Classifying User Forum Participants: Separating the Gurus from the Hacks, and Other Tales of the Internet

Marco Lui and Timothy Baldwin
NICTA VRL
Department of Computer Science and Software Engineering
University of Melbourne, VIC 3010, Australia
saffsd@gmail.com, tb@ldwin.net

Abstract

This paper introduces a novel user classification task in the context of web user forums. We present a definition of four basic user characteristics and an annotated dataset. We outline a series of approaches for predicting user characteristics, utilising aggregated post features and user/thread network analysis in a supervised learning context. Using the proposed feature sets, we achieve results above both a naive baseline and a bag-of-words approach, for all four of our basic user characteristics. In all cases, our best-performing classifier is statistically indistinct from an upper bound based on the inter-annotator agreement for the task.

1 Introduction

The most natural form of communication is through dialogue, and in the Internet age this manifests itself via modalities such as forums and mailing lists. What these systems have in common is that they are a textual representation of a *threaded discourse*. The Internet is full of communities which engage in innumerable discourses, generating massive quantities of data in the process. This data is rich in information, and with the help of computers we are able to archive it, index it, query it and retrieve it. In theory, this would allow people to take a question to an online community, search its archives for the same or similar questions, follow up on the contents of prior discussion and find an answer. However, in practice, search forum accessibility tends to be limited at best, prompting recent interest in information access for user forums (Cong et al., 2008; Elsas and Carbonell, 2009; Seo et al., 2009).

One problem with current approaches to accessing forum data is that they tend not to take into account the structure of the discourse itself, or other characteristics of the forum or forum participants. The bag-of-words (BOW) model common in information retrieval (IR) and text categorisation discards all contextual information. However, even in IR it has long been known that much more information than simple term occurrence is available. In the modern era of web search, for example, extensive use is made of link structure (Brin and Page, 1998), anchor text, document zones, and a plethora of other document (and query, click stream and user) features (Manning et al., 2008).

The natural question to ask at this point is, *What additional structure can we extract from web forum data?* Previous work has been done in extracting useful information from various dimensions of web forums, such as the post-level structure (Kim et al., 2010). One dimension that has received relatively little attention is how we can use information about the identity of the participants to extract useful information from a web forum. In this work we will examine how we can utilize such *user-level* structure to improve performance over a user classification task.

We have used the term *threaded discourse* to describe online data that represents a record of messages exchanged between a group of participants. In this work, we examine data from LinuxQuestions, a popular Internet forum for Linux-related troubleshooting. Aside from a limited set of features specific to the Linux-related troubleshooting domain, however, our techniques are domain-inspecific and expected to generalize to any data that can be interpreted as a threaded discourse.

This work is part of ILIAD (Baldwin et al., 2010), an ongoing effort to improve information access in linux forums. Our contribution to the project is techniques to identify characteristics of forum users, building on earlier work in the space (Lui, 2009). The problem that we face here is two-fold: Firstly, there is no established ontology for

Marco Lui and Timothy Baldwin. 2010. Classifying User Forum Participants: Separating the Gurus from the Hacks, and Other Tales of the Internet. In *Proceedings of Australasian Language Technology Association Workshop*, pages 49–57

characteristics of forum users. To address this, we have designed a set of attributes that we expect to be helpful in improving information access over forum data. Secondly, in order to exploit user characteristics we would need to evaluate a large number of users. This quantity of data would be much too large to be processed manually. We therefore apply supervised machine learning techniques to allow us to effectively discover the characteristics of a large number of forum users in an automated fashion.

2 Related Work

Lui and Baldwin (2009b) showed that user-level structure is useful in predicting percieved quality of forum posts. The data they evaluate over is extracted from Nabble, where the ratings provided by users are interpeted as the gold-standard for a correct classification. The task was originally proposed by Weimer et al. (2007) and further explored by Weimer and Gurevych (2007). In both cases, the authors focus on heuristic post-level features, which are used to predict perceived quality of posts using a supervised machine learning approach. Lui and Baldwin (2009b) showed that features based on user-level structure outperformed the benchmark set by Weimer and Gurevych (2007) on a closely-related task, by using user-level structure to inform a post-level classification task. We build on this work by utilizing the user-level structure to perform our novel user-level classification task.

