The Importance of Pronouns to Sentiment Analysis: Online Cancer Survivor Network Case Study

Nir Ofek, Lior Rokach
Dept. of Information Systems Eng., Ben-Gurion University of the Negev, Israel
{nirofek, liorrk}@bgu.ac.il

Cornelia Caragea
Computer Science and Engineering, University of North Texas, Denton, TX, USA
ccaragea@unt.edu

John Yen
College of Information Sciences and Technology, Pennsylvania State University, University Park, PA, USA
jyen@ist.psu.edu

ABSTRACT

Online health communities are a major source for patients and their informal caregivers in the process of gathering information and seeking social support. The Cancer Survivors Network of the American Cancer Society has many users and presents a large number of user interactions with regards to coping with cancer. Sentiment analysis is an important process in understanding members' needs and concerns and the impact of users' responses on other members. It aims to determine the participants' subjective attitude and reflect their emotions. Analyzing the sentiment of posts in online health communities enables the investigation of various factors such as what affects the sentiment change and discovery of sentiment change patterns. Since each writer has his or her own personality, and temporal emotional state, behavioral traits can be reflected in the writer's writing style. Pronouns are function-words which often convey some unique styling patterns into the texts. Drawing on a lexical approach to emotions, we conduct factor analysis on the use of pronouns in self-descriptions texts. Our analysis shows that the usage of pronouns has an effect on sentiment classification. Moreover, we evaluated the use of pronouns in our domain, and found it different than standard English usage.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text Analysis

Keywords

Pronouns, Behavioral traits, Sentiment analysis, Self-descriptions

1. INTRODUCTION

Text classification aims to label documents or any other text fragment into one or more class buckets. In many classification tasks, the classes are categories such as *economy* and *sport*. Since these tasks do not depend on the subjectivity of the writer, often function-words are omitted in the learning process. However, other classification tasks such as sentiment classification aim to capture the emotional state, or personality of the writer. For these tasks, some function-words such as pronouns could be beneficial signals, since they often convey some unique styling patterns into their texts, and may imply subjectivity [4]. It is known that the writer's language is valuable in diagnosing his/her personality or emotional state and classifying dimensions of traits [2].

We hypothesize that by discovering the pronouns that are more heavily associated with certain class types, we are able to distill personality and emotional states from language. This could be useful for many text classification analyses, especially, when the writer's personality or emotional state are at focus [1, 3]. We evaluated our hypothesis on a sentiment classification task by analyzing an online health forum.

2. CASE STUDY DATA: CANCER SURVIVORS NETWORK

People diagnosed with cancer, as well as caregivers for individuals with cancer, join the Cancer Survivors Network[1] (CSN) of the American Cancer Society to seek social support and information from members who have experienced a particular situation first hand, as well as to seek emotional support. Initiated in June 2000, the Cancer Survivors Network currently has more than 164,000 member participants and offers a way to share people's experiences about cancer and cancer treatments and support for one another. Network members can benefit from the understanding of emotional impacts of online participation on survivors [5]. This may help survivors themselves as well as their informal caregivers since it allows providing useful insight into the design of new features or the enhancement of the existing ones in improving the facilitation of emotional support. Sentiment analysis is an important process in understanding participants' needs and concerns and the impact of users' responses on other members. It aims to determine the participants' subjective attitude and reflect their emotions. Analyzing the sentiment of posts in online health communities enables the investigation of various factors, such as what affects sentiment change and the discovery of sentiment change patterns [5].

3. EXPERIMENTS

Qiu et al. [5] collected forum posts from July 2000 to October 2010 comprised of 48,779 forum threads and more than 468,000 posts from 37,922 participants. In order to train a sentiment classifier, they randomly sampled 298 posts from the CSN breast cancer forum and each post was manually annotated as having a positive (204 posts) or a negative sentiment (94 pots). Examples from each class are shown in Table 1 (see [5] for further details on data sampling and annotation).

Table 1. Examples of posts and their sentiment labels.

Label	Post
Positive	Make me go to itunes to see if i can find it ... . lol starts out sad - reads to me that it ends on a somewhat positive vein.
Negative	I'm afraid there is an environmental problem you should see about

For each pronoun we calculated the positive to negative frequency ratio and the negative to positive frequency ratio as follows: The ratio score was calculated by dividing the frequency of the pronoun in each class by its frequency in the other class. In

[1] http://csn.cancer.org

the case that a pronoun was observed only in a single class example, then it was considered as observed once in the second class, for calculating the ratio. Table 2 describes the frequency ratio of prominent pronouns in each dominant class.

Table 2. Example of prominent pronouns, by their frequency ratio in sentiment classes.

