
SentiStrength	0.516	0.539	0.484	0.747	0.509	0.519	0.480	0.723
DepecheMood	0.530	0.549	0.519	0.745	0.497	0.501	0.454	0.700
HealthAffect	**0.590**	**0.595**	***0.580***	***0.793***	***0.578***	***0.589***	**0.582**	**0.801**

The best *F*-score is **0.582**, the second best—***0.580***

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 wide spectrum of sentiments related to healthcare issues.

Table 11 Classification results for 4 classes, the baseline *F*-score = 0.353

	DMNBText				NBMultinomial			
	P	*R*	*F*	AUC	*P*	*R*	*F*	AUC
SentiWordNet	*0.625*	*0.645*	*0.611*	*0.777*	*0.598*	*0.605*	*0.552*	*0.751*
MPQA	0.585	0.603	0.550	0.746	0.583	0.605	0.556	0.744
SenticNet 3	0.615	0.618	0.557	0.757	0.575	0.601	0.556	0.741
SentiStrength	0.625	0.622	0.566	0.761	0.599	0.598	0.553	0.745
DepecheMood	0.625	0.648	0.618	0.789	0.596	0.614	0.552	0.757
HealthAffect	***0.670***	***0.669***	***0.657***	***0.809***	**0.676**	**0.678**	**0.667**	**0.828**

The best *F*-score is **0.667**, the second best—***0.657***

Table 12 Classification results for 3 classes, the baseline *F*-score = 0.353

	DMNBText				NBMultinomial			
	P	*R*	*F*	AUC	*P*	*R*	*F*	AUC
SentiWordNet	*0.660*	*0.694*	*0.665*	*0.787*	*0.662*	*0.680*	*0.651*	*0.773*
MPQA	0.652	0.677	0.645	0.770	0.626	0.641	0.618	0.764
SenticNet 3	0.648	0.681	0.643	0.768	0.637	0.654	0.631	0.749
SentiStrength	0.652	0.669	0.630	0.766	0.632	0.643	0.617	0.755
DepecheMood	0.669	**0.702**	***0.675***	0.802	0.679	0.693	0.663	0.781
HealthAffect	***0.681***	0.681	0.672	***0.807***	**0.697**	***0.697***	**0.697**	**0.827**

