Can I hear you? Sentiment Analysis on Medical Forums

Tanveer Ali
University of Ottawa
tali028@uottawa.ca

David Schramm
University of Ottawa and CHEO
dschramm@ottawahospital.on.ca

Marina Sokolova
University of Ottawa and CHEO
sokolova@uottawa.ca

Diana Inkpen
University of Ottawa
Diana.Inkpen@uottawa.ca

Abstract

Text mining studies have started to investigae relations between positive and negative opinions and patients' physical health. Several studies linked the personal lexicon with health and the health-related behavior of the individual. However, few text mining studies were performed to analyze opinions expressed in a large volume of user-written Web content. Our current study focused on performing sentiment analysis on several medical forums dedicated to Hearing Loss (HL). We categorized messages posted on the forums as *positive*, *negative* and *neutral.* Our study had two stages: first, we applied manual annotation of the posts with two annotators and have 82.01% overall agreement with kappa 0.65 and then we applied Machine Learning techniques to classify the posts.

1 Introduction

Natural language statements can be divided into two categories: factual and emotional. Factual statements can be expressed with a few topic keywords, while emotional statements express sentiments of the statement's author and require a more complex analysis than the factual ones.

Sentiment Analysis is often regarded as classifying and identifying the subjective information in the natural language text. In its application, Sentiment Analysis aims to detect the sentiments (e.g., opinions and emotions) of the speaker of the statement. Sentiments are characterized by polarity, intensity, strength and immediacy.

In the current study, we focus on the polarity of sentiments that are expressed in messages posted on medical forums. Polarity can be binary (e.g., positive vs. negative) or multi-categorical (e.g., positive, negative and unknown). Below we list examples found in online discussions about hearing aids.

Positive[1]

This has the beneficial effect of making the quieter sounds audible but not blowing your head off with the louder sounds.

Neutral/Unknown

Now, you'll hear some people saying that compression is bad, and linearity is good, especially for music.

Negative

Someone with 50 DB hearing aid gain with a total loss of 70 DB may not know that the place is producing 107 DB since it may not appear too loud to him since he only perceives 47 DB.

In this work, we classified the subjective sentences into positive, negative and neutral. We have identified different syntactic features, i.e., patterns / rules (Yi and Nasukawa, 2003) which can indicate subjectivity and polarity of the sentences. The dataset of 3515 sentences from 26 threads were manually annotated by two annotators having overall agreement of 82.01% and kappa 0.65 which indicates substantial agreed data.

Our experiments with different combinations of features using different classifiers have shown significant improvement in performance over the baseline. For example, with the Naïve Bayes classifier, the F1-score was10.5%better.

The rest of the paper is organized as follows: we discuss the sentiment analysis of health-related online messages, then we introduce our data; next we discuss the Subjectivity Lexicon and the features we use to represent the data, the analysis of the manual annotation and the machine learning classification results, before we conclude the presentation.

[1]All textual examples keep the original spelling and grammar.

International Joint Conference on Natural Language Processing, pages 667–673,
Nagoya, Japan, 14-18 October 2013.

2 Related Work

Very little work has been done in sentiment analysis on health-related forums. In (Goeuriot et al., 2012), the authors have built a medical domain lexicon in order to perform classification on a dataset that they collected from a website called Drug Expert. The dataset contains user reviews on drugs with ratings from 0 to 10 (Negative to positive) and they achieved F-score of 0.62 for the positive class, 0.48 for the negative class and 0.09 for the neutral class. The authors have performed the polarity detection on this dataset which already contains subjective information (opinions) about users' experience with particular drugs. However, in our case, we have extracted messages from health forums which contain mixed subjective and non-subjective information.

Users express their sentiments differently on forums compared to the way they express opinions when providing reviews or sharing messages on social networks. Bobicev et al. (2012) have analyzed sentiments in Twitter messages using some statistical features based on the occurrence and correlation among words with the class labels of the training set. However, we have identified the correlation of phrases within sentences for predicting subjectivity and polarity.

3 Building the Dataset

Surgeries related to HL are the most common surgeries in North America; thus, they affect many patients and their families.

However, there are only a few health forums dedicated to Hearing Loss (HL). Hence, we did not have an access to a high volume of data. Also, we need forum discussions, i.e., threads, which consist of more opinionated messages rather than questions and answers about the medical problems.

