Improving the Classification of Newsgroup Messages through Social Network Analysis

Blaz Fortuna
Dept. Knowledge Technologies
Institute Jožef Stefan, Jamova 39
1000 Ljubljana, Slovenija
+386 (01) 477 39 00
blaz.fortuna@ijs.si

Eduarda Mendes Rodrigues
Microsoft Research Ltd
7 JJ Thomson Avenue
Cambridge, CB3 0FB, UK
+44 (0) 1223 479 700
eduarda@microsoft.com

Natasa Milic-Frayling
Microsoft Research Ltd
7 JJ Thomson Avenue
Cambridge, CB3 0FB, UK
+44 (0) 1223 479 700
natasamf@microsoft.com

ABSTRACT

Newsgroup participants interact with their communities through conversation threads. They may respond to a message to answer a question, debate a topic, support or disagree with another person's point, or digress and write about a different subject. Understanding the structure of threads and the sentiment of the participants' interaction is valuable for search and moderation of newsgroups.

In this paper, we focus on automatic classification of message replies into several types. For representing messages we consider rich feature sets that combine the standard author reply-to network properties with features derived from four additional structures identified in the data: 1) a network of authors who participate in the same threads, 2) network of authors who post similar content, 3) network of threads sharing common authors, and 4) network of content-related threads.

For selected newsgroups we train linear SVM classifiers to identify agreement and disagreement with the original message, and question and answer patterns in the threads. We show that the use of newly defined features substantially improves classification of messages in comparison with the SVM model based only on the standard reply-to network.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *abstracting methods, indexing methods.*
I.5.2 **[Pattern Recognition]**: Design Methodology – *feature evaluation and selection.*

General Terms: Algorithms, Theory, Human Factors.

Keywords: Message classification, social networks, newsgroups, communities.

1. INTRODUCTION

Newsgroup communities have been around since the early days of the Internet. They are formed around a variety of topics and users are evermore turning to them to share their opinions, find specific information, seek community support, provide answers, and a like.

A large percentage of people in online communities (often over 90% [12]) are 'lurkers' who search and read available content but rarely communicate with others [12,14]. They typically look for detailed answers that can hardly be found elsewhere on the Web. Thus, providing effective support for search and browsing through message threads is of great value to the users. In particular, for finding information it is helpful to understand the structure of a discussion thread and quickly zoom onto the 'answer' messages. For those joining in a long discussion it is useful to get a sense of the dynamics and agreement level among participants. In order to devise optimal algorithms and user interfaces for accessing newsgroup content, we thus need to understand the characteristics of messages and message threads, including the structure, type, and sentiment of the participants' interaction.

In this paper we present a novel work on automatic identification of response types in message threads: agreement or disagreement with the previous message, insult, question, and answer. To the best of our knowledge, newsgroup message classification without constraints on the topicality or social structure of the user community has not been attempted in previous research. The problem is a challenging one because topics of distinct threads may vary considerably even within a single newsgroup. At the same time, it is a not clear whether replies to messages may exhibit sufficient commonality across topics to make their type recognizable and automatically detectable.

Our work shows that we can create robust classifiers for several types of replies regardless of the topics they cover. We achieve this by exploring a rich set of features from newly defined *author* and *thread networks*. We demonstrate a significant improvement over the baseline classifier that uses only the *reply-to network* model. In the following sections we review related research (§2), provide a detailed description of our experimental work (§§3-5), including the characteristics of the data set and feature sets and the study findings (§6). Finally, we discuss application areas and future work (§7).

2. RELATED RESEARCH

Discussion groups, blogs, online product reviews, and other community-generated content are rich sources of users' sentiment and opinion and have been a subject of a considerable body of research on opinion polarity and sentiment analysis. Techniques that have been used include text classification methods [1,13], linguistic analysis [4,10], and social network analysis [1,16]. The reply-to network, in particular, has been considered for identifying topical

CIKM'07, November 6--8, 2007, Lisboa, Portugal.

polarity of newsgroup participants [1, 8] and categorizing newsgroup types and author roles [6].

