Personal Health and Sentiments in Tweets

Victoria Bobicev[1], Yasser Jafer[2], and Marina Sokolova[3,4]

[1] Department of Applied Informatics, Technical University of Moldova
[2] School of Electrical Engineering and Computer Science, University of Ottawa
[3] Faculty of Medicine, University of Ottawa,
[4] Electronic Health Information Lab, CHEO Research Institute
[5] vika@rol.md, {yjafe089,sokolova}@uottawa.ca

Abstract. This paper presents results of sentiment analysis in Twitter messages that disclose personal health information. The messages discuss ailment, treatment, medications, 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, process of sentiment recognition 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 multi-class and binary classification of sentiments. Our reported Machine Learning results are comparable with previous results in the subjectivity analysis of user-written Web content.

1 Personal Health on the Web

Introduction of Web 2.0 technologies allowed the general public to actively participate in web content creation. In 2009, 44.6% of broadband users had posted some content on the Internet, with projected increase to 60% by 2013.[6] Blogosphere, social networks, message boards are awash with contributors' personal news that, in most cases, can be read without limitation by a global community. It had been shown that over 80% of Internet users are recipients of other users' content [6]. Those readers are influenced by emotional appeal of the content, as emotion-rich text affects a public mood stronger than rational arguments [1].

Twitter, the world's tenth most popular Web site, is a micro-blogging service with instant message postings.[7] It has > 200 $mln.$ users.[8] A user can post publicly visible messages ≤ 140 characters, often with shortenings: On my way c [*see*] vicki Shes recovering frm [*from*] surgery. Other users can subscribe to these tweets and respond with their messages.

Analysis of Twitter messages (i.e., tweets) presents a rapid means of estimating public mood on various subjects [16, 14]. Importance of user sentiments regarding health had become evident during H1N1 pandemic, the first pandemic when tweets' content has influenced behavior of significant group of population

[6] http://www.iab.net/insights research/947883/1675/669304
[7] http://twitter.com/
[8] alexa.com/topsites

[9, 12]. However, the reported work did not involve large-scale text analysis nor applied sentiment and text mining methods.

Sentiment analysis has become a major research topic for Computational Linguistics. There were several studies of emotions expressed by tweets. In [7], the authors explored happiness as a function of time, space and demographics using Twitter as a data source. A study of monthly English Twitter posts is reported in [23]. It investigates whether popular events are typically associated with increases in sentiment strength. In [11], the authors compared sentiment classification in microblogs containing branding comments, sentiments and opinions with manual coding. In [13], the authors study happy and sad moods on a corpus of annotated blogposts from the LiveJournal community.

At the same time, corpora annotation studies did not attract as mush attention. Topic-specific opinions in blogs were evaluated in [17]. Agreement among seven raters was computed for five classification categories, including positive, negative, mixed opinions and non-opinionated and non-relevant categories. Several publications were focused on subjectivity annotation of traditional media [2, 3, 21, 22].

There are few publications that consider sentiments in relation with personal health information posted on the Web. In [20], the authors analyzed opinions and sentiments expressed in the sci.med messages of *20 NewsGroups*. They evaluated concordance of the manual annotation by computing three measures: p_{pos}, p_{neg} and *kappa*. The results show that annotators stronger agree on what type of sentences do *not* belong to positive or negative subjective categories. For sentiments, an inter-rater agreement reached p_{pos}= 0.667, p_{neg} = 0.956, κ = 0.621, the best *Fscore*= 70.8 % was obtained by Support Vector Machines. 16 categories of opinions and emotions in tweets were presented in [5]. The extraction method traced tweets that contained H1N1 and its synonyms (e.g., swine flu). Neither methodological background nor numerical evaluation of the method were reported by the authors.

Our current work applies sentiment analysis methods to study sentiments in tweets related to personal health. The subjectivity and sentiment analysis is an interdisciplinary problem that is challenging not only for automated methods but also for humans. At the same time, Machine Learning (ML) methods have shown to be suitable for sentiment research of their ability to learn new information from empirical data. Our paper reports results obtained in solving both tasks: (a) manual sentiment analysis performed by multiple annotators on health-related tweets; (b) machine learning multi-class and binary classification of sentiments on health-related tweets.