In work on thread classification, Baldwin et al. (2007) attempted to classify forum threads scraped from Linux-related newsgroups according to three attributes: (1) Task Oriented: *is the thread about a specific problem?*; (2) Complete: *is the problem described in adequate detail?*; and (3) Solved: *has a solution been provided?* They manually annotated a set of 250 threads for these attributes, and extracted a set of features to describe each thread based on the aggregation of features from posts in different sections of the thread. We provide a novel extension of this concept, whereby we aggregate posts from a given user.

Wanas et al. (2008) develop a set of post-level features for a classification task involving post and rating data from Slashdot. Their task involves classifying posts into one of three quality levels (High, Medium or Low), where the gold-standard is provided by user annotations from the forum. This is conceptually very similar to our task, and we build on this feature set.

Extracting community structure from networks can yield insights into the relationships between users in a forum (Newman and Girvan, 2004; Drineas et al., 2004; Chapanond et al., 2005), and could in turn aid in engineering descriptions of the users more suited to a particular task. Agrawal et al. (2003) describe a technique for partitioning the users in an online community based on their opinion on a given topic. They find that basic text classification techniques are unable to do better than the majority-class baseline for this particular task. They then describe a technique based on modeling the community as a *reply-to* network, with users as individual nodes, and edges indicating that a user has replied to a post by another user; using this representation, they are able to do much better than the baseline. Fortuna et al. (2007) build on this work, defining additional classes of networks that represent some of the relationships present in an online community. Part of our feature set is derived from modelling Internet forum users on the basis of the interactions that exist between them, such as a tendency to reply to each other or to co-participate in threads. We extend the social network analysis of Agrawal et al. (2003) and Fortuna et al. (2007) to generate user-level features.

Malouf and Mullen (2008) present the task of determining the political leaning of users on a U.S. political discussion site. They apply network analysis to the task, based on the observation that users tend to quote users of opposing political leaning more than they quote those of similar political leaning. They found that standard text categorisation methods performed poorly over their task, and that the results were improved significantly by incorporating network-derived features.

In a similar vein, Carvalho et al. (2007) used a combination of textual features (in the form of "email speech acts") and network-based features to learn which users were team leaders. They found that the network-based features enhanced classification accuracy.

Sentiment analysis (Pang and Lee, 2008) relates to this work as one of our user characteristics (POSITIVITY) is an expression of user sentiment. However, sentiment analysis has tended to focus on individual documents, and rarely takes into account the author. An exception to this is the work of Thomas et al. (2006), who attempted to predict

which way each speaker in a U.S. Congressional debate on a proposed bill voted, on the basis of both what was said and the indication of agreement between speakers. Their task is related to ours in that it involves a user-level classification, but it focused on extracting information identifying where the speakers agree and disagree.

Expert finding is the task of ranking experts relative to each of a series of queries, and has been part of the TREC Enterprise Track (Craswell et al., 2005; Soboroff et al., 2006; Balog et al., 2006; Fang and Zhai, 2007). The challenge is to estimate the likelihood of a given individual being an expert on a particular topic, on the basis of a document collection. There is certainly scope to evaluate the utility of the user characteristics proposed in this research in the context of the TREC expert finding task, although only a small fraction of the document collection (the mailing list archives) has the threaded structure requisite for our methods, and our focus is on the general characteristics of the user rather than their topic-specific expertise.

3 User Characteristics

We have designed a set of user-level attributes which we expect to be useful in improving information access over forum data. The attributes were selected based on our personal experiences in interacting with online communities. In this, we sought to capture the attributes of users who provide meaningful contributions, as follows:

CLARITY: How clear is what the user meant in each of their posts, in the broader context of the thread?

PROFICIENCY: What level of perceived technical competence does the user have in their posts?