Pronoun	Positive frequency ratio	Pronoun	Negative frequency ratio
us	4.07	they	2.68
you	1.75	mine	2.43
she	1.75	who	1.54
we	1.71	I	1.45
her	1.59	me	1.41
he	1.32	my	1.23

Our goal was to classify the sentiment of each post into one of the two classes. The positive to negative frequency ratio was calculated for each term. To avoid the curse-of-dimensionality, which may lead to overfitting, instead of using each term as a feature, for every class we compute three features: The first is the maximum positive to negative frequency ratio among the words in the post. The second is the sum of the highest three ratios, and the third is the sum of the highest six. Similarly, to compute the other three features, the maximum is taken on the values of negative to positive frequency ratios. For example, the first post in Table 1 is taken. The first feature to correspond with a negative class is set to 1.45 if the term 'I' has the maximum negative to positive frequencies ratio, among all terms in the post. We trained a classification model by using all terms, and a second model by using all terms after pronouns were removed. A 10-fold cross validation set of experiments was conducted.

3.1 Results

Table 3 shows classification results of four different classifiers, which were found adequate for the task. In all the experimented classifiers, there is a drop in performance when the pronouns are not participating as predictors by most measurements, i.e., the results are better by accounting pronouns as well. This finding shows that pronouns play a role in sentiment classification. A two-way ANOVA (algorithm, pronouns) revealed a significant effect of pronouns to sentiment classification on the AUC ($p = 0.074$) and Accuracy ($p=0.08$), but not for the F-Measure ($p=0.43$). Figure 1 shows a comparison between pronoun frequencies in general English, as reported by [1], and in the Cancer Survivors Network.

4. DISCUSSION AND CONCLUSIONS

This first work discusses the community behavior in terms of pronouns usage. The importance of this work is in demonstrating that the text styling, in terms of pronouns usage, is useful for some text analyses, which relates to emotional states.

Table 3. Drop in sentiment classification results, when pronouns are omitted from the text. The drop is in percentage from the original performance.

Classifier	Drop in Accuracy	Drop in F-measure	Drop in AUC
adaboost_LMT	-0.01	**0.14**	**5.84**
J48	**0.94**	-0.84	**3.37**
Logistic Regression	**1.84**	**1.51**	**0.21**
Rotation Forest	**1.41**	**0.94**	**3.02**

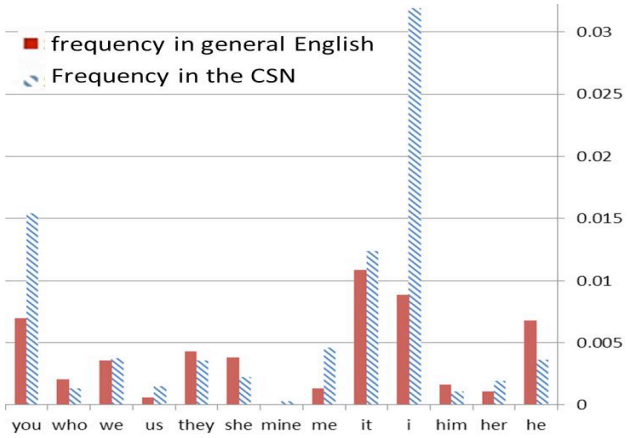

Figure 1. Frequency of pronouns in general English and in the Cancer Survivors Network (CSN).

However, while our findings provide some insights on whether pronouns are important to the description of personality or emotional state, in terms of sentiment, they also raise some questions regarding the psychological meanings and dimensions of the findings.

ACKNOWLEDGMENTS

We are grateful to the American Cancer Society (ACS) for making the Cancer Survivors' Network data available to us. We also want to thank Ken Portier and Greta Greer from ACS for their support and discussions related to this work.

5. REFERENCES

[1] Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2013, May). Steering user behavior with badges. In Proceedings of the 22nd international conference on World Wide Web (pp. 95-106). International World Wide Web Conferences Steering Committee.

[2] Chung, C. K., & Pennebaker, J. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. Journal of Research in Personality, 42, 96-132.

[3] Davison,K.E, & Pennebaker,J.W.(1997).Virtual narratives: Illness representations in online support groups. In K.J. Petrie & J.A. Weinman (Eds.), Perceptions of health and illness: Current research and applications (pp.463-486).Singapore:Harwood.

[4] Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In LREC.

[5] B. Qiu, K. Zhao, P. Mitra, D. Wu, C. Caragea, J. Yen, G. E. Greer, K. Portier. (2011) Get Online Support, Feel Better--Sentiment Analysis and Dynamics in an Online Cancer Survivor Community. In: SocialCom 2011.