2 Tweets with Personal Health Information

We had an access to 30, 164 Twitter threads (i.e., consequent tweets posted by a user). [9] An average length of a thread is 560 words, albeit some words can be

[9] http://caw2.barcelonamedia.org/node/7

Table 1. Tweets extracted from 200 x 5 random threads.

annotation	preceding tweets		PHI tweets		next tweets		total	
	#	words	#	words	#	words	#	words
fold 1	60	873	61	1,042	58	910	179	2,825
fold 2	54	770	54	828	53	783	161	2,381
fold 3	48	761	49	844	47	724	144	2,329
fold 4	46	605	47	709	46	543	139	1,857
fold 5	49	647	49	757	46	677	144	2,081
total	257	3,656	260	4,180	250	3,637	767	11,473

very short (e.g., "u","4"). The data set had only conversational tweets; spam, ads, organizational and promotional tweets were cleaned up. We collected 1000 random threads, by doing five rounds of random selection, 200 threads per round. We examined individual tweets within a thread and extracted those tweets which referred to personal health.

Our procedure used two lexical resources. First, we used ontology of personal health terms which lists terms related to body organs, symptoms, treatment, medical professional designations, etc. [19]. Semantic information from WordNet [10] helped us to identify terms that hold only health-related meaning (radiology, hernia, dermatologist) and more ambiguous terms (cavity, back, heart). Second, we used ontology of personal references. We have observed that in personal health related discussions, a person usually talks about his/her personal health and personal health of relations, relatives and non-relatives alike. The personal references, then, included personal pronouns (I, he, her), nouns representing relations (son, daughter, parents), and most frequent verbs of belonging (has, have, was).

In the current study, we used only unambiguous health terms to find health-related tweets. If an unambiguous term was not found in a tweet, the tweet was discharged, and the next tweet within a thread was processed. If at least one unambiguous term was found within a tweet, we marked the tweet as a potential PHI. We, then, searched through the tweet for a personal term. If a personal term was found, we put the tweet and its preceding and next tweets into the list A, otherwise we put the tweet into the list B. Such separation made easier the next step of manual procession of the tweets. Although the list B had a high redundancy, it did contain messages with personal health information, for example, Headache is not going away. Such messages are often more informal, then those which contain personal terms. The number of extracted health-related tweets was consistent for all the five folders of data. For each such tweet, we then extracted the preceding tweet and the next after it tweet. Table 1 presents the resulting data sets.

It should be emphasized that, the presence of one or more health ontology term(s) does not necessarily guarantee that the message refers to personal health. In well Im keeping my eye on you just so you know, eye indicates "anatomical body part" but the message does not refer to personal health. Therefore, manual

[10] http://wordnet.princeton.edu/

screening of the extracted messages was a complementary and necessary step in order to remove un-relevant messages and keep the personal health related tweets for future analysis.

3 Sentiment Annotation

Model Annotation of subjectivity can be centered either on perception of a reader [21] or the author of a text [2]. Our annotation model was author-centric and followed the model we used for sentiment annotation of health-related web messages [20]. We suggested that an annotator imagined sentiments and attitudes that the author possibly had while writing.

Separation of good and bad news from sentiments is challenging in the health-related text analysis. As subjective expressions are highly reflective of the text content and context [4], health-related messages can be distressing when written about illnesses, sick relatives and friends. Hence, we asked annotators not to mark descriptions of symptoms and diseases as subjective; only author's sentiments should be annotated. For example, I am hot I am sweating It is below freezing and I have to be going through menopause or somthing is a description of symptoms and should not be annotated as subjective. In contrast, it wasnt the stomach flu it was the nora virus yay me exposes the author's sentiment.

We considered essential to advice annotators not to agonize over the annotation and, if doubtful, leave the example un-annotated. The rule is especially important for annotation of tweets, when annotators can be destructed and even annoyed by misspellings, simplified grammar, informal style and unfamiliar terminology specific to an individual user. Another specific problem was the message shortness. For instance, the tweet What did you tell your parents The flu lol cause us to imagine different situations; the only indicator of sentiment is lol which allows us to interpret the whole tweet as humorous hence positive. In few cases, one tweet consisted of several sentences with different sentiments. Dentist tomorrow to fix the smile hopefully Ugh Anyway that was my night Hope urs was better LOL had three sentences Dentist tomorrow to fix the smile hopefully Ugh (negative), Anyway that was my night (neutral) and Hope urs was better LOL (positive). Such tweets were identified and excluded from further experiments.