POSITIVITY: How positive is the user in their posts?

EFFORT: How much effort does the user put into their posts?

Each user-level attribute is quantified by way of a 5 point ordinal scale, as detailed in Table 1.

While we have described the four attributes as if they were orthogonal to each other, in reality there are obvious overlaps. For example, high clarity often implies high effort, but the reverse is not necessarily true. For simplicity, we do not consider the interactions between the characteristics in this work, leaving it as a possibility for further research.

4 Dataset

We created a new dataset specifically for this work based on data crawled from LinuxQuestions,[1] a popular Internet forum for Linux troubleshooting. From this forum, we scraped a background collection of 34157 threads, spanning 126094 posts by 25361 users.

In order to evaluate how well we can automatically rate forum users in each of our four user characteristics (from Section 3), we randomly selected 50 users who had each participated in more than 15 different threads in the full dataset. We asked four independent annotators to annotate the 50 users over each of the 4 attributes. The annotators all had a computer science background, and had participated in Linux-related online communities. For each attribute, the annotators were asked to choose a rating on a five-point scale, based on the description of user attributes from Section 3.

For each of the 50 users, we randomly selected 15 threads that they had participated in, and partitioned these into 5 separate annotation instances as follows: for the first instance, we selected 1 thread; for the second instance we selected 2 threads; and so on, giving us 5 instances, each with 1 to 5 threads. This gave us a total of 250 annotation instances (with 5 instances per user). We chose to annotate each user multiple times in order to build a more complete picture of the user. Each instance presented a different number of threads to the annotator, in order to give the annotators maximal context in annotating a user while still minimizing the number of threads we required the user to have participated in.

Each annotator was asked to rate all 250 annotation instances, meaning that they actually saw each of the 50 users a total of five times each. Annotators were not alerted to the fact that they would annotate each user five times, and all usernames were removed from the threads before being displayed to the annotator. However, for a given annotation instance, the annotator was alerted to which posts the user being annotated had authored. The posts of other users in those threads where also presented to provide the full thread context, but the annotators were instructed to use those posts only to interpret the posts of the user in question.

Since each annotator annotated each user 5 times for each attribute, we compute a score for each user–annotator–attribute combination, which

[1] `http://www.linuxquestions.org`

Attribute	*Value*		*Description*
CLARITY	1	**Unintelligible**	It is impossible to make sense of the user's posts; clear as mud!
	2	**Somewhat confused**	The meaning of the user's posts is ambiguous or open to interpretation
	3	**Comprehensible**	With some effort, it is possible to understand the meaning of the post
	4	**Reasonably clear**	You occasionally question the meaning of the user's posts
	5	**Very clear**	Meaning is always immediately obvious relative to the thread; sparkling clarity!
PROFICIENCY	1	**Hack**	The posts of this user make it patently obvious that they have no technical knowledge relevant to the threads they participate in; get off the forum!
	2	**Newbie**	Has limited understanding of the very basics, but nothing more
	3	**Average**	Usually able to make a meaningful technical contribution, but struggles with more difficult/specialized problems
	4	**Veteran**	User gives the impression of knowing what they are talking about, with good insights into the topic of the thread but also some gaps in their knowledge
	5	**Guru**	The posts of this user inspire supreme confidence, and leave the reader with a warm, fuzzy feeling!
POSITIVITY	1	**Demon**	Deliberately and systematically negative with no positive contribution; the prince/princess of evil!
	2	**Snark**	The user is somewhat hurtful in their posts
	3	**Dull**	The user's posts express no strong sentiment
	4	**Jolly**	The user's posts are generally pleasant
	5	**Solar**	Goes out of his/her way in trying to make a positive contribution in all possible ways; positively radiant!
EFFORT	1	**Loser**	Zero effort on the part of the user
	2	**Slacker**	Obvious deficiency in effort
	3	**Plodder**	User's posts are unremarkable in terms of the effort put in
	4	**Strider**	Puts obvious effort into their post
	5	**Turbo**	Goes out of his/her way in trying to make a contribution; an eager beaver!