Based on the hypothesis that a message response is most likely to disagree with the parent message, Agrawal *et al.* [1] applied constrained and unconstrained graph partitioning techniques to cluster authors who share similar opinions into two opposing camps. Kelly *et al.* [8] clustered participants with similar opinions within a newsgroup and found that, regardless of the underlying distribution of participants into the clusters, the ratio of messages on each side of the discussion is balanced. Indeed, the traffic of the minority opinion was found to be larger in order to make up for the smaller number of people.

Fisher *et al.* [6] used graph properties of the reply-to network to study social aspects of online communities. They categorized user roles and newsgroup types through analysis of in- and out-link degree distributions of key authors. Motivated by the observations of this study we consider multiple representations of the social network for message classification, including the *reply-to* network.

Our research differs from the previous work. Instead of studying the opinion of a newsgroup community on a single topic we analyze the structure of discussion threads and types of message replies across topics. We expect that the message thread properties reflect the polarity of the community but we do not assume that individual participants always express a strong and consistent opinion about a topic.

3. LEARNING METHOD

Support Vector Machine (SVM) [3] has gained a wide popularity as one of the state-of-the-art machine learning methods for tasks such as classification and regression. Among specific applications, it has proven to perform particularly well for text categorization [7] and sentiment analysis of movie reviews posted to newsgroups [13]. We apply linear SVM classifiers to:

1) Predict the agreement level between a message and its parent message within discussion threads. Messages are classified as 'agree', 'disagree', or 'insult'.
2) Identify questions and answers to a parent message within technical discussion threads. Messages are classified as 'question' or 'answer'.

We describe each message-parent pair by a vector of features and we use *one-vs-all* multi-class approach for classifying unseen message pairs (details provided in §6). We use the linear SVM implementation included in the Text Garden library [15].

4. EXPERIMENTAL SETUP

4.1 Data Sets

Our data set consists of message threads and header information from four Usenet newsgroups. The first two newsgroups, *alt.politics.immigration* and *talk.politics.guns* host mostly political discussions and debates. These same groups were used in [1] for social network analysis. The other two groups *microsoft public. internetexplorer.general* and *microsoft.public.windowsxp.general*, host mostly Q&A-type threads. Table 1 shows summary information about these data sets, hereafter referred to as *immigration*, *guns*, *iexplorer* and *winxp.* It lists the total number of threads, messages, replies and authors per newsgroup. It also indicates the period of time in which all messages were collected.

Table 1. Description of the newsgroup data sets.

Newsgroup	Threads	Messages	Replies	Authors	Collection Period
immigration	1,367	10,095	8,728	463	Aug 31 to
guns	874	6,776	5,902	844	Oct 19'06
iexplorer	3,631	10,934	7,303	3,443	Jul 19 to
winxp	10,280	42,052	31,772	8,145	Oct 19'06

Table 2. Message judgments.

Label	Description
agree	Message agrees with the point of view of the parent message. Adding clarifications or extra info also counts.
disagree	Message disagrees with the point of view of the parent message. Sarcastic comments also count.
insult	Author of the message is purely insulting the author parent message. Insults replying to insults are *disagree* messages.
question	Message is a question or a clarification of a previously asked question by the same author.
answer	Message is an answer to a question in the parent message or a request for further information about the question.
off-topic	The message has no connection to the parent message and is not a question message.
don't know	If none of the above labels apply.

Table 3. Number of labeled messages in the training sets.

Label	Newsgroup			
	immigration	*guns*	*iexplorer*	*winxp*
agree	179	42	20	12
disagree	284	128	21	28
insult	37	32	-	-
question	30	36	45	47
answer	17	22	79	147
off-topic	65	15	2	8
don't know	121	23	10	11
Total	733	298	177	253

4.2 Training Data

We created training data sets from several samples of threads randomly selected from each newsgroup. The sample messages were annotated by experts with one of the labels listed in Table 2. When two or more labels were considered valid, the experts were asked to select the label that applied the most, or in case of ambiguity, to annotate as "don't know". Table 3 shows the total number of messages annotated with each label.

