Recognition of Sentiment Sequences in Online Discussions

Victoria Bobicev
Technical University of Moldova
vika@rol.md

Marina Sokolova
University of Ottawa,
Institute for Big Data Analytics, Canada
sokolova@uottawa.ca

Michael Oakes
University of Wolverhampton, UK
Michael.Oakes@wlv.ac.uk

- 19%-28% of Internet users participate in online health discussions.
- In North America, 59% of all adults have looked online for information about a disease or treatment.
- Up to **49%** of the users are most interested in personal testimonials related to health

Personal Health Information

- Personal health information (PHI) is information about one’s health discussed by a patient in a clinical setting
- PHI is the most vulnerable private information posted online
  - *I have a family history of Alzheimer's disease. I have seen what it does and its sadness is a part of my life. I am already burdened with the knowledge that I am at risk.*
  - *We're going for the basic blood tests, the NT scan, and the "Ashkenazi panel" since both XX and I are Jewish from E. European descent.*

Motivation

- Health information posted by the general public is important for the development of health care policies
  - *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.*

- Previous to emergence of social networks, subjective health information had been analysed on restricted and controlled groups (e.g., nuns from the same monastery, patients of the same clinic)

- Data harvested from social networks provide an opportunity for development of social mining techniques

Outline

- We present sentiment analysis of messages posted on a medical forum.
- We categorize posts into five categories: *encouragement, gratitude, confusion*, *facts,* and *endorsement*.
- Our empirical results are obtained on 1438 messages from 130 discussions dedicated to infertility treatments.
- Our analysis concentrates on sequences of sentiments in the forum discourse.

Example

- *Alice*: Jane - whats going on??
- *Jane:* We have our appt. Wednesday!! EEE!!!
- *Beth*: Good luck on your transfer! Grow embies grow!!!!
- *Jane*: The transfer went well - my RE did it himself which was comforting. 2 embies (grade 1 but slow in development) so I am not holding my breath for a positive. This really was my worst cycle yet ; it was the Antagonist protocol which is supposed to be great when you are over 40 but not so much for me!!

Data

- We looked for discussions where the forum participants discussed only one topic.
  - A preliminary analysis showed that discussions with ≤ 20 posts satisfied this condition.
- We wanted discussions be long enough to form a meaningful discourse.
  - This condition was satisfied when discussion had ≥ 10 messages.
- As a result, 80 discussions were selected for a manual analysis; average of 17 messages per discussion.

Challenges of Sentiment and Opinion Mining in Health-related Messages

Sentiment: *I am sickened by the thought …*	Ailment: *I feel sick for awhile; should see my physician*
Opinion: *I think it is evident that …*	Improvement: *The benefit is usually evident within a few days of starting it*
Humor: *don't forget that it's better for your health to enjoy your steak than to resent your sprouts*	Complain: *After that my health deteriorated …*

Modus Operandi

- Data annotation by 2 annotators
- Minimal text pre-processing
- Domain-specific resource (i.e., HealthAffect lexicon)
- Use of *robust* Machine Learning methods
  - Naive Bayes, Logistic Regression
- Appropriate evaluation metrics

Annotation Process

We used 292 random posts to verify whether the messages were self-evident for sentiment annotation or required an additional context.

The annotators reported that posts were long enough to convey emotions and in most cases there was no need for a wider context

Two raters annotated each post with the dominant sentiment.

Only author's subjective comments were marked as such; if the author conveyed sentiments of others, we did not mark it.

We obtained Fleiss Kappa = 0.737 which indicated a strong agreement between annotators.

Class distribution of the IVF posts

Classification category	**# posts**	**Per-cent**
Facts	494	34.4%
Encouragement	333	23.2%
Endorsement	166	11.5%
Confusion	146	10.2%
Gratitude	131	9.1%
Ambiguous	168	11.7%
Total	1438	100%

The most frequent sequences of sentiments

Sentiment pairs	**Occurrence**	**Percent**
facts, facts	170	19.5%
encouragement, encouragement	119	13.7%
facts, encouragement	55	6.3%
endorsement, facts	53	6.1%
encouragement, facts	44	5.1%