The best *F*-score is **0.697**, the second best—***0.675***

References

1. Kareklas I, Muehling DD, Weber TJ. Reexamining health messages in the digital age: a fresh look at source credibility effects. J Advert. 2015. Available at SSRN: http://ssrn.com/abstract=2556998.
2. Skowron M, Rank S, Świderska A, Küster D, Kappas A. Applying a text-based affective dialogue system in psychological research: case studies on the effects of system behaviour, interaction context and social exclusion. Cogn Comput. 2014;6(4):872–91.
3. Sillence E, Briggs P. Trust and Engagement in Online Health A Timeline Approach. Handb PsycholCommun Technol. 2015;33:469–87.
4. Chee B, Berlin R, Schatz B. Measuring population health using personal health messages. In: Proceedings of AMIA symposium; 2009. p. 92–6.
5. Sudau F, Friede T, Grabowski J, Koschack J, Makedonski P, Himmel W. Sources of information and behavioral patterns in online health forums: observational study. J Med Internet Res. 2014;16(1):e10. doi:10.2196/jmir.2875.
6. Pennebaker JW, Chung CK. Expressive writing, emotional upheavals, and health. In: Evans JF, editor. Wellness & writing connections: writing for better physical, mental, and spiritual health. Enumclaw, WA: Idyll Arbor, Inc.; 2010. p. 33–112.
7. Smith CA. Consumer language, patient language, and thesauri: a review of the literature. J Med Libr Asso. 2011;99(2):135.
8. Zafarani R, Cole W, Liu H. Sentiment propagation in social networks: a case study in live journal. Advances in social computing (SBP 2010). Springer Berlin Heidelberg; 2010. p. 413–20.
9. Malik S, Coulson N. Coping with infertility online: an examination of self-help mechanisms in an online infertility support group. Patient Educ Couns. 2010;81(2):315–8.
10. Bobicev V, Sokolova M, Oakes M. Recognition of sentiment sequences in online discussions, SocialNLP-COLING; 2014.
11. Bisio F, Gastaldo P, Peretti C, Zunino R, Cambria E. Data intensive review mining for sentiment classification across heterogeneous domains. In: Advances in social networks analysis and mining (ASONAM). 2013 IEEE/ACM International Conference, IEEE; 2013. p. 1061–67.
12. Poggi I, D'Errico F. Multimodal acid communication of a politician ESSEM@AI*IA, vol. 1096 of CEUR workshop. In: Proceedings, CEUR-WS.org; 2013. p. 59–70.
13. Cieliebak M, Dürr O, Uzdilli F. Potential and limitations of commercial sentiment detection tools. In: Battaglino C, Bosco C, Cambria E, Damiano R, Patti V, Rosso P, editors. Proceedings of the First International Workshop on Emotion and sentiment in social and expressive media: approaches and perspectives from AI (ESSEM 2013) A workshop of the XIII International Conference of the Italian Association for Artificial Intelligence (AI*IA 2013). 2013.
14. Biyani P, Bhatia S, Caragea C, Mitra P. Using non-lexical features for identifying factual and opinionative threads in online forums. Knowl-Based Syst. 2014;69:170–8.
15. Dodds P, Harris K, Kloumann I, Bliss C, Danforth C. Temporal patterns of happiness and information in a global social network: hedonometrics and twitter. PLoS One. 2011;6:e26752.
16. Chmiel A, Sienkiewicz J, Thelwall M, Paltoglou G, Buckley K, Kappas A, Hołyst JA. Collective emotions online and their influence on community life. PloS One. 2011;6(7):e22207.
17. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-Based Methods for Sentiment Analysis. Comput Linguist. 2011;37(2):267–307.
18. Saif H, Fernandez M, He Y, Alani H. Evaluation datasets for twitter sentiment analysis. A survey and a new dataset, the STS-gold. First ESSEM workshop; 2013.
19. Ekman P. An argument for basic emotions. Cogn Emot. 1992;6:169–200.
20. Strapparava C, Mihalcea R. Semeval-2007 task 14: affective text. In: Proceedings of the 2008 ACM symposium on applied computing; 2008.
21. Cambria E, Hussain A. Sentic computing: techniques, tools, and applications. New York: Springer; 2012.