Our annotation schema was implemented as follows:
(a) annotation was performed on a sentence level; one sentence expressed only one assertion; this assumption held in a majority of cases; annotators were informed that the annotation was sentence-level and examples of annotated texts presented them were also with annotated sentences;
(b) only author's subjective comments were marked as such; if the author conveyed 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.

Table 2. Examples of tweets and their labelling.

Tweet	labelling
It's already Christmas Eve? Whoa this sure snuck up on me, lol! Merry Ho everyone!'	three positives
OMG Mitchs dad was in the hospital for the last days and we just found out today now that hes home	three negatives
Morning all. I feel like i've been beaten up.	two negative and one neutral
Hiya! How are you today? What u up to?'	two neutral and one positive
working cough at home cough today guh	one negative and two neutral
Boy I sure had fun at the dentist today Psyche	ambiguous

Process The data annotation was a practical work for the course "Semantic Interpretation of Text" which pre-requisites include Computational Linguistics and Natural Language Processing courses. [11] The students had a sound knowledge of academic and practical English (they participated in "Work and Travel in USA" during summer terms.) 10 annotators were selected through a rigorous process. Our goal was to label each tweet independently by 3 annotators.

We have divided the set of annotated tweets into 3 categories: (a) tweets with strong agreement: all three annotators picked up the same tag (positive, negative or neutral); (b) tweets with weak agreement: two of three annotators picked up the same tag; (c) uncertain tweets: all three annotators picked up different tags (positive, negative and neutral or different parts of the message were annotated with several tags). Table 2 presents some examples.

Discussion The first our annotation experiment was carried out with a set of tweets which contained health-related terms. One of our conclusions was that in many cases it was extremely difficult to annotate scattered tweets without knowing context of a longer discussion. There were many messages which could be understood by the addressee but did not make sense for others. For example,opera 10 feels pretty dang fas or You mean Madman Muntz? What has he got to do with us? True, Don used to sell cars, like Muntz, but long ago, before we met or is so ready for "oh nine" and is so over "oh ate". Great-now he's hungry.

As a result, the next experiment we carried with sequences of three messages: one preceding message, the message with health-related terms and one following message. However, these consequent tweets were not always related. For example, Writing more crack. Draco's gonna break his hand punching stalker!Edward. *evil laugh* preceded a tweet with PHI: Have developed an allergy to fried okra and Arbys chicken Joy, which, in turn, was followed by Beatrice hates me and needs new sparkplugs. All the three messages are somehow ambiguous. Also, humor and irony were difficult for sentiment classification, e.g. Headache good night appeared to be problematic for annotators. The listed challenges, however, did not prevent the annotators from reaching strong agreement in many cases. Table 3 presents results of the the annotated data. Corresponding κ values point to fair to moderate to substantial agreement as we show in Section 4.

[11] http://lilu.fcim.utm.md/teaching.html

Table 3. Distribution of tweets among annotation categories.

annotation	preceding tweets		PHI tweets		next tweets		total	
	#	words	#	words	#	words	#	words
strong agreement	148	1,940	124	2,005	137	1,801	409	5,746
weak agreement	80	1,154	96	1,480	84	1,285	260	3,919
uncertain	29	562	40	695	29	551	98	1,808
total	257	3,656	260	4,180	250	3,637	767	11,473

4 Manual Analysis Results

The similarity of raters' categorization of items into group categories helps to estimate possible risks of future decision making. In a sentiment analysis study, we consider that the raters' agreement can estimate a possible degree of sentiment classification and be a tentative predictor of values of performance evaluation metrics *Fscore* and *Accuracy*. Our task for the assessment of manual evaluation is formulated as follows:

reviewers evaluate multiple rankers' agreement on assigning short messages into sentiment categories; having multiple reviewers (i) reduces an impact of a singular reviewer on the text's sentiment label, (ii) allows to choose a few levels of certainly about the assigned labels: all reviewers agree, some reviewers agree, all disagree.

categories are positive and negative sentiments and neutral; the three categories imply a level of certainty about the assigned sentiments, whereas two categories often signify positive and non-positive sentiments or negative and non-negative sentiments.