5. FEATURE SETS

Our aim was to classify messages posted to two classes of groups, political discussions and Q&A, and to investigate the impact of particular features on the classifiers performance. Thus, we considered a varied set of features both of structural and content nature. For content analysis, we cleaned each message to remove headers and any quoted text from parent messages. We derived features from multiple networks: 3 author networks and 2 thread networks. Past research has used thread-level message features for analysis of newsgroup data [2,5,6,17]. We also ran experiments with such kind of features, but our results did not show much improvement. Thus, in this and subsequent sections we will only refer to the multi-network features.

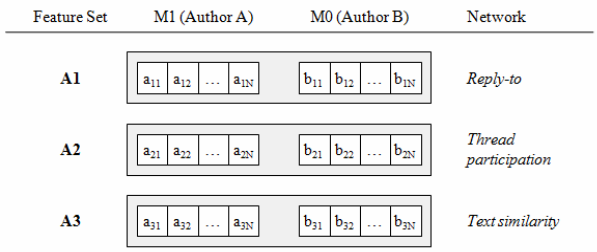

Figure 1. Feature sets extracted from the author networks.

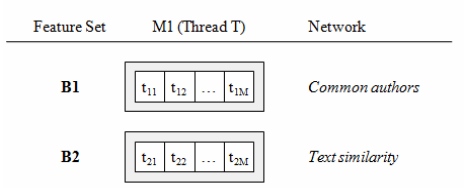

Figure 2. Feature sets extracted from the thread networks.

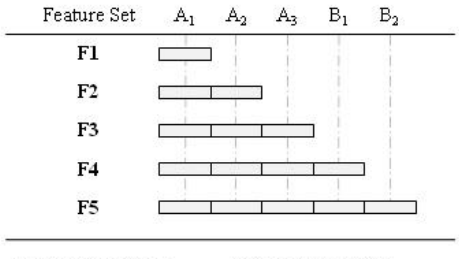

Figure 3. Combinations of feature sets.

5.1 Author Networks

We created 3 types of author networks for each newsgroup: *reply-to*, *thread participation*, and *text similarity*. In all of these, the nodes represent authors, but the edges carry distinct semantics:

- A *reply-to network* edge from author A to author B indicates that A has replied to at least one message posted by B.
- A *thread participation network* edge from author A to author B indicates that both authors have participated in the same thread in at least *k* occasions; we chose *k=5* for this study.
- A *text similarity network* edge indicates similarity between the content of connected authors' messages. The messages of each author were represented by a single centroid keyword vector and *author-author* edges were created for centroid cosine similarities of at least η; we set $\eta=0.3$ for this study.

We described each *message reply* by a vector of features extracted from the three author networks, which we will refer to as **A1**, **A2** and **A3**. Given a message, **M1**, and the message it replies to, **M0**, 3 feature vectors were created for M1 and another 3 for M0 (see Figure 1). Individual features of each vector are associated with nodes in the networks, i.e., authors. The values of vector components $\{a_{i1},a_{i2},\ldots,a_{iN}\}$ for the author of the message M1 are 0, unless they correspond to authors who are direct or 2nd-level neighbors of M1's author, in which case the values are set to 1 and 0.5, respectively. A similar author node vector was created for the author of the parent message M0. The final feature set for a *message reply* concatenated the two vectors.

5.2 Thread Networks

In [1], it was assumed that, for a broad topic, people who engage in political discussions usually reply to present an opposing view. Our work, however, makes no assumptions on the consistency of an author's view on a topic. Thus, we acknowledge that people's opinions across or even within the broad topic are not always on the same side of the argument.

However, the very involvement of the author in the discussion thread may reveal information about the topic that is discussed. Thus, we looked for associations of topics and, to that effect, generated two types of thread networks for each newsgroup: *common authors network* and *text similarity network*. The nodes of both networks represent threads but the edges have a different meaning in each case:

- A *common authors network* edge between thread T and Q indicates "thread T has at least *m* authors in common with thread Q". In this study we set *m=3*.
- A *text similarity network* edge between thread T and Q indicates similarity between the content of their messages. The cosine similarity between centroid keyword vectors is used. An edge between thread T and Q indicates a similarity of at least η; we set $\eta=0.3$ in this study.