22. Staiano J, Guerini M. DepecheMood: a Lexicon for emotion analysis from crowd-annotated news. In: Proceedings of ACL-2014; 2014.
23. Osman D, Yearwood J, Vamplew P. Automated opinion detection: Implications of the level of agreement between human raters. Inf Process Manag. 2010;46:331–42.
24. Sokolova M, Bobicev V. What sentiments can be found in medical forums? In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP. Shoumen, Bulgaria: INCOMA Ltd; 2013. p. 633–39.
25. Bobicev V, Sokolova M, Jaffer Y, Schramm D. Learning sentiments from tweets with personal health information. In: Proceedings of Canadian AI 2012. Springer; 2012. p. 37–48.
26. Balahur A, Steinberger R. Rethinking sentiment analysis in the news: from theory to practice and back. In: Proceedings of the 1st workshop on opinion mining and sentiment analysis; 2009.
27. Goeuriot L, Na J, Kyaing W, Khoo C, Chang Y, Theng Y and Kim J. Sentiment lexicons for health-related opinion mining. In: Proceedings of the 2nd ACM SIGHIT international health informatics symposium, ACM; 2012. p. 219–25.
28. Xia R, Zong C, Hu X, Cambria E. Feature ensemble plus sample selection: domain adaptation for sentiment classification. Intell Syst IEEE. 2013;28(3):10–8.
29. Weichselbraun A, Gindl S, Scharl A. Extracting and grounding context-aware sentiment lexicons. IEEE Intell Syst. 2013;28(2): 39–46.
30. Hung C, Lin HK. Using objective words in SentiWordNet to improve word-of-mouth sentiment classification. IEEE Intell Syst. 2013;28(2):47–54.
31. Smith P, Lee M. Acknowledging discourse function for sentiment analysis. In: Proceedings of CICLing; 2014.
32. Tsai ACR, Wu CE, Tsai RTH, Hsu JYJ. Building a concept-level sentiment dictionary based on commonsense knowledge. IEEE Intell Syst. 2013;28(2):22–30. doi:10.1109/MIS.2013.25.
33. Tan C, Lee L, Tang J, Jiang L, Zhou M, Li P. User-level sentiment analysis incorporating social networks. In: Proceedings of the 17th ACM SIGKDD international conference on KDDM; 2011.
34. Hassan A, Abu-Jbara A, Radev D. 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; 2012.
35. Esposito A, Fortunati L, Lugano G. Modeling emotion, behavior and context in socially believable robots and ict interfaces. Cogn Comput. 2014;6:623–7.
36. Cambria E. An introduction to concept-level sentiment analysis. In: Proceedings of micai 2013, Springer; 2013. p. 478–83.
37. Baayen H. Analysing linguistic data: a practical introduction to statistics using R. New York: Cambridge University Press; 2008.
38. Stanley DJ, Meyer JP. Two-dimensional affective space: a new approach to orienting the axes. Emotion. 2009;9(2):214–37.
39. Havasi C, Speer R, Alonso J. ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In: Proceedings of recent advances in natural language processing; 2007.
40. Agarwal B, Poria S, Mittal N, Gelbukh A, Hussain A. Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. Cogn Comput. 2015.41.
41. Mantikou E, Youssef MAFM, van Wely M, van der Veen F, Al-Inany HG, Repping S, Mastenbroek S. Embryo culture media and IVF/ICSI success rates: a systematic review. Hum Reprod Update. 2013;19(3):210–20.
42. Pantasri T, Norman RJ. The effects of being overweight and obese on female reproduction: a review. Gynecol Endocrinol. 2013;30(2):90–4.
43. Zillen N. Internet use of fertility patients: a systemic review of the literature. J Reprod Med Endocrinol. 2011;8(4):281–7.
44. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PloS One. 2010;5(11):e14118.
45. Cambria E, Hussain A, Havasi C, Eckl C, Munro J. Towards crowd validation of the UK national health service. In: ACM WebSci. Raleigh; 2010.
46. Cambria E, Hussain A, Eckl C. Bridging the gap between structured and unstructured health-care data through semantics and sentics. In: ACM Web Sci, 3rd International Conference on Web Science. Germany; 2011.
47. 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. 2012;39(12):10533–43.
48. Nichols T, Wisner P, Cripe G, Gulabchand L. Putting the kappa statistic to use. Qual Assur J. 2010;13:57–61.
49. Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the 7th conference on international language resources and evaluation; 2010. p. 2200-04.
50. Wiebe Janyce, Wilson T, Cardie C. Annotating expressions of opinions and emotions in language. Lang Resour Eval. 2005;39: 165–210.
51. Thelwall M, Buckley K, Paltoglou G. Sentiment strength detection for the social Web. J Am Soc Inf Sci Technol. 2012;63(1): 163–73.
52. Riloff E, Wiebe J. Learning extraction patterns for subjective expressions. EMNLP-2003; 2003.
53. Turney PD. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of ACL'02. Philadelphia, Pennsylvania, p. 417–24.
54. Cai Q, He H, Man H. Imbalanced evolving self-organizing learning. Neurocomputing. 2014;133:258–70.
55. Jurman G, Riccadonna S, Furlanello C. A comparison of MCC and CEN error measures in multi-class prediction. PloS One. 2012;7(8):e41882.