Concordance measure For agreement evaluation, we used *Fleiss kappa* (κ). The κ assesses agreement among n users assigning $i = 1, \ldots, N$ items into $j = 1, \ldots, K$ categories [10, 15]. We start with computing how many raters assigned the i_{th} item into the j_{th} category (n_{ij}). Then we compute p_i that evaluates raters' agreement on the i_{th} item and p_j that shows the ratio of all items assigned into the j_{th} category.

$$p_i = \frac{1}{n(n-1)}(\sum_{j=1}^{K} n_{ij}^2 - n) \quad (1) \qquad p_j = \frac{1}{N \cdot n} \sum_{i=1}^{N} n_{ij} \quad (2)$$

An example is given in Table 4. The individual values are averaged with respect to the set of items:

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} p_i \quad (3) \qquad \bar{P}_e = \sum_{j=1}^{N} p_j^2 \quad (4)$$

Finally, the *kappa* is calculated as follows:

Table 4. Examples of text ranking.

Tweets		Sentiment categories			raters	
#	text	pos	neg	neut	#	p_i
1	She should go, as long as it's not his place. Unless she wants that ;)	2	0	1	3	0.333
2	Hooray no insomnia last night Almost finished with cabin web site	1	1	1	3	0.000
3	Helped put away leftovers and feed all the kitties, and now I'm trying to avoid another night of watching crappy Hallmark movies.	0	2	0	2	0.167
4	I didnt know I was pregnant The news numbed me for a while I havent given up riding yet but jumping I had to let go	0	3	0	3	1.000
	p_j	0.623	0.831	0.416		

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{5}$$

where the numerator $\bar{P} - \bar{P}_e$ shows the degree of agreement achieved above chance, and the denominator $1 - \bar{P}_e$ shows the degree of agreement obtainable above chance.

Concordance evaluation *Fleiss Kappa*'s computation is based on the number of reviewers per text(n), the number of texts (N), the number of categories (K). To eliminate a possible evaluation noise, we can introduce a threshold for the agreement on an individual category per text (n_{ij}). Our computation of κ takes advantage of these options:

preliminary agreement in this case, we use all ranked tweets to calculate the agreement; as some tweets were labeled by only two raters, we average $\bar{n} = 2.83$; $N = 767$, $K = 3$;

three raters agreement we calculate the agreement on tweets that have been ranked by three raters: $n = 3$, $N = 686$, $K = 3$; from examples in Table 4, tweet # 3 will be excluded from the data.

strong agreement the agreement is calculated on tweets where two raters agree on the same sentiment: $n_{ij} \geq 2$, $n = 3$, $N = 669$, $K = 3$; from examples in Table 4, only tweets # 1 and 4 will be counted in the data.

κ_F has been used in opinion evaluation in blogs [17]. Agreement among seven raters was computed for five classification categories, including positive, negative, mixed opinions and non-opinionated and non-relevant categories. In that work, the κ scores were divided into 11 groups: from less than chance (< 0) to moderate ($0.51-0.60$) to perfect ($0.91-1.00$). We use the same scale to interpret the scores. We report the obtained scores in Table 5. We also present ranking agreement of individual sentiment categories.

The listed *kappa* scores show the raters' agreement is consistently moderate when all the three tweets' rankings are counted. Agreement on the individual

Table 5. *Fleiss Kappa*, $p_{pos}, p_{neg}, p_{neu}$ scores, $\times 10$; 3 tweets' values were obtained on sequences of three tweets, other values were obtained on sets of individual tweets (Preced. – on tweets preceding the PHI tweets, PHI – the PHI ones, Next – the next after the PHI tweets). **Bold** illustrates the best agreement value for a given sentiment category; we do not emphasize values when there is a tie.

Tweets	Agreement											
	preliminary				three raters				strong			
	κ	p_{pos}	p_{neg}	p_{neu}	κ	p_{pos}	p_{neg}	p_{neu}	κ	p_{pos}	p_{neg}	p_{neu}
3 tweets	*52*	29	24	46	*57*	28	25	47	*59*	28	24	47
Preced.	*54*	33	18	49	*60*	33	18	49	*62*	33	17	49
PHI	*46*	22	33	46	*50*	21	34	46	*47*	22	**35**	48
Next	*55*	32	22	43	*58*	32	29	46	*60*	32	23	46

tweet subsets depends on the tweet category: fair/moderate – for the tweets with PHI, moderate/substantial – for the tweets preceding the PHI, moderate – for the tweets next to the PHI.

