Sentiments and Opinions in Health-related Web Messages

Marina Sokolova
Faculty of Medicine, University of Ottawa
and
Electronic Health Information Lab,
CHEO Research Institute
sokolova@uottawa.ca

Victoria Bobicev
Department of Applied Informatics,
Faculty of Computers, Informatics
and Microelectronics,
Technical University of Moldova
vika@rol.md

Abstract

In this work, we analyze sentiments and opinions expressed in user-written Web messages. The messages discuss health-related topics: medications, treatment, illness and cure, etc. Recognition of sentiments and opinions is a challenging task for humans as well as an automated text analysis. In this work, we apply both the approaches. The paper presents the annotation model, discusses characteristics of subjectivity annotations in health-related messages, and reports the results of the annotation agreement. For external evaluation of the labeling results, we apply Machine Learning methods on the annotated data and present the obtained results.

1 Motivation

In recent years, Text Data Mining (TDM) and Natural Language Processing (NLP) intensively studied sentiments and opinions in user-written Web texts (e.g., tweets, blogs, messages). Researchers analyzed sentiments and opinions that appear in consumer-written product reviews, financial blogs, political discussions (Blitzer et al, 2007; Ferguson et al, 2009; Kim and Hovy, 2007). Health care and medical delivery service is another area where practitioners become interested in what users write in their Web posts. Importance of knowing user opinions had became evident during H1N1 pandemic, the first pandemic when Web discussions influenced the general public (Eysenbach, 2009); Figure 1 presents an example.[1]

The shift from contrived medical text to less rigorously written and edited user-written texts is a challenge for TDM and NLP methods. The current techniques were primarily designed to analyze medical publications in traditional media (e.g., journal articles) and organizational documents (e.g., hospital records) or task-dependent (e.g., information retrieval related to insurance claims)(Angelova, 2009; Cohen et al, 2010; Konovalov et al, 2010).

> Posted by Kristi: I really dont know why everyones freaking out about the H1N1 vaccine. I got it the first day it came out (about a week and a half ago) and so did 4 of my family members. None of us had any problems and were all really glad we got the vaccine.

Figure 1: A user post about H1N1 vaccination.

The goal of this work is to study sentiments and opinions in health-related Web messages. We start with building a data set of annotated sentences. We present an opinion and sentiment annotation scheme and its application to tag sentences harvested from the Web messages. We report evaluation of manual annotation agreement. Finally, machine learning methods are applied to automatically assess the sentence labeling.

2 Opinions and Sentiments

We are interested in the expressions of user *private state* which is not open to objective observation or verification (Quirk et al, 1985). These personal views are revealed through thoughts, perceptions and other subjective expressions that can be found in text (Wiebe, 1994).

We assume that the private states can be revealed by emotional statements, *sentiments*, and subjective statements that may not imply emotions, *opinions*. In this work, statements are considered within the sentence bounds; thus, sentences are the units of our language analysis. We agree with Lasersohn (2005) and Kim and Hovy (2007) that opinion can be expressed about a fact of matter, and should not be treated as identical to sentimental expression.

[1] http://www.gocoldflu.info/archives/, accessed April 25, 2011

Proceedings of Recent Advances in Natural Language Processing, pages 132–139,
Hissar, Bulgaria, 12-14 September 2011.

We further sub-categorize sentiments into *positive* and *negative*, and opinions – into *positive*, *negative* and *neutral*. Sentences that do not bear opinions or sentiments are considered objective by default and are left for future studies.

3 Opinion and Sentiment Annotation

3.1 Annotation Model

Annotation of subjectivity can be centered either around perception of a reader/annotator (Strapparava and Mihalceal, 2008) or the author of a text (Balahur and Steinberger, 2009). Our model is author-centric. Our guidelines for annotators defined that a subjective statement contains information which has not been taken by the author from some external source but rather his/her personal thoughts (as defined in Section 2). We requested that annotators do not impose their own sentiments and attitudes towards information in the text (Balahur and Steinberger, 2009). Instead we suggested that an annotator imagined sentiments and attitudes that the author possibly had while writing.

