Learning Sentiments from Tweets with Personal Health Information

Victoria Bobicev[1], Marina Sokolova[2,3,4], Yasser Jafer[3], and David Schramm[4,5]

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

Abstract. We present results of sentiment analysis in Twitter messages that disclose personal health information. In these messages (tweets), users discuss ailment, treatment, medications, etc. We use the author-centric annotation model to label tweets as positive sentiments, negative sentiments or neutral. The results of the agreement among three raters are reported and discussed. We then use Machine Learning methods on multi-class and binary classification of sentiments. The obtained results are comparable with previous results in the subjectivity analysis of user-written Web content.

Keywords: sentiment analysis, personal health information, Twitter

1 Introduction

Web 2.0 technologies allowed the general public actively contribute to the Web content. Blogosphere, social networks, message boards are awash with users' personal news that, in most cases, can be read without limitation by a global community. Those readers are influenced by emotional appeal of the content, as emotion-rich text affects a public mood stronger than rational arguments [1]. Previous studies had shown that expressed sentiments in text relate to the author's personal health [12]. The relations, however, were studied on a small number of texts written by few authors. In the current study, we analyze the relations on a massive amount of text contributed by many authors.

Twitter, the world's tenth most popular Web site, is a micro-blogging service with instant message postings.[6] It has > 200 $mln.$ users.[7] 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.

Twitter messages (i.e., tweets) present a real-time means of estimating public interest in various subjects, including personal health and expressed sentiments.

[6] http://twitter.com/
[7] alexa.com/topsites

The health-related messages often reveal information that previously was discussed in clinical or family settings: ailment, treatment, medications [8, 10].

Our study focusses on sentiments in tweets related to personal health. Although sentiment analysis is a popular text mining discipline, there are few publications that consider sentiments in relation with personal health information posted on the Web. In [15], the authors analyzed opinions and sentiments expressed in the sci.med messages of *20 NewsGroups*. They evaluated concordance of the manual annotation by two raters. The results show that raters stronger agree on what type of sentences do *not* belong to positive or negative subjective categories. 16 categories of opinions and emotions in tweets were manually analyzed [4]. The extraction method traced tweets with H1N1 and its synonyms (e.g., swine flu).

Our current work includes both manual and automated components: sentiment tagging of tweets performed by multiple raters and machine learning multi-class and binary classification of sentiments. The following sections discuss the Twitter data, our annotation model and process, the rater agreement evaluation, and the application of machine learning methods and the obtained classification results.

2 Tweets with Personal Health Information

We had an access to 30,164 Twitter threads (i.e., consequent tweets posted by a user).[8] An average length of a thread is 560 words, albeit some words can be 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.

We used two lexical resources for identification of tweets containing personal health information (PHI). First, we used ontology of personal health terms which lists terms related to body organs, symptoms, treatment, medical professional designations [14]. Semantic information from WordNet [9] 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 tweets with PHI. 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

[8] http://caw2.barcelonamedia.org/node/7
[9] http://wordnet.princeton.edu/

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

annotation	preceding tweets		tweets with PHI		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

PHI. We, then, manually confirmed the presence of personal health information in the extracted tweets. Some tweets explicitly referred to a person and his/her health, Mitchs dad was in the hospital for the last days, some tweets disclose personal health information without direct reference to a person, Headache is not going away. The latter messages are often more informal, then those which contain personal terms. We found that the number of tweets with PHI was consistent for all the five folders of data. For each tweet with PHI, we then worked with the thread from which it was extracted and retrieved 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 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 [16] or the author of a text [2]. Our annotation model was author-centric and followed the model we used for sentiment annotation of user messages on online patient forums [15]. We suggested that a rater imagined sentiments and attitudes that the author possibly had while writing.

Subjective expressions are highly reflective of the text content and context [3], and text related to personal health brings in an additional challenge of separating good and bad news from sentiments. Health-related messages can be distressing when written about personal illnesses or sick relatives and friends. Hence, we asked raters 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 raters not to agonize over the annotation and, if doubtful, leave the example un-annotated. The rule is especially impor-

tant for annotation of tweets, when raters 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; raters 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.