Discussion For the sentiment categories, we conclude that annotators find a stronger agreement on *positive* tweets when they either precede or follow the PHI tweet. This mutual understanding holds across all the three agreement assessments. For the PHI tweets, however, the reverse tendency is true: raters stronger agree on *negative* sentiments than on *positive* ones.

To assess the impact of changes in the ranked tweet selection, we applied the paired *t-test* to estimate commonalities between the obtained scores. In our case, the test examines the null hypothesis that there is no mean difference between two sets of the *kappa* scores (i.e., the difference mean is equal to 0). Difference between the sets of *kappa* values in preliminary and three raters' agreement was deemed statistically significant (P = 0.0061). The further tightening of ranking conditions did not significantly alter the rater agreement (P= 0.5908). Hence, the null hypothesis was rejected for the preliminary – three raters comparison pair and accepted for the three raters – strong comparison pair.

The positive sentiment ranking p_{pos} was uniform across a given tweets' choice and near independent from the agreement case: 0.33 – for preceding tweets, 0.32 – for next tweets, 0.21-0.22 – for PHI tweets, and 0.28-0.29 – for the 3 tweets' set. The negative sentiment ranking p_{neg}, too, was uniform across a given tweets' choice and near independent from the agreement case: 0.18 – for preceding tweets, 0.22 – for next tweets, 0.33-0.35 – for PHI tweets, and 0.24-0.25 – for the 3 tweets' set. Agreement on neutral tweets was 0.43-0.49 for all the sets.

Presence of health information in tweets had a major impact on the sentiment ranking, as those tweets contain more negative sentiments than the preceding or next ones. When we compare tweets with health information and tweets without health information, we see that raters' agreement has been reversed for both positive and negative sentiments. On tweets *without* health information, raters' p_{pos} was 0.32-0.33 and p_{neg} was 0.17-0.23. On other hand, on tweets *with* health information, p_{pos} was 0.22 and p_{neg} was 0.33-0.35. As a result, the κ scores

changed from fair/moderate on tweets with PHI to moderate/substantial on other tweets.

5 Sentiment Classification Results

For the machine learning part of our studies, we used tweets with the strong ranking agreement. The data set contained all the three types of tweets: tweets with personal health information, tweets preceding them and tweets next to them. Each tweet was labeled with the sentiment assigned by the majority of raters. We investigated the ability of learning algorithms to distinguish between positive and negative sentiments and neutral ones. We applied Naive Bayes (NB), Decision Trees (DT), K-Nearest Neighbor (KNN) and Support Vector Machines (SVM). Average *Fscore*(F), *Precision*(Pr), *Recall*(R) and *ROC* were used to evaluate the performance.

We represented the data set through all the words that appear in the set more than twice. We opted for the statistical feature selection approach instead of semantic, as tweets are short texts, with a high variety of lexical units and semantic generalization can be challenging. The following sets of features were selected for ML experiments: a) bag of words occurred > 2 – 1015 features (BoW2) , b) bag of words occurred > 5 – 312 features (BoW5); c) words that, individually, are highly correlated with the class label and have a low inter-correlation (CorrelatW); the former evaluates the predictive power of an individual word and the latter estimates the word redundancy; d) words that form a subset better consistent with the class labels when evaluated on the training set (ConsistSubs). Two learning settings were considered: 1) three-class classification of positive, negative and neutral tweets; 2) binary classification of positive and negative tweets. We combined *Fscore* and *Accuracy* to measure the goodness of results: from several algorithm settings that output the same *Recall* we chose the one that gave us a higher *Fscore*.

Table 6 reports the best results of *three class* classification:
for BoW2: DT – learning coefficient $\alpha = 0.10$, K-NN – 2 neighbors, inverse-distance-weighted; the multinomial NB; SVM – complexity parameter C = 3.0, kernel polynomial K= 1.0.
for BoW5: DT – learning coefficient $\alpha = 0.10$, K-NN – 1 neighbor, Euclidean distance; the updateable multinomial NB; SVM – complexity parameter C = 3.0, kernel polynomial K= 4.0.
for CorrelatW: DT – learning coefficient $\alpha = 0.20$, K-NN – 1 neighbor, similarity-weighted distance; NB – with kernel estimates; SVM – complexity parameter C = 3.0, kernel polynomial K= $\sum_{i=1}^{4} i$.
for ConsistSubs: DT – learning coefficient $\alpha = 0.30$, K-NN – 1 neighbor, similarity-weighted distance; NB – with kernel estimates; SVM – complexity parameter C = 5.0, kernel polynomial K= $\sum_{i=1}^{4} i$.