Separation of good and bad news from the author attitude is important in the health-related analysis. We know that subjective expressions are highly reflective of the text content and context (Chen, 2008). Health-related messages are often written about illnesses and medical treatment. Users write about diseases, symptoms, sick relative and friends. This information is naturally distressing and may cause a negative attitude in annotators. We asked annotators not to mark descriptions of symptoms and diseases as subjective; only author's opinion or sentiment should be annotated. For example, "For a very long time I've had a problem with feeling really awful when I try to get up in the morning" is a description of some symptoms and should not be annotated as subjective. In contrast, "I don't know if that makes sense, it seems to me that the new drug which stimulates red blood cell production would be a more logical approach, erythropoiten (sp?)" exposes the author's thoughts and ideas. It should be annotated as an opinion though without an emotional attitude. Another example, "Alas, I didn't record the program, but wish I had" expresses the author's regret and should be annotated as a negative opinion about the action (i.e., not recording the program).

We considered essential to advise annotators not to agonize over the annotation and, if doubtful, leave the example un-annotated (Balahur and Steinberger, 2009). The rule is especially important for annotation of user-written texts, when annotators can be destructed and even annoyed by misspellings, simplified grammar and informal style and unfamiliar terminology specific to an individual user..

3.2 Schema

Our annotation schema is based on the following assumptions:

(a) annotation was performed on a sentence level; one sentence expressed only one assertion; this assumption held in a majority of cases;

(b) only author's subjective comments were marked as such; if the author conveyed opinions or sentiments of others, we did not mark it as subjective as the author was not the holder of these opinions or sentiments;

(c) we did not differentiate between the objects of comments; author's attitude towards a situation, an event, a person or an object were considered equally important.

Annotators were informed that the annotation was sentence-level and examples of annotated texts presented them were also with annotated sentences. Thus they tended to annotate sentences. If consecutive sentences were subjective, every one was marked. In some cases, only a subjective part of a sentence was tagged, whereas the other part, containing factual information was not included in the sentiment tag.

3.3 Mode

User-written messages usually have opening, body, and closure. Opening can be email subject, parameters of the message, body presents the main content, and closure can be signature or a link to a personal web site.

We used the markup tags HEADER, FOOTER and BODY (Figure 2). HEADER referred to the parameters of the message, FOOTER marked the closing part which started with the signature; this part was marked FOOTER regardless of its length and omitted from the processing. BODY marked the message between HEADER and FOOTER.

To comply with our annotation schema, we divide BODY into CITATION and TEXT. CITATION marked embedding of the previous messages in

the current one, TEXT marked the text of the message written by the author. In the current study, we are interested in the TEXT part; other parts are left for future work. TEXT was divided in sentences and further analyzed for opinions and sentiments.

HEADER:
Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.edu!ogicse!emory!gatech!pitt.edu!pitt!geb
From: geb@cs.pitt.edu (Gordon Banks)
Newsgroups: sci.med
Subject: Re: vangus nerve (vagus nerve)
Message-ID: <19397@pitt.UUCP
Date: 5 Apr 93 14:27:13 GMT
Article-I.D.: pitt.19397
References: <52223@seismo.CSS.GOV
Sender: news@cs.pitt.edu
Reply-To: geb@cs.pitt.edu (Gordon Banks)
Organization: Univ. of Pittsburgh Computer Science
Lines: 16
BODY:
CITATION:
In article <52223@seismo.CSS.GOV bwb@seismo.CSS.GOV (Brian W. Barker) writes:
> mostly right. Is there a connection between vomiting
> and fainting that has something to do with the vagus nerve?
TEXT:
Stimulation of the vagus nerve slows the heart and drops the blood pressure.
FOOTER:

Gordon Banks N3JXP | "Skepticism is the chastity of
geb@cadre.dsl.pitt.edu | the intellect, and it is shameful
to surrender it too soon."

Figure 2: Example of a message.

4 Empirical Application

4.1 Data

For our empirical part, we used the sci.med texts of *20 Newsgroups* [2]. It is a benchmark data set of 20,000 messages, popular in applications of machine learning techniques, such as text classification and text clustering. There are 1000 sci.med messages. Most sci.med messages were posted by people who wanted to know something about an illness, drugs or treatment (e.g., questions on tuberculosis, haldol prescription to elderly). After the question appeared on the message board, other people could reply and add comments (Figure 2).