We described each *thread* by a vector of features extracted from the two thread networks, referred to as **B1** and **B2**, respectively. Given a message **M1** belonging to the thread **T**, we created two feature vectors, where individual components were associated with other nodes the networks, i.e. threads – see Figure 2. The values of vector components $\{t_{i1},t_{i2},\ldots,t_{iM}\}$ for thread T_i are 0, unless the respective thread node is a direct or 2nd-level neighbor of thread node T_i in the network, in which case the values are set to 1 and 0.5, respectively.

6. EXPERIMENTS AND RESULTS

We conducted a comprehensive set of experiments with linear SVM classifiers to investigate the effectiveness of individual feature sets and their combinations in:

1) Predicting the level of agreement of messages posted to political discussion newsgroups.
2) Identifying question and answer messages in technical discussion newsgroups.

For evaluation we used 10-fold cross-validation: the data was randomly split into 10 folds of equal size and in turn, the classifier was trained on 9 folds and evaluated on the remaining fold. The results were averaged over the 10 iterations. The performance of the classifier was measured based on the *break-even-point* (BEP) from the ranked list of messages scored by the classifier. The BEP value is associated with the rank at which classification precision and recall are equal.

Selected experiments are outlined in Figure 3. Sets **F1** to **F3** consist of features extracted from the author networks (A1-A3). Features sets **F4** and **F5** additionally include features from the thread networks (B1 and B2). In §6.1 and §6.2, we present the experimental results for the two newsgroups types present in our data set: discussion and Q&A. In §6.3, we discuss briefly the classifiers' performance. Given that *reply-to network* features have been extensively used in previous work [1,6,8], we use **F1** as baseline for our analysis.

6.1 Classification of Discussion newsgroups

To predict the level of agreement between a message and its parent message in discussion threads, we only used the relevant training data, i.e. we used messages labeled as 'agree', 'disagree' or 'insult'. The classification results for the various combinations of feature sets (F1-F5) are shown in Table 4. In this type of groups, an increased performance due to thread network features was particularly evident in the 'insult' class, where such messages seem

to be strongly predicted through the co-participation in threads (B1): increase from 68% to 74% for *guns* and from 38% to 85% for *immigration*. Using threads text similarity features (B2) gave further boost to the *guns* category: from 74% to 81%.

6.2 Classification of Q&A newsgroups

To identify questions and answers in technical newsgroups, we only used the relevant training data: messages labeled as 'question' or 'answer'. The classification results are shown in Table 5. Unlike the previous case, features derived from the thread networks did not lead to better classification performance. Connections among authors that participated in the same threads (A2) were particularly beneficial to predict 'questions': increase of 59% to 64% for *iexplorer* and 78% to 80% for *winxp*. Content-based author similarity features (A3) improved the prediction of 'answers' for *iexplorer*: from 71% to 75%.

6.3 Discussion

We observed that the co-participation of authors across threads (feature set B1) was a particularly relevant factor for improving the classification of messages of discussion threads. Text similarity features further improved classification. These results hint that authors seem to be consistent in their opinions, when participating with the same authors on multiple discussion threads. Fisher *et al.* [6] observed that participants of political discussion groups tend to form a fairly closed community, responding to each other often and mostly ignoring people who are not in core participants. They also observed that, the mostly connected participants of technical newsgroups, on the contrary, respond essentially to the outsiders, whose messages are generally questions. Our results indirectly support this analysis, since thread network features did not enhance the classification performance in the Q&A case.

7. CONCLUSIONS

In this paper we demonstrated that it is possible to train a robust message classifier to automatically detect messages of selected response types, including agreement and disagreement among discussion participants. This is a rather significant result from both research and application point of view. Previous research tried to limit the topicality of messages in order to perform classification [1]. We have shown that with well selected author and thread network features we can achieve very good classification results (measured through BEP) for any topic that participants of a newsgroup may be discussing. The results clearly demonstrate the superiority of the thread network features over the standard reply-to network alone.

Our results open doors to various applications that can benefit from deeper understanding of newsgroup participants interaction. For example, client applications could use our message classifier to identify and label insulting messages. Verbal attacks can be derived from social context of the messages' authors and the relationship between discussion topics. Our findings offer also the foundation for the design of ranking functions for newsgroup search that take into account the types of messages, given a search goal, such as, finding answers to a question, finding a similar question, or finding strong positive and negative opinions about a topic.