For the sentiment analysis, we have chosen a critical domain of HL problems: opinions about Hearing Aids. To the best of our knowledge, no relevant previous work was done in this area. For our dataset, we have collected individual posts from 26 different threads on three health forums[2].

3.1 Data Description

The initial collection of data contains about 893 individual posts from 34 threads. They were extracted using the XPath query by using the Google Chrome extension "XPathHelper".

This data was filtered and reduced to 607 posts in 26 threads (Table 1), by removing the threads where people discussed the factual information about a specific problem or disease and which do not contain any sentiments or opinions. Statistics, like average posts per person, were measured for filtering the data. For example, threads with more than 100 posts were removed, as threads with a large number of posts deviated from the main topic of discussion.

	Threads	Posts	Avg. posts per person
www.hearingaidforums.com	7	185	2.9
www.medhelp.org	9	105	2.77
www.alldeaf.com	10	317	1.93
Total	26	607	2.53

Table 1. Filtered dataset collection statistics

We split the data from individual threads into sentences using our version of a regular expression-based sentence splitter. We partly removed noise from the text by removing sentences containing very few words (i.e., less than 4 in our case). The remaining sentences from the 26 threads were manually annotated by two independent annotators into three classes (Positive, Negative and Neutral/Unknown).

4 Subjectivity Lexicon

For our experiments, we used the Subjectivity Lexicon (SL) built by Wilson, Wiebe, and Hoffman (2005). The lexicon contains 8221 subjective expressions manually annotated as strongly or weakly subjective, and as positive, negative, neutral, or both. We have chosen this lexicon over other large automatically-generated dictionaries like SentiWordNet (Baccianella, Esuli, and Sebastiani, 2010), as it has been manually annotated and provides rich information with the subjectivity strength and prior polarity for each word considering the context of the word in the form of part of speech information.

The quality of this Subjectivity Lexicon is higher than the quality of other large automatically generated dictionaries; for example, SentiWordNet includes more than 65,000 entries. Some papers (Taboada et al., 2011) have shown that larger dictionaries contain information which is not detailed and include more words which may lead to more noise.

[2] http://www.medhelp.org, http://www.alldeaf.com, http://www.hearingaidforums.com

Below is a sample entry from the lexicon:

type=weaksubjlen=1 word1=ability pos1=noun stemmed1=n priorpolarity=positive

This entry contains the term *ability*, which is a noun. Its length is 1 (single term); it is not stemmed; it is weakly subjective and positive.

	Positive	Negative	Neutral	Both	Total	Percent
Adjective	1171	1838	235	5	3249	39.52
Noun	677	1346	144	3	2170	26.40
Verb	380	869	68	8	1325	16.12
anypos	362	676	104	5	1147	13.95
Adverb	128	183	19	0	330	4.01
Total	2718	4912	570	21	8221	100
Percent	33.06	59.75	6.93	0.26	100	

Table 2. Distribution of prior polarities within Subjectivity Lexicon

The Subjectivity Lexicon contains only single term expressions. Table 2 shows that about 60% of the words are negative and 33% are positive. Also, this resource contains 40% adjectives, 26.4% nouns, 16.12% verb, 13.95% anypos (could be in any part of speech) and only 4% adverbs. Table 3 shows that about 67.74% of the words are strong subjective and the rest of 32.2% are weak subjective in nature.

	Strong Subj	Weak Subj	Total	Percent
Adjective	2006 (61.74%)	1243 (38.25%)	3249	39.52
Noun	1440 (25.85%)	730 (33.6%)	2170	26.40
Verb	861 (15.46%)	464 (35.01%)	1325	16.12
Anypos	1043 (18.72%)	104 (9.06%)	1147	13.95
Adverb	219 (3.93%)	111 (33.6%)	330	4.01
Total	5569	2652	8221	100
Percent	67.74	32.26	100	

Table 3. Distribution of subjectivity clue within the Subjectivity Lexicon

The lexicon contains only 21 words having polarity "both". Out of these 21, only 10 words were found unique with their part of speech. As these both polarity words are neutral in our case, we decided to merge them with the neutral words. Table 4 shows the relation between strong and weak subjectivity with the polarity lexicon.