Sentiment triads	**Occurrence**	**Percent**
factual, factual, factual	94	12.8%
encouragement, encouragement, encouragement	63	8.6%
encouragement, gratitude, encouragement	18	2.4%
factual, endorsement, factual	18	2.4%
confusion, factual, factual	17	2.3%

HealthAffect

- We adapted the Pointwise Mutual Information (PMI) of *word*1 and *word*2 (Turney, 2002):

PMI(*word*1, *word*2) = $\log_2$(p(*word*1 & *word*2)/(p(*word*1) p(*word*2)))

- First, we created a list of *phrases, i.e.,* all unigrams, bigrams and trigrams, of words with frequency ≥ 5 from the unambiguously annotated posts.
- Then, for each class, we calculated

PMI(*phrase*, *class*) = $\log_2$(p(*phrase* in *class*)/(p(*phrase*) p(*class*))).

- Finally, we calculated Semantic Orientation (SO) for each term:

SO(*phrase, class*) = PMI(*phrase*, *class*) - Σ PMI(*phrase*, *other_classes*)

- 431 unigrams, 555 bigrams, 214 trigrams

Sentiment Recognition

- We calculated the number of HealthAffect terms from each category in the post and classified the post in the category for which the maximal number of terms was found.
- The algorithm's performance was evaluated through two multiclass classification results:
  - 4-class classification where all 1269 unambiguous posts are classified into (*encouragement, gratitude, confusion*, and neutral, i.e., *facts* and *endorsement*), and
  - 3-class classification (positive: *encouragement, gratitude;* negative: *confusion*, neutral: *facts* and *endorsement*).

Classification Accuracy

Metrics	4-class classification	3-class classification
microaverage F-score	0.633	0.672
macroaverage Precision	0.593	0.625
macroaverage Recall	0.686	0.679
macroaverage F-score	0.636	0.651

Sentiment Classification

- The most accurate classification occurred for *gratitude*. It was correctly classified in 83.6% of its occurrences. It was most commonly misclassified as *encouragement* (9.7%).
- The second most accurate result was achieved for *encouragement.* It was correctly classified in 76.7% of cases. It was misclassified as neutral, i.e. *facts + endorsement,* in 9.8%.
- The least often correctly classified class was neutral (50.8%). One possible explanation is the presence of the sentiment bearing words in the description of facts in a post which is in general objective and which was marked as factual by the annotators.

Related Work

- 16 categories of opinions and emotions in health-related tweets were presented in (Chew and Eysenbach, 2010).
- Sokolova and Bobicev (2011) studied positive and negative opinions and positive and negative sentiments in the health-related sci.med messages from *20 NewsGroups*
- Sentiments in health-related tweets were studied in (Bobicev et al, 2012).
- sentiment propagation among related semantic concepts has been studied by Tsai et al, 2013.

Discussion

- We obtained a strong inter-annotator agreement between two independent annotators: Fleiss Kappa = 0.73. The Kappa values demonstrated an adequate selection of classes of sentiments and appropriate annotation guidelines.
- A specific set of sentiments on the IVF forum suggested that we applied the PMI approach to build a domain-specific lexicon HealthAffect.
- Manual analysis of a sample of data showed that discussion contained a coherent discourse. Some unexpected shifts in the discourse flow were introduced by a new participant joining the discussion.
- In future work, we may include the post’s author information in the sentiment interaction analysis.
- One future possibility is to construct a Markov model for the sentiment sequences. However, in any online discussion there are random shifts and alternations in discourse which complicate application of the Markov model.

Markov Model: the transition matrix of sentiments (rows = previous, columns = next) for the thread “Anyone in the group TTC?”

	factual	encourage-ment	gratitude	confusion	end
Start	48	9	0	108	0
Factual	1583	837	270	181	87
Encourage-ment	874	836	260	129	59
Gratitude	229	224	91	54	22
confusion	333	283	16	99	7