Table 4. Classification results for discussion groups.

Features	*guns*			*immigration*		
	agree	*disagree*	*insult*	*agree*	*disagree*	*insult*
F1	61%	80%	62%	65%	75%	37%
F2	69%	82%	72%	66%	76%	45%
F3	65%	84%	68%	68%	77%	38%
F4	67%	86%	74%	73%	80%	85%
F5	66%	85%	81%	72%	80%	85%

Table 5. Classification results for Q&A groups.

Features	*iexplorer*		*winxp*	
	answer	*question*	*answer*	*question*
F1	70%	59%	93%	78%
F2	71%	64%	94%	80%
F3	75%	66%	94%	79%
F4	75%	66%	94%	79%
F5	75%	65%	94%	77%

8. REFERENCES

[1] Agrawal, R., Rajagopalan, S., Srikant, R., Xu, Y., "Mining Newsgroups Using Networks Arising from Social Behavior," *Proc. of* WWW'03, pp. 529-535, 2003.

[2] Borgs, C., Chayes, J., Mahdian, M. and Saberi, A., "Exploring the Community Structure of Newsgroups," In: *Proc. of* KDD'04, 2004.

[3] Cortes, C. and Vapnik, V. Support Vector Networks. *Machine Learning*, vol. 20, pp. 273–297, 1995.

[4] Glance, N., Hurst, M., Nigam, K., Siegler, M., Stockton, R., and Tomokiyo, T., "Deriving Marketing Intelligence from Online Discussion," *Proc. of* KDD'05, pp. 419-428, 2005.

[5] Fiore, A., Teirnan, S.L., Smith, M. "Observed Behavior and Perceived Value of Authors in Usenet Newsgroups: Bridging the Gap," *Proc. of* CHI'02, pp. 323-330, 2002.

[6] D. Fisher, M. Smith, H. Welser, "You Are Who You Talk To: Detecting Roles in Usenet Newsgroups," *Proc. of the 39th* HICSS, 2006.

[7] Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *Proc. of* ECML'98, Nédellec, C. and Rouveirol, C. (Eds.), *Lecture Notes in Computer Science*, vol. 1398, pp. 137-142, Springer-Verlag, 1998.

[8] Kelly, J.W., Fisher, D., and Smith, M., "Friends, Foes, and Fringe: Norms and Structure in Political Discussion Networks," *Proc. of the Int. Conf. on Digital Government Research '06*, pp. 412-417, 2006.

[9] Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R., and Riedl, J., "Applying Collaborative Filtering to Usenet News," *Communications of the* ACM, vol. 40, no. 3, pp. 77-87, 1997.

[10] B. Liu, M. Hu, , J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," *Proc. of* WWW'05, pp.342-351, 2005

[11] Netscan Microsoft Research: http://netscan.research.microsoft.com.

[12] Nonnecke, B. and Preece, J., "Lurker demographics: counting the silent," *Proc. of* CHI '00, pp.73-80, Apr 01-06, 2000.

[13] Pang, B., Lee, L. and Vaithyanathan, S., "Thumbs up? Sentiment Classification using Machine Learning Techniques," *Proc. of* EMNLP'02, pp. 79-86, 2002.

[14] Soroka, V. and Rafaeli, S., "Invisible participants: how cultural capital relates to lurking behavior," *Proc. of* WWW'06, pp. 163-172, May 23-26, 2006.

[15] *Text Garden* - Available at: http://www.textmining.net/

[16] Tuulos, V. and Tirri, H., "Combining topic models and social networks for chat data mining," *Proc. of the* IEEE/WIC/ACM WI'04, pp. 206–213, 2004.

[17] W. Xi, J. Lind and E. Brill, "Learning effective ranking functions for newsgroup search," *Proc. of SIGIR'04*, pp 394-401, 2004.

[18] B. Zhang, H. Li, Y. Liu, W. Xi, W. Fan, Z. Chen, W.Y. Ma, "Improving Web Search Results using Affinity Graph," *Proc. of* SIGIR'05, pp 504-511, 2005.