To group messages by their content, we merged the messages with the same topic. A script automatically placed all messages with the same Subject line in the file with the same title. Thus, we obtained 365 files named "Arrhythmia", "arthritis and diabetes", "Athletes Heart", etc. Essentially, a file stored the whole discussion thread on the title topic. Many files contained only one question and one or two answers. Several topics raised interest of many list members. Such files contained rather hot discussions (e.g., "Candidayeast Bloom"," MSG sensitivity", "Homeopathy"). In contrast, some files contained newsletters, conference announcements, other announcements that were considered objective (Section 2); these files were deleted from annotation. Finally, 357 files were left for the annotation.

4.2 Annotation results

10 undergraduate and 10 master students were involved in the process. A master student had 30 files to annotate. The results of the annotation were examined; students with better annotations received more files. An undergraduate student had 10 files to annotate; only students with the satisfactory quality annotations were given more files. Finally, the 357 files have been annotated by at least one annotator.

216 have been tagged by two annotators, and 21 have been tagged by three annotators. 120 files have been tagged by only one annotator. A majority of these files did not contain subjective information, e.g., a question and a factual answer . We have divided the final tags into 3 categories:[3]:

subjective sentences : both annotators identified them as subjective, sentiment or opinion, and marked either the same polarity or neutral;

weak subjective sentences : only one annotator identified them as subjective;

non-subjective and uncertain sentences : sentences that the annotators did not mark as subjective or marked with the opposite polarity.

[2] http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html

[3] The labelled sentences are posted on www.ehealthinformation.ca/ap0/opendata.asp

Subjective sentences		
1st annotator	2nd annotator	#
negative sentiment	negative sentiment	92
neutral opinion	neutral opinion	85
positive opinion	neutral opinion	57
negative opinion	neutral opinion	53
negative sentiment	negative opinion	48
negative opinion	negative opinion	43
positive sentiment	positive sentiment	41
negative sentiment	neutral opinion	41
positive opinion	positive opinion	27
positive sentiment	positive opinion	21
positive sentiment	neutral opinion	20
Weak subjective sentences		
1st annotator	2nd annotator	#
no annotation	neutral opinion	655
no annotation	negative sentiment	331
positive opinion	no annotation	212
negative opinion	no annotation	201
positive sentiment	no annotation	172
no annotation	unspecif. sentiment	12
Non-subjective and uncertain subjectivity		
1st annotator	2nd annotator	#
no annotation	no annotation	4190
positive sentiment	negative opinion	34
negative sentiment	positive sentiment	28
positive opinion	negative sentiment	9
positive opinion	negative opinion	9

Table 1: Annotation results for sentiment and opinion sentences in the sci.med texts.

Table 1 lists the results for the three sentence groups.

4.3 Discussion

6408 sentences were annotated in total. The majority – 4190 sentences – were considered non-subjective by both annotators. *Neutral opinion* was the most frequent subjective label, some persons asked questions and some replied in many cases expressing their own opinions. 85 sentences were marked *neutral opinion* by both annotators. In 655 cases, it was a weak subjectivity (i.e., identified by one annotator). The latter set contained ambiguous sentences, without clear indicators was the expressed statement author's thought or just information taken from some sources. We report some examples: "Symptoms can be drastically enhanced by food but not inflammation", "The low residue diet is appropriate for you if you still have obstructions", "Then they may be able to crowd out garbage genes"

Negative sentiment was another large set of the ambiguous annotation. In Section 3.1, we wrote that the texts were about diseases, so it was natural that sometimes annotators marked descriptions of symptoms or sickness as *negative sentiment*. Often *negative sentiment* was attributed to sentences that were interpreted as subjective only in the message context. For example, "I said that I PERSONALLY had other people order the EXACT SAME FOOD at TWO DIFFERENT TIMES from the SAME RESTAURANT" was marked *negative sentiment* in context of a very opinionated discussion. For the annotator, it was clear that the author of the text had been really angry, and the sentence did carry negative emotion even if it did not contain indicative words.

We have found that sarcasm was a strong factor for the polarity disagreement between annotators. "I'm forever in your debt" was marked as *positive sentiment* and *negative sentiment*, because it was positive as is but was used in a sarcastic answer to another message; one annotator took the whole context in consideration but another one did not. "Surprise surprise different people react differently to different things." and "Subject: Scientific Yawn" (denouncing an alternative medicine) are two other illustrations of opposite polarity labeling. Perhaps, a more complex set of sentiment annotation tags can help to capture such sentiments.