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. 10 raters were selected through a rigorous process. Our goal was to label each tweet independently by 3 raters.

Process Our annotation started on 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 to other readers. 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, we followed with annotation of 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 raters.

After the tweets were labelled, we have divided them into 3 categories: (a) tweets with strong agreement: all three raters picked up the same tag (positive,

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	uncertain

Table 3. Distribution of tweets among annotation categories.

annotation	preceding tweets		tweets with PHI		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

negative or neutral); (b) tweets with weak agreement: two of three raters picked up the same tag; (c) uncertain tweets: all three raters picked up different tags. Table 2 presents examples of each category.

The data challenges, however, did not prevent the raters from reaching strong agreement in many cases. Table 3 presents results of the the annotated data.

4 Assessment of the Annotation

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

- evaluate agreement among three rankers on assigning tweets into sentiment categories; having multiple raters (i) reduces an impact of a single rater on the text's sentiment label, (ii) allows to choose a few levels of certainly about the assigned labels: all the raters agree, some raters agree, all disagree.
- differentiate positive sentiments, negative sentiments and neutral; the three categories imply a level of certainty about the assigned sentiment tags, 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 raters assigning $i = 1, \ldots, N$ tweets into $j = 1, \ldots, 3$ sentiment categories [6, 9]. We start with computing how many

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		

raters assigned the i_{th} tweet into the j_{th} sentiment category (n_{ij}). Then we compute p_i that evaluates raters' agreement on the i_{th} tweet and p_j that shows the ratio of all tweets assigned into the j_{th} sentiment category.

$$p_i = \frac{1}{n(n-1)}(\sum_{j=1}^{3} n_{ij}^2 - n) \quad (1) \qquad p_j = \frac{1}{N \cdot n}\sum_{i=1}^{N} n_{ij} \quad (2)$$

We show the p_i, p_j computation example in Table 4. The individual values p_i, p_j are then summarized and averaged, to show an average agreement per tweet ($\bar{P}$) and per sentiment category ($\bar{P_{sent}}$):

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

Finally, the *kappa* is calculated as follows:

$$\kappa = \frac{\bar{P} - \bar{P}_{sent}}{1 - \bar{P}_{sent}} \quad (5)$$

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

Concordance evaluation We assessed the rater agreement under three types of rating conditions:

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 $\bar{n} = 2.83$; $N = 767$, $K = 3$;

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

To eliminate a possible evaluation noise, we can introduce thresholds for n, the average number of raters per tweet, and for n_{ij}, the agreement on an individual sentiment category per tweet:

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.

We report the obtained scores in Table 5. We also present rater agreement of positive, negative and neutral categories:

$$p_{pos} = \frac{1}{N}\sum_{i=1}^{N} n_{ipos} \quad (6) \qquad p_{neg} = \frac{1}{N}\sum_{i=1}^{N} n_{ineg} \quad (7) \qquad p_{neu} = \frac{1}{N}\sum_{i=1}^{N} n_{ineu} \quad (8)$$

Fleiss Kappa has been used in opinion evaluation in blogs [11]. 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.

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

5 Sentiment Classification Results

For the machine learning (ML) 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) [19].

Average $Fscore(F)$, $Precision(Pr)$, $Recall(R)$ and $AreaUnderCurve(AUC)$ were used to evaluate the performance.

$$Precision = \frac{tp}{tp + fp} \quad (9) \qquad Recall = \frac{tp}{tp + fn} \quad (10)$$

$$Fscore = \frac{2tp}{2tp + fn + fp} \quad (11) \qquad AUC = \frac{1}{2}(\frac{tp}{tp + fn} + \frac{tn}{tn + fp}) \quad (12)$$

where tp – correctly recognized positive examples, tn– correctly recognized negative examples, fp – negative examples recognized as positives, fn- - positive examples recognized as negatives.

We used two supervised learning settings: 1) three-class classification of positive, negative and neutral tweets; 2) binary classification of positive and negative tweets. We combined *Recall* and *Fscore* to determine the optimal classifier: from the set of adjustable parameters that output a classifier with the same *Recall*, we chose the parameters that gave us a higher *Fscore*.