Table 7 reports the best results of *binary* classification:
for BoW2: DT – learning coefficient $\alpha = 0.35$, K-NN – 1 neighbor, Euclidean distance; the multinomial NB; SVM – complexity parameter C = 2.0, kernel

Algor	BoW2				BoW5				CorrelatW				ConsistSubs			
	Pr	R	F	ROC	Pr	R	F	ROC	Pr	R	F	ROC	Pr	R	F	ROC
DT	49.7	51.9	48.4	58.9	49.1	51.6	47.8	58.1	55.7	53.6	46.1	54.5	56.1	53.6	45.9	54.2
K-NN	55.2	56.0	51.3	59.2	53.7	.54.4	51.3	61.5	70.0	67.9	66.0	73.5	72.8	69.6	67.8	74.2
NB	60.1	60.3	59.2	72.9	60.3	60.8	60.0	71.4	70.7	68.1	66.2	75.7	71.7	68.9	67.0	**75.7**
SVM	62.4	62.9	62.1	69.9	59.6	60.3	58.9	65.3	72.5	69.2	67.2	73.3	**75.3**	**71.0**	**69.2**	74.2

Table 6. Multi-class classification results for positive, negative and neutral tweets (%). Best values are in **bold**. Baseline is calculated if all the sentences are into the majority class (%): Pr = 24.2, R = 49.2, F = 32.5, ROC = 49.9.

polynomial K= 1.0.
for BoW5: DT – learning coefficient $\alpha = 0.35$, K-NN – 1 neighbor, similarity-weighted distance; NB – multinomial; SVM – complexity parameter C = 4.0, kernel polynomial K= $\sum_{i=1}^{4} i$.
for CorrelatW: DT – learning coefficient $\alpha = 0.30$, K-NN – 1 neighbor, Euclidean distance; NB – with kernel estimates; SVM – complexity parameter C = 1.0, kernel polynomial K= 2.0.
for ConsistSubs: DT – learning coefficient $\alpha = 0.30$, K-NN – 1 neighbor, Euclidean distance; NB – with kernel estimates; SVM – complexity parameter C = 5.0, kernel polynomial K= $\sum_{i=1}^{2} i$.

Algor	BoW2				BoW5				CorrelatW				ConsistSubs			
	Pr	R	F	ROC	Pr	R	F	ROC	Pr	R	F	ROC	Pr	R	F	ROC
DT	64.8	64.5	64.6	67.0	65.5	65.3	65.3	69.2	61.2	60.0	57.5	65.1	66.9	59.7	53.0	56.2
K-NN	60.3	57.6	56.1	63.5	62.7	61.9	61.8	68.0	72.2	71.1	70.4	82.7	80.0	74.4	72.7	74.4
NB	75.7	75.3	75.1	**83.1**	71.7	71.4	71.3	77.5	78.6	75.6	74.5	**85.9**	**82.4**	**77.8**	**76.6**	76.1
SVM	71.4	71.5	71.4	71.3	68.0	67.8	67.8	67.9	76.4	73.9	72.9	72.9	80.5	73.9	71.9	72.5

Table 7. Binary classification results for positive and negative tweets (%). Best values are in **bold**. Baseline is calculated if all the sentences are into the majority class (%): P = 27.9, R = 52.8, F = 36.5, ROC = 50.0.

Discussion In the three-class classification, in terms of *Precision*, *Recall*, *Fscore*, SVM consistently outperformed other methods on BoW2, CorrelatW and ConsistSubs features sets. At the same time, NB was the best in terms of *ROC* and on the BoW5 feature set. In binary classification, NB obtained better results for all given feature sets. Our results are competitive with previously obtained results. As reported in [18], 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 *Fscore* 39% –67%, *Fscore* increased to 53% – 75% when the posts were enriched with the Web information.