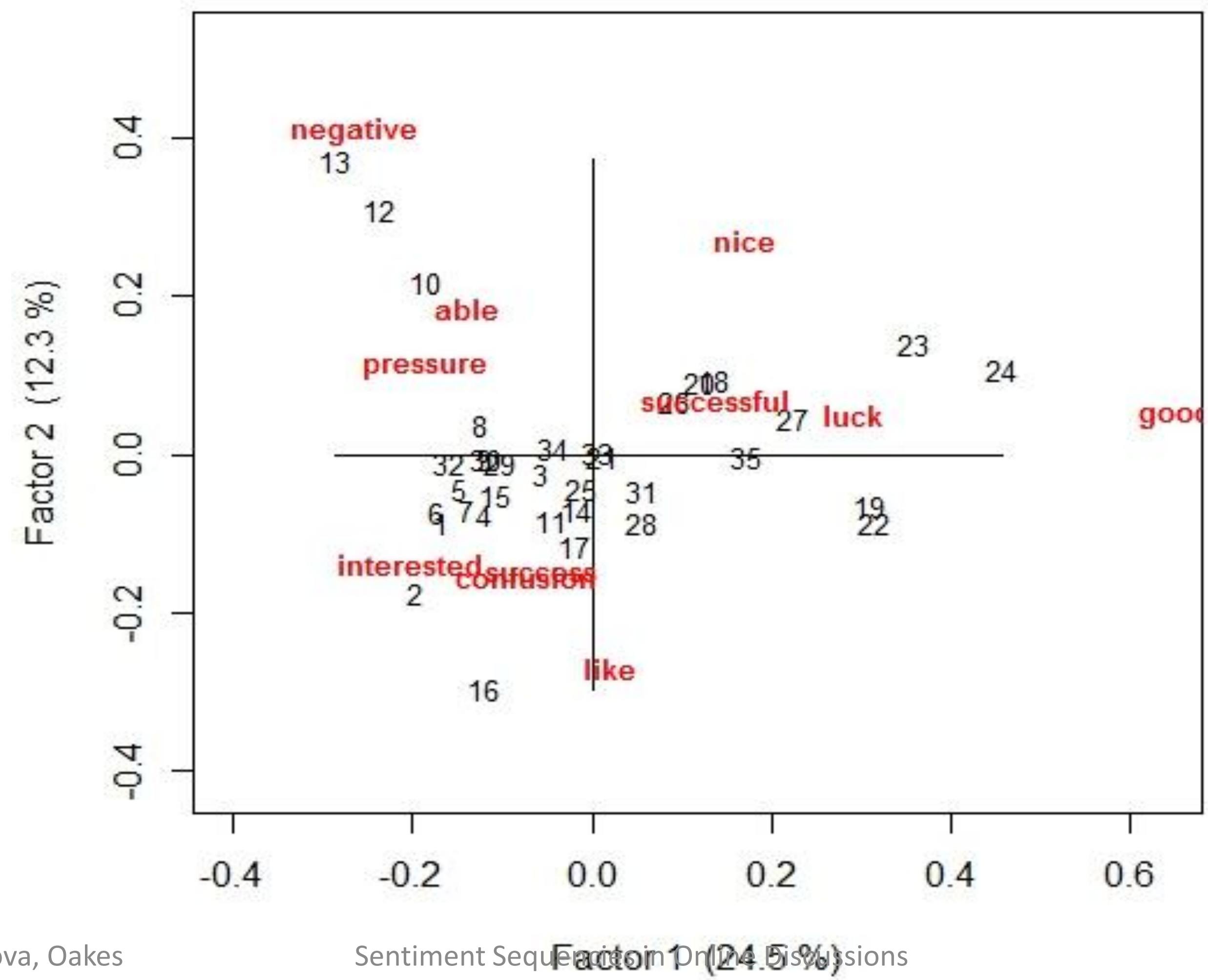
Transform into a matrix of probabilities, by dividing each value by the row total

	factual	encourage-ment	gratitude	confusion	end
Start	0.291	0.055	0	0.655	0
Factual	0.535	0.283	0.091	0.061	0.029
Encourage-ment	0.405	0.387	0.120	0.060	0.027
Gratitude	0.363	0.371	0.144	0.086	0.035
confusion	0.451	0.383	0.022	0.134	0.009

Markov Model: analogy with CLAWS part-of-speech tagger

- Imagine annotators / machine classifiers cannot decide between confusion and factual for the first post.
- Which is more likely, a) start → confusion → factual, or b) start → factual → factual?
- a) 0.655 x 0.451 = 0.295
- b) 0.291 x 0.535 = 0.155

Correspondence Analysis of a sequence of postings



Thank you!

Questions?