Content-wise, we found that several types of sentences created problems while annotation: advices, suggestions ("go and see a doctor"); courtesy ("thank you in advance", "I would greatly appreciate any reply", "good luck"); questions and indirect questions ("can somebody point me", "I am interested in", "I would like to find any information"). An appropriate remedy can be to divide subjective sentences into categories, e.g., reporting, advice, judgment and sentiment (Asher et al, 2009). Rhetorical relations formed another influential factor. However, correct identification of this phenomena requires a higher proficiency of annotations.

Additionally, annotators faced challenges intrinsic to the user-written text (Section 3.1). Indeed, syntactic rules were not strictly respected and there were mistypes and misspellings. Other challenges were recognitions of trade-mark and proprietary names ("itraconazole", "Oodles of Noodles"), public health and related services ("AMA", "FDA", "State Licensing Board", "ABFP") and medi-

Table 2: Concordance matrix.

	1st observer		
2nd observer	YES	NO	Totals
YES	a	b	g_1
NO	c	d	g_2
Totals	f_1	f_2	N

cal and scientific terms ("Candida", "sinusitis", "yeast bloom").

5 Empirical Evaluation

5.1 Concordance evaluation

To assess the quality of subjective labeling, we computed two types of measures. First, we separately assessed agreement between the annotator labeling of positive and negative sentiments and opinions. We opted for two, positive and negative, measures because annotators may agree on *what constitutes* a subjective label and disagree on *what does not*, e.g., their understanding of *positive* may be close and their understanding of *not positive* may be far apart. We find the two-dimensional values being more informative than the one-dimensional value (Bhowmick et al, 2008; Murakami et al, 2010).

We applied two measures introduced in (Cicchetti and Feinstein, 1990a):

$$p_{pos} = 2a/(f_1 + g_1) \quad (1)$$

$$p_{neg} = 2d/(N - (a - d)) \quad (2)$$

Next, we computed a commonly used *kappa* to evaluate a ratio between the chance-corrected observed agreement and the chance-corrected perfect agreement (Cicchetti and Feinstein, 1990a):

$$kappa = \frac{\frac{a+d}{N} - \frac{f_1 g_1 + f_2 g_2}{N^2}}{1 - \frac{f_1 g_1 + f_2 g_2}{N^2}} \quad (3)$$

Notations are presented in Table 2.

We report the assessment results in Table 3.

The reported results show that annotators find a common ground on sentences that *do not* belong to the categories. This mutual understanding holds across all the subjective categories. We interpret this as a possibility of correct identification of negative examples for all the categories. Annotators also agree on what belongs to positive and negative sentiments; for these two categories, we expect correct identification of positive and negative examples.

Annotation	p_{pos}	p_{neg}	$kappa$
Pos Sentiment	0.667	0.956	0.621
Neg Sentiment	0.674	0.886	0.562
$Average$	0.671	0.921	0.592
Assessment	p_{pos}	p_{neg}	$kappa$
Pos Opinion	0.409	0.892	0.350
Neg Opinion	0.460	0.884	0.365
Neut Opinion	0.497	0.761	0.280
$Average$	0.455	0.846	0.332

Table 3: Concordance assessment.

5.2 Statistical language analysis

To analyze the lexical indicators of subjectivity, we built N-gram models ($N = 1, 2, 3, 4$). The N-gram models estimate the probability of a word sequence $w_1 \dots w_n$ as a conditional probability of the word w_n appearing after the sequence of words $w_1 \dots w_{n-1}$:

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-N+1}^{n-1}) \quad (4)$$

The models were built for subjective sentences and weak subjective sentences (upper parts of Table 1). We analyzed most frequent words (occurrence $\geq$ 3) and word combinations output by the models. To make the task feasible, we deleted stop words (i.e., pronouns, prepositions, articles, determiners and auxiliary verbs).

Uni- and bi-gram outputs had shown that very few emotionally charged words appear among the most frequent words. Examples of such words are "good", "happy", "hard", "unfortunately"; "good", "happy", however, may indicate courtesy expressions more than sentiments. For instance, their most frequent bi-grams are "very good", "am happy". Tri- and quadri-gram outputs were very sparse (i.e., occurrences < 5), thus, not reliable for semantic generalization. Important to note that words listed in SentiWordNet (Denecke, 2008) and WordNet-Affect (Strapparava and Mihalceal, 2008) as a rule do not appear in our data .