6 Conclusions and Future Work

We have presented a study of sentiments and opinions in tweets related to personal health. In those tweets, users discussed health and ailment, treatments of themselves and their relations. We have used an author-centric annotation model first introduced in [20]. The annotation model shows how positive, negative and neutral sentiments can be identified in health-related tweets.

To assess the quality of sentiment classification, we have decided on *positive, negative, neutral* categories. We chose three categories to better see on what annotators may agree on *what constitutes* a subjective label and disagree on *what does not*; in other words, their understanding of *positive* may be close and their understanding of *not positive* may be far apart. We have applied *Fleiss Kappa* to evaluate the inter-rater agreement. The obtained κ scores indicated fair/moderate and moderate/substantial agreement.

In the machine learning studies, we ran three-class and binary classification experiments. Tweets were represented through the individual words appeared in them. Bag-of-word representation provided an estimate for expected results. We have applied statistical feature selection methods that allowed us to obtain results on subsets of words. In three-class classification, SVM had performed better than other algorithms. In binary classification, the best results were obtained by NB.

Our future work will focus on studies of threads which contain tweets with personal information. On that stage, we will analyze a thread as an entity and look for patterns of subjectivity expressions. We also plan to analyze user posts on other types of social media (e.g., social networks).

Acknowledgements

This work is in part funded by an NSERC Discovery grant.

References

1. Allan, K. *Explorations in Classical Sociological Theory: Seeing the Social World.* Pine Forge Press, 2005.
2. 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
3. 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.
4. Chen, W. Dimensions of Subjectivity in Natural Language (Short Paper). In *Proceedings of ACL-HLT*, 2008.
5. Chew, C. and G. Eysenbach. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLoS One*, **5**(11), 2010.

6. Dijck, van J. Users like you? Theorising agency in user-generated content. *Media, Culture & Society*, **31**(1): p.p. 41 58, 2009.
7. Dodds, P., K. Harris, I. Kloumann, C. Bliss, C. Danforth. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PLoS ONE*, **6**, e26752, 2011.
8. Doing-Harris, K. and Q. Zeng-Treiler. Computer-Assisted Update of a Consumer Health Vocabulary Through Mining of Social Network Data. *Journal of Medical Internet Research*, **13**(2):e37, 2011.
9. Eysenbach, G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behaviour on the Internet. *Journal of Medical Internet Research*, **11**(1), 2009.
10. Green, A. Kappa statistics for multiple raters using categorical classifications. *Proceedings of the 22nd Annual Conference of SAS Users Group*, 1997.
11. Jansen, B.J., Zhang, M., Sobel, K., and Chowdury, A. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, **60**(11), 2169–2188, 2009.
12. Lampos, V. and N. Christianini. "Tracking the flu pandemic by monitoring the social web". *2nd Workshop on Cognitive Information Processing*, 2010.
13. Mihalcea, R. and H. Liu, A corpus-based approach to finding happiness, *Proceedings of the AAAI Spring Symposium on Computational Approaches to Weblogs*, 2006.
14. Mislove, A., S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. Niels Rosenquist. Understanding the Demographics of Twitter Users. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, 2011.
15. Nichols, T., P. Wisner, G. Cripe, and L. Gulabchand. Putting the Kappa Statistic to Use. *Qual Assur Journal*, **13**, 57–61, 2010.
16. O'Connor, B., R. Balasubramanyan, B. Routledge, and N. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM'10)*, 122–129, 2010.
17. Osman, D., J. Yearwood, P. Vamplew. Automated opinion detection: Implications of the level of agreement between human raters. *Information Processing and Management*, **46**, 331–342, 2010.
18. Sokolova, M., G. Lapalme. Learning opinions in user-generated Web content. *Journal of Natural Language Engineering*, 2011.
19. Sokolova, M. and D. Schramm. Building a patient-based ontology for mining user-written content. *Recent Advances in Natural Language Processing*, p.p. 758–763, 2011.
20. Sokolova, M. and V. Bobicev. Sentiments and Opinions in Health-related Web messages. *Recent Advances in Natural Language Processing*, p.p. 132– 139, 2011.
21. Strapparava, C., R. Mihalcea Learning to Identify Emotions in Text, *Proceedings of the 2008 ACM symposium on Applied Computing* 2008
22. 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
23. Thelwall, M., K. Buckley, and G. Paltoglou. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, 2010.