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. First, we represented the data set through all the words that appear in the set more than twice (BoW2) – 1015 features. Next, reduced sets of features were selected for following experiments: bag of the words that occurred > 5 – 312 features (BoW5); words that are highly correlated with the class label, but have a low inter-correlation among themselves (CorrelatW); words that form a subset of words which showed a better consistency with the class labels on the training set (ConsistSubs).

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.

Algor	BoW2				BoW5				CorrelatW				ConsistSubs			
	Pr	R	F	AUC	Pr	R	F	AUC	Pr	R	F	AUC	Pr	R	F	AUC
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, AUC = 49.9.

Algor	BoW2				BoW5				CorrelatW				ConsistSubs			
	Pr	R	F	AUC	Pr	R	F	AUC	Pr	R	F	AUC	Pr	R	F	AUC
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, AUC = 50.0.

for CorrelatW: DT – learning coefficient α = 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 α = 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 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= $\sum_{i=1}^{2} i$.

6 Discussion 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. Twitter data became a subject of sentiment analysis research. In [5], 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 [17]. It investigates whether popular events are typically associated with increases in sentiment strength. In [7], the authors compared manual coding and sentiment classification of tweets containing branding comments. Topic-specific opinions in blogs were evaluated in [11]. Agreement among seven raters was computed for five classification categories, including positive, negative, mixed opinions and non-opinionated and non-relevant categories. Results of manual emotion classification of 819 MySpace messages were reported in [18].

We have used an author-centric annotation model first introduced in [15]. 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 raters 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.

For the sentiment categories, we conclude that raters find a stronger agreement on *positive* tweets when they either precede or follow the PHI tweet. This mutual understanding holds across all the three types of 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.

In the supervised learning studies, we applied DT,K-NN,NB,SVM. We ran three-class and binary classification experiments. Tweets were represented through the individual words appeared in them. Bag of the words with occurrence > 2 provided an estimate for expected results (BoW2). We have applied statistical feature selection methods that allowed us to represent tweets by subsets of the words. We considered a subset of more frequent words with occurrence > 5 (BoW5), the words with a high prediction of the sentiment class and a low redundancy among them (CorrelatW), and a subset of words which showed a high consistency with the sentiment class labels when evaluated on the training set (ConsistSubs).

In three-class classification, SVM had performed better than other algorithms. 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 *AUC* 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 [13], 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.

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 and CHEO Research 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. Chen, W. Dimensions of Subjectivity in Natural Language (Short Paper). In *Proceedings of ACL-HLT*, 2008.
4. 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.
5. 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.
6. Green, A. Kappa statistics for multiple raters using categorical classifications. *Proceedings of the 22nd Annual Conference of SAS Users Group*, 1997.
7. 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.
8. Lampos, V. and N. Christianini. "Tracking the flu pandemic by monitoring the social web". *2nd Workshop on Cognitive Information Processing*, 2010.
9. Nichols, T., P. Wisner, G. Cripe, and L. Gulabchand. Putting the Kappa Statistic to Use. *Qual Assur Journal*, **13**, 57–61, 2010.
10. 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.
11. 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.
12. Pennebaker, J. and Chung, C. Expressive Writing, Emotional Upheavals, and Health. *Handbook of Health Psychology*, Friedman, H. and R. Silver. (eds.), Oxford University Press, 2006.
13. Sokolova, M., G. Lapalme. Learning opinions in user-generated Web content. *Journal of Natural Language Engineering*, 2011.
14. 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.
15. Sokolova, M. and V. Bobicev. Sentiments and Opinions in Health-related Web messages. *Recent Advances in Natural Language Processing*, p.p. 132– 139, 2011.
16. Strapparava, C., R. Mihalcea Learning to Identify Emotions in Text, *Proceedings of the 2008 ACM symposium on Applied Computing* 2008
17. Thelwall, M., K. Buckley, and G. Paltoglou. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*,**62**(2), 406–418, 2010.
18. Thelwall, M., Wilkinson, D. and S. Uppal. Data Mining Emotion in Social Network Communication: Gender Differences in MySpace. *Journal of the American Society for Information Science and Technology*,**61**(1), 190–199, 2010.
19. Witten, I., E. Frank, M. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufman, 2011.