We computed a significant relative frequency difference (Rayson and Garside, 2000) to find words and word combinations ($N = 2, 3, 4$) on which two sets of sentences differ. The difference was computed as follows:

$$LL(w) = 2(a \log \frac{a(a+b)}{c} + b \log \frac{b(a+b)}{d}) \quad (5)$$

where w – the word, a and b are the occurrences of w in sets A and B respectively, c and d – sizes of

A and B in words. We chose LL because the measure allows two-tailed comparison of w's position in sets A and B.

This method, too, output a few emotionally charged words: "trouble", "hard", "problem", "expensive" are content words that differentiate between positive and negative opinions; "bad", "problem", "hard", "better" appear among words that differentiate between positive and negative sentiments. Word combinations on which the sets differ do not contain emotionally charged words.

5.3 Machine Learning Experiments

Sentiment and opinion classification results are highly susceptible to the classification task, the data characteristics and selected text features. Consequently, the data characteristics affect the classification accuracy. We wished to assess how well algorithms discriminate between

(a) positive and negative sentiment sentences,

(b) positive and negative opinion sentences.

Our hypothesis was that if algorithms achieved a competitive accuracy of learning then it confirmed a good quality of labels.

5.4 Data

We used the labeled sentences without any additional pre-processing. As a result, two sentence sets have been built:

Sentiments 62 positive and 179 negative sentences;

Opinions 169 positive and 74 negative sentences.

We represented each set through all the words that appear in the set more than twice. Two types of attributes were used in experiments: bag of all the words (binary representation) and occurrences of all the words (numeric representation). The two representations provided similar results. We further report the numeric representation results, which were slightly better than binary.

5.5 Learning Results

We applied *Naive Bayes (NB)*, *Decision Trees (DT)*, *K-Nearest Neighbor (KNN)* and *Support Vector Machines (SVM)*. $Fscore$, $Precision(Pr)$, $Recall(R)$ and $BalancedAccuracy(ROC)$ were used to evaluate the performance.

Sentiments				
Algorithm	Pr	R	Fscore	ROC
NB	0.679	0.726	0.686	**0.611**
K-NN	0.649	0.705	0.664	0.578
SVM	**0.714**	**0.751**	**0.708**	0.574
DT	0.552	0.743	0.633	0.485
Baseline	0.552	0.743	0.633	0.485
Opinions				
Algorithm	Pr	R	Fscore	ROC
NB	0.791	0.790	0.767	**0.805**
K-NN	0.744	0.753	0.720	0.586
SVM	**0.850**	**0.848**	**0.839**	0.777
DT	0.734	0.741	0.737	0.682
Baseline	0.484	0.695	0.571	0.481

Table 4: Classification results for positive and negative sentence classification. The values are averaged for positive and negative classes. Best values are in **bold**. Baseline is calculated if all the sentences are into the majority class.

Table 4 reports the best results. For positive and negative sentiments, the reported results were obtained with the following parameters: *DT* – learning coefficient $\alpha = 0.15$, *NB* used kernel estimates; *K-NN* – 9 neighbors, Euclidean distance; *SVM* – complexity parameter C = 0.65, kernel polynomial = 0.52. For positive and negative opinions, the reported results were obtained with the following parameters: *DT* – learning coefficient $\alpha = 0.40$; *NB* – with kernel estimates; *K-NN* – 1 neighbor, Euclidean distance; *SVM* – complexity parameter C = 2.75, kernel polynomial K= 1.0.

Our results are competitive with previously obtained results. As reported in (Sokolova and Lapalme, 2011), opinion-bearing sentences are classified against facts with $Precision$ 80% – 90% (Yu and Hatzivassiloglou, 2003); for consumer reviews, opinion-bearing text segments are classified into positive and negative categories with $Precision$ 56% – 72%; for online debates, posts were classified as positive or negative with $F-score$ 39% –67%, $F-score$ increased to 53% – 75% when the posts were enriched with the Web information, . 90% $BalancedAccuracy(ROC)$ was obtained in opinion spam reviews versus genuine reviews classification. For positive and negative review classification, $Accuracy$ is 75.0% –81.8% when data sets are represented through all the uni- and bigrams.

6 Text Mining and Corpora Annotation in the Domain

Opinion mining and sentiment analysis have become a major research topic for Computational Linguistics. A high demand for knowledge sources prompted development of semantic resources SentiWordNet (Denecke, 2008), WordNet-Affect (Strapparava and Mihalceal, 2008), MicroWNOp (Balahur et al, 2010), as well as lists of affective words or collocations created ad-hoc (Whitelaw et al, 2005; Yu and Hatzivassiloglou, 2003) and even non-affective words (Sokolova and Lapalme, 2011). Sometimes positive and negative text rating was available and used in machine-learning experiments (Pang et al, 2002). At the same time, there are no available sources for sentiment and opinion analysis of user-written health discussions. We work to build such a source.