	Strong Subj	Weak Subj	Total	Percent
Positive	1717 (30.8%)	1001 (37.74%)	2718	33.06
Negative	3621 (65%)	1291 (48.6%)	4912	59.75
Neutral	231 (4.14%)	360 (13.57%)	591	7.18
Total	5569	2652	8221	100
Percent	67.74	32.26	100	

Table 4. Distribution among subjectivity and polarity in the lexicon

5 Methodology

In this work, we have used several different features for the sentiment analysis of the sentences. Section 4.2 lists all these features. These features are computed and presented for each sentence in a data file format used by the WEKA tool (Hall et al., 2009). Classification is performed based on the computed features and accuracy is measured using different combinations of features in order to improve the classification performance.

5.1 Parts of Speech in Lexicon Matching

Words can have different polarity when they represent different parts of speech; e.g., novel is positive when it is in adjective form; however it is a neutral as a noun. To minimize this problem, we have matched the words in the lexicon with their part-of-speech information. That helped us to use the correct polarity and subjectivity indication considering the correct part of speech.

Nouns

In our lexicon, nouns have the second most coverage, with 26.4%.

Verbs

Verbs are the next common in the lexicon and give good indication of subjectivity. However, as verbs are used in many different forms and have many meanings, just relying on the verb polarity will misguide the prediction in cases where the verbs are used in some other senses, e.g., *he uses a car* is neutral, when *he was used* has a negative sense.

Lemmatization

For all nouns and verbs, we have used the lemmatization from the GATE [3] morphological plugin, which provides the root word. In case of nouns, the root word is the singular form of the plural noun, e.g., *bottles* become *bottle*, etc. In the case of verbs, the plugin provides the base form for infinitive, e.g., *helping* becomes *help*, and *watches* become *watch*. After performing lemmatization, we found 158 more words that were detected with the same part of speech considered as the original. There were still 175 words which were found with the root word in the lexicon, but with different part of speech, e.g., *senses* was used as noun in the data; after lemmatization it becomes *sense*, which exists as verb in the lexicon. Therefore it cannot be matched, as the context and meaning of the word is different.

Adjectives

Early research in sentiment analysis focused mainly on adjectives and phrases containing adjectives, e.g., *what a blessed relief*. Adjectives are good indicators for the positivity or negativity of the sentences, but they are not sufficient for identifying the subjectivity in the sentences, as we will see in the experiments.

Adverbs

Adverbs are words that modify the verbs, adjectives and other phrases or clauses, e.g., *I am usually a contributing adult, and am happily sane and I say whoa how did that happen?* Adverbs have the lowest concentration in the lexicon, only 4%, and as many adverbs are identified by their characteristic "ly" suffix, we have removed the suffix–ly and then matched the new word in the lexicon by considering it as adjective. In English, most of the adverbs with suffix –ly such as *badly*, *softly*, *carefully*, *extremely* are forms of adjectives; therefore considering these provides better results in predicting the polarity of words in their correct senses.

Features

All the features considered for the experiment are based on sentence level. Table 5 shows the final features selected for the experiments. The most common features were pronouns, followed by weak subjective clues, adjectives, and adverbs. There were more words that matched with the lexicon's positive words than those that matched with the lexicon's negative words. This led to classifier's performance become slightly better for positive in the experiments.

STRONGSUBJ	# of words found as strong subjective in current sentence
WEAKSUBJ	# of words found as weak subjective in current sentence
ADJECTIVE	# of adjectives
ADVERBS	# of adverbs
PRONOUN	# of pronouns
POSITIVE	# of words found having prior polarity as positive
NEGATIVE	# of words found having prior polarity as negative polarity as negative
NEUTRAL	# of words found having prior polarity as neutral
PRP_PHRASE	# of phrases containing pronouns found in current sentence

Table 5. Final features considered for the experiments

6 Sentiment Categories

The dataset of 3515 sentences from 26 threads were manually annotated by two annotators. The annotators were asked to tag each sentence into positive, negative and neutral (where both positive and negative sentiments are discussed). All the sentences which do not contain any opinions are left blank and they are removed, as we focus on sentences containing sentiments. According to Table 6, annotator1 and annotator2 did not label a large number of sentences, i.e., 2939 and 2728, respectively; therefore these sentences are removed. Due to the large number of unlabeled sentences, the data is reduced, as we consider only those sentences labeled as positive, negative and neutral. Since the positive and negative dataset is already balanced, no data balancing is performed.