Sentiment and opinion analysis intensively studied consumer-written product reviews (Blitzer et al, 2007). Somewhat lesser attention was given to political discussion boards (Kim and Hovy, 2007). In (Ferguson et al, 2009), financial blogs were annotated on the document and paragraphs level with their sentiment towards the same topic using a five-point scale *Very Negative, Negative, Neutral, Positive, Very Positive*, in addition to the labels *mixed*, which indicates a mixture of positive and negative sentiment, and *not relevant*. It seemed intuitive that paragraph -level annotation should be useful in providing more accurate information which can be leveraged by a machine learning module. However, the results did not show any improvement. To the best of our knowledge there was only one corpus of blogs with fine-grained annotation of subjectivity (Boldrini et al, 2009). A multilingual corpus of blog posts on different topics of interest in three languages - Spanish, Italian and English was annotated using a fine-grained annotation schema in order to capture the different subjectivity/ objectivity, emotion/opinion/ attitude aspects.

Unlike the listed above work, we concentrate on discussions of health-related topics. There are few dedicated work on polarity of health and medical text. In (Niu et al., 2005; Niu et al., 2006), the authors analyzed textual expressions corresponding to *positive, negative, neural* clinical outcomes. In our work, however, clinical outcomes are set apart from user sentiments and opinions.

So far, experiments in corpora annotation attracted considerably less attention. In (Wiebe et al, 2005), the authors annotated articles at the word- and phrase-level by using fine-grained annotation scheme. Another experiment on news annotation was carried on for the SemEval 2007 Affective Text Task (Strapparava and Mihalceal, 2008). The subjectivity annotation of newspaper articles was discussed in (Balahur and Steinberger, 2009) and (Bhowmick et al, 2008). In the former, the researchers extracted 1592 quotes (reported speech) from newspaper articles and annotated for the sentiment on the target of the quotes. The annotation guidelines allowed increase of the inter-annotator agreement from $< 50\%$ up to 60%. In the latter, the authors collected 1000 affective sentences and categorized them into *direct* and *indirect* affect categories. Our work, instead, is focused on positive and negative sentiments and opinions in user-written Web messages.

7 Conclusion and Future Work

In this paper, we have presented a study of sentiments and opinions in user-written Web messages. We focused on messages posted on health discussion boards. In those messages, users discussed health and ailment, treatments and drugs, asked questions about possible cures. Without having precedents of subjectivity analysis in health discussions, we have designed an author-centric annotation model. The model shows how positive and negative sentiments and positive, negative and neutral opinions can be identified in health discussions.

We applied the annotation model to the sci.med messages of *20 NewsGroups*. We have evaluated concordance of the manual annotation by computing three measures : p_{pos}, p_{neg} and $kappa$. The results show that annotators better identify sentiments than opinions and stronger agree on what type of sentences do *not* belong to positive or negative subjective categories. Our Machine Learning results are comparable with previous results in the subjectivity domain.

Our future plans are to continue the annotation; the final aim is to have all texts annotated by at least five persons. We also plan to study objective, factual statements expressed by users in their messages.

Acknowledgements

The first author's work is in part funded by a Discovery grant of Natural Sciences and Engineering Research Council of Canada. The second author thanks the conference organizers for the RANLP grant.

References

Angelova, G. Ontological Approach to Terminology Learning. *Comptes rendus de l'Academie bulgare des Sciences*, **62**(10), pp. 1319–1326, 2009.

Asher N., Benamara F., Y. Y. Mathieu. Appraisal of opinion expressions in discourse, *Lingvisticae Investigationes*, **32**(2), 2009.

Balahur, A., R. Steinberger, M. Kabadjov, V. Zavarella, E. van der Goot, M. Halkia, B. Pouliquen, J. Belyaeva. Sentiment Analysis in the News, *Proceedings of the Seventh conference on International Language Resources and Evaluation*, 2010.