Annotator 2	Annotator 1				
	Pos	Neg	Neut	No Label	Total
Pos	226				329
Neg		214			296
Neutral			117		162
No Label				2720	2728
Total	230	218	128	2939	3515

Table 6. Annotations statistics of Sentences between the two annotators

[3] http://gate.ac.uk/sale/tao/splitch21.html#x26-52600021.11

The overall agreement for the two datasets is computed through the commonly used kappa statistic to evaluate the agreement ratio between the two annotators, in the same form used in (Sokolova & Bobicev, 2011):

$$\text{kappa} = \frac{\frac{a+d}{N} - \frac{f1g1+f2g2}{N^2}}{1 - \frac{f1g1+f2g2}{N^2}}$$

The overall percentage agreement between the annotators for the positive/negative dataset was 82.01% and kappa was 0.65. This indicates a substantial agreement between the taggers.

Positive / Negative dataset									
	Naïve Bayes			**SVM**			**Logistics-R**		
	P	R	F-1	P	Re	F-1	P	R	F-1
positive, negative	0.656	0.65	**0.644**	0.661	0.641	**0.625**	0.649	0.645	**0.641**
all features	0.595	0.584	0.565	0.641	0.618	0.596	0.657	0.657	**0.656**
Baseline	0.540	0.541	0.539	0.586	0.586	0.586	0.585	0.584	**0.584**

Table 7. Comparison of performance between different features among three classifiers for both datasets

Positive / Negative dataset with lemmatization									
	Naïve Bayes			**SVM**			**Logistic-R**		
	P	R	F-1	P	Re	F-1	P	R	F-1
positive, negative	0.644	0.625	**0.607**	0.636	0.607	**0.578**	0.688	0.686	**0.685**
all features	0.589	0.580	0.560	0.627	0.600	0.570	0.671	0.670	0.670
Baseline	0.540	0.541	0.539	0.586	0.586	0.586	0.585	0.584	**0.584**

Table 8.Comparison of performance with lemmatization between different features among three classifiers for both datasets

7 Experiments

The output files generated by the system for the dataset are classified using the WEKA tool (Hall et al., 2009). For our evaluation, we used 10-fold cross validation which is a standard classifier selection for classification purpose. Experiments were performed using three different classifiers: Naïve Bayes, because it is known to work well with text, support vector machine (SVM) because of its high performance in many tasks, and logistic regression (logistic-R), in order to try one more classifier based on a different approach.

Performance was evaluated using the F1-measure between the three classifiers on the given datasets. We found that the performance of logistics regression was the best on the features selected for our evaluation.

For the baseline, the feature vector of bag of words is considered for both the datasets. We have not considered the unique words for the bag of words because eliminating the words that appeared only once halves the size of the vectors, thus it makes it easier for the classifier to handle bag-of-words; also the unique words do not contribute much in classification since they appear only once, in one class. Table 7 shows significant improvement with positive, and negative features over the baseline and the difference was much higher with Naïve Bayes and SVM than with logistic-R, i.e., 10.5%, 3.9% and 5.7%, respectively.

In Table 8, for the positive/negative dataset, the classifiers Naïve Bayes and SVM underperformed with the lemmatization and their best performance decreased by 3.7% and 4.7%, respectively. However, the performance of logistic-R increased significantly, by 4.4%, and its F1-measure reached 68.5%, which indicates the benefit of lemmatization in matching within the lexicon.

8 Analysis

The results from the experiments have provided several insights about the sentiment analysis in health-related forums. Note that the bag-of-word representation (BOW) is a high baseline that is hard to beat in many texts classification problems. The Subjectivity Lexicon clues for polarity such as positive, negative and neutral have shown significant improvement for the identification of positive and negative sentences. As a result, the performance has increased by 4.2% on average among the three classifiers.

We have noticed that for the semantic orientation of the sentences, the combination of lexicon clue features with other basic counting features such as the number of adjectives, the number of adverbs, etc., decreased the performance of classification, as all the three classifiers have performed best with only positive and negative features.

Our results are comparative to other related studies for sentiment classification of medical forums. Sokolova & Bobicev (2011) achieved the best F-score of 0.708 using SVM; similarly Goeuriot et al. (2012) for drug reviews achieved F-score of 0.62 for the positive class, 0.48 for the negative class and 0.09 for the neutral class.