Balahur, A., R. Steinberger Rethinking Sentiment Analysis in the News: from Theory to Practice and back. *Proceedings of the 1st Workshop on Opionion Mining and Sentiment Analysis*, 2009

Bhowmick, P., P. Mitra, A. Basu. An Agreement Measure for Determining Inter-Annotator Reliability of Human Judgements on Affective Text, *Proceedings of Workshop on Human Judgements in Computational Linguistics*, COLING, p.p. 58–65, 2008.

Blitzer, J., M. Dredze, F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *Proceedings of ACL*, 440–447, 2007

Boldrini, E., A. Balahur, P. Martnez-Barco, A. Montoyo EmotiBlog: a finer-grained and more precise learning of subjectivity expression models. In *Proceedings of LAW IV, ACL*, 2009

Chen, W. Dimensions of Subjectivity in Natural Language (Short Paper). In *Proceedings of ACL-HLT*, 2008.

Cicchetti, D., A. Feinstein. High Agreement but Low Kappa: The Problems of Two Paradoxes, *Journal of Clinical Epidemiology*, **43**(6),p.p. 543–549 and p p. 551–558, 1990.

Cohen, K., C. Roeder, W. Baumgartner Jr., L. Hunter, and K. Verspoor Test suite design for biomedical ontology concept recognition systems. *Proceedings of LREC*, pp. 441–446, 2010.

Denecke, K. Using SentiWordNet for multilingual sentiment analysis, *Data Engineering Workshop, IEEE 24th International Conference*, 2008.

Eysenbach, G. Infodemiology and infoveillance. *Journal of Medical Internet Research*, **11**(1), 2009.

Ferguson, P., O'Hare, N., Davy, M., Bermingham, A., Tattersall, S., Sheridan, P., Gurrin, C., Smeaton, A. Exploring the use of Paragraph-level Annotations for Sentiment Analysis of Financial Blogs. *WOMAS 2009 - Workshop on Opinion Mining and Sentiment Analysis*, 2009.

Kim, S.-M., E. Hovy. Crystal: Analyzing predictive opinions on the web. *Proceedings of the 2007 EMNLP-CoNLL*, pages 1056–1064, 2007.

Konovalov S, M. Scotch, L. Post, C. Brandt. Biomedical Informatics Techniques for Processing and Analyzing Web Blogs of Military Service Members, *Journal of Medical Internet Research*, **12**(4), 2010.

Lasersohn, P. Context Dependence, disagreement, and predicates of personal taste *Linguistics and Philosophy*, **28**, pages 643–686, 2005

Murakami, K., E. Nichols, J. Mizuno, Y.Watanabe, H. Goto, M. Ohki, S. Matsuyoshi, K. Inui, Y. Matsumoto. Automatic Classification of Semantic Relations between Facts and Opinions, *Proceedings of NLP Challenges in the Information Explosion Era*, COLING, p.p. 21–31,2010,

Niu, Y., X. Zhu, J. Li, G. Hirst. Analysis of polarity information in medical text, in *Proceedings of the AMIA Annual Symposium*, 2005, 500–574.

Niu, Y., X. D. Zhu, G. Hirst. Using outcome polarity in sentence extraction for medical question-answering. In *Proceedings of the AMIA Annual Symposium*, 2006, 599–603.

Pang, B., L. Lee, S. Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques *Proceedings of EMNLP'-02*, pages 79-86, 2002.

Quirk, R., S. Greenbaum, G. Leech, J. Svartvik *A Comprehensive Grammar of the English Language* Longman, 1985.

Rayson, P., R. Garside. Comparing corpora using frequency profiling. *Proceedings of Comparing Corpora Workshop, ACL*, p.p. 1–6, 2000.

Sokolova, M., G. Lapalme. Learning opinions in user-generated Web content. *Journal of Natural Language Engineering*, to appear.

Strapparava, C., R. Mihalcea Learning to Identify Emotions in Text, *Proceedings of the 2008 ACM symposium on Applied Computing* 2008

Whitelaw, C., N. Garg, S. Argamon Using Appraisal Groups for Sentiment Analysis. *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 625 – 631, 2005.

Wiebe, J. Tracking point of view in narrative. *Computational Linguistics*, **20**, pp. 233–287, 1994

Wiebe, J., T. Wilson, C. Cardie Annotating Expressions of Opinions and Emotions in Language, *Language Resources and Evaluation*, **39** (2–3), pp. 165–210, 2005

Yu, H., V. Hatzivassiloglou Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings of EMNLP-03*, 2003.