In general, for consumer reviews, opinion-bearing text segments are classified into positive and negative categories with Precision 56%−72% (Hu & Liu 2004). For online debates, the complete texts (i.e., posts) were classified as positive or negative stance with F-score 39%−67% (Somasundaran & Wiebe, 2009); when those posts were enriched with preferences learned from the Web, the F-score increased to 53%−75%.

It is also noted that the classification for semantic orientation depends heavily on the quality of the lexicon used, rather than the size of the lexicon, as the results show that the classification of the sentences into positive and negative reached 70% by using only the polarity clues for individual words within the lexicon.

9 Conclusion and Future Work

In this work, we performed the sentiment analysis of sentiments expressed in online messages related to Hearing Loss.

We used several lexicon-based features together with rule-based features like pronoun phrases for our classification of the dataset for detecting semantic orientation within the subjective data using different features based on the subjective lexicon.

The dataset of 3515 sentences from 26 threads were manually annotated by two annotators and achieved 82.01% overall agreement with kappa 0.65. Evaluations have been made for the classification of the substantial agreed data using three different supervised learning-based classifiers and it is shown that our proposed features outperformed the baseline of bag-of-word features by 10.5% with Naïve Bayes, 3.9% with SVM and 5.7% with logistic-R.

In future work, we could consider several directions. The lexicon could be improved, as the domain lexicon created in (Goeuriot et al., 2012) has shown better results over other dictionaries for polarity detection of the sentences.

Also, techniques and features presented in (Taboada et al., 2011), (Kennedy and Inkpen, 2006) such as intensification (e.g., *very good*) increase the polarity of *good* and negation (e.g., *not good*) which reverses the polarity of *good*, can be used for the semantic orientation or polarity detection of the sentences.

Another direction for future work could be to investigate changes of sentiments in threads. We want to analyze what linguistic events may prompt polarity to reverse (e.g., from *positive* to *negative)* and under what conditions the same polarity is sustained. To the best of our knowledge, this task was not addressed before.

Acknowledgments

This work in part has been funded by a Natural Sciences and Engineering Research Council of Canada Discovery Research Grant and by a Children's Hospital of Eastern Ontario Department of Surgery Research Grant.

References

Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). *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 (LREC'10), Valletta, Malta, May.

Bobicev, V., Sokolova, M., Jafer, Y., & Schramm, D. (2012).*Learning sentiments from tweets with personal health information.*In Advances in Artificial Intelligence (pp. 37-48).Springer Berlin Heidelberg.

Eysenbach, G. (2009). *Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet.* Journal of medical Internet research, 11(1).

Gillick, D. (2009, May). *Sentence boundary detection and the problem with the US.* In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers (pp. 241-244). Association for Computational Linguistics.

Goeuriot, L., Na, J. C., Min Kyaing, W. Y., Khoo, C., Chang, Y. K., Theng, Y. L., & Kim, J. J. (2012, January). *Sentiment lexicons for health-related opinion mining.*In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (pp. 219-226).ACM.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009).*The WEKA data mining software: an update.* ACM SIGKDD Explorations Newsletter, 11(1), 10-18.

Hu, M., & Liu, B. (2004, August).*Mining and summarizing customer reviews.*In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177).ACM.

Kennedy, A., & Inkpen, D. (2006).*Sentiment classification of movie reviews using contextual valence shifters.* Computational Intelligence, 22(2), 110-125.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003).*Lying words: Predicting deception from linguistic styles.* Personality and Social Psychology Bulletin, 29(5), 665-675.

Rhodewalt, F., & Zone, J. B. (1989).*Appraisal of life change, depression, and illness in hardy and nonhardy women.* Journal of Personality and Social Psychology, 56(1), 81.

Sokolova, M., &Bobicev, V. (2011).*Sentiments and Opinions in Health-related Web messages.*In Recent Advances in Natural Language Processing (pp. 132-139).

Somasundaran, S., &Wiebe, J. (2009, August).*Recognizing stances in online debates.* In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1 (pp. 226-234). Association for Computational Linguistics.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., &Stede, M. (2011).*Lexicon-based methods for sentiment analysis.* Computational linguistics, 37(2), 267-307.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October).*Recognizing contextual polarity in phrase-level sentiment analysis.*In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (pp. 347-354).Association for Computational Linguistics.

Yi, J., Nasukawa, T., Bunescu, R., &Niblack, W. (2003, November). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Data Mining, 2003.*ICDM 2003. Third IEEE International Conference on (pp. 427-434).