Table 1: A detailed description of the user-level attribute values

is simply the sum across the 5 annotations. Using this score, we then rank the users for each pairing of annotator–quality.

We formulated the user-level classification task as four separate classification tasks, across the four attributes. In order to account for subtle variance in annotators' interpretations of the ordinal scale, we took a non-parametric approach to the data: we pooled all of the annotator ratings and established a single ranking over all the annotated users for each attribute. We then discretized this ranking into 5 equal-sized bins, in order to provide a more coarse-grained view of the relative ordering between users. Therefore, our task can be interpreted as assigning each user to their corresponding uniformly-distributed quintile on each attribute.

4.1 Inter-annotator Agreement

We calculate inter-annotator agreement on each of the four attributes via leave-one-out cross-validation. For each user-annotator-attribute combination, we calculate two scores: the sum of ratings given by the annotator being considered, and the sum of ratings given by all the other annotators. For each of the four attributes, we rank the users based on each of these two scores, and compute Kendall's τ (Kendall, 1938) between the two ranklists (Table 2), as well as the p-value for the significance of the τ value.

Attribute	*Annotator*	τ	p
Clarity	Annotator 1	0.235	0.016
	Annotator 2	0.221	0.024
	Annotator 3	0.292	0.003
	Annotator 4	0.307	0.002
Effort	Annotator 1	0.517	0.000
	Annotator 2	0.707	0.000
	Annotator 3	0.682	0.000
	Annotator 4	0.610	0.000
Proficiency	Annotator 1	0.582	0.000
	Annotator 2	0.460	0.000
	Annotator 3	0.536	0.000
	Annotator 4	0.407	0.000
Positivity	Annotator 1	0.009	0.924
	Annotator 2	0.434	0.000
	Annotator 3	0.473	0.000
	Annotator 4	0.436	0.000

Table 2: Inter-annotator agreement, based on Kendall's τ and associated p-value

We see that for all attributes, there is a statistically significant correlation between the annotations. This correlation is strongest in the EFFORT and PROFICIENCY attributes, and weakest in the

CLARITY attribute. This is partly to be expected, since CLARITY is more subjective than EFFORT or PROFICIENCY. POSITIVITY shows an interesting quirk, where the ratings from one annotator appear completely uncorrelated with those of all the others. This suggests that POSITIVITY as an attribute is slightly more subjective than the others.

5 Feature Extraction

We extract features for each user based on aggregating post-level features and via social network analysis.

5.1 Post-Aggregate Features

The most basic feature set we consider is a simple bag-of-words (BOW), computed as the sum of the bag-of-words model over each of the user's individual posts.

We also make use of two post-level feature sets from the literature on web user forum classification. The first is that of Baldwin et al. (2007) (BALDWIN^{Post}), and outlined in Table 3. It was designed to represent key posts in a thread for a thread-level classification (see Section 2) task. We compute this feature set for each of a user's posts.

The second is that of Wanas et al. (2008), and is described in Table 4. In this case, it was developed for a post-level classification task rating post quality, and thus lends itself readily to our post-aggregate user representation.

From each of BALDWIN^{Post} and WANAS, we derive a user-level feature set by finding the mean of each feature value over all of the user's posts in the full dataset. For boolean features, this can be directly interpreted as the proportion of the user's posts in which the feature is present. These feature sets are referred to as BALDWIN$^{Post}_{\text{AGG}}$ and WANAS$_{\text{AGG}}$ respectively.

Whereas it is possible for us to engineer a novel post-level feature set, our aim in this research is not to analyze the feature sets themselves, but rather to show that our techniques utilizing user-level structure perform better than techniques which ignore this information. We leave post-level feature engineering as an open avenue of further work.

5.2 Network Features

Fortuna et al. (2007) present a method of describing forum data using Social Network Analysis. The network is a graph representation of relationships within the forum. Building on Fortuna et al. (2007), we consider *User Networks*, where each node represents a user, and *Thread Networks*, where each node represents a thread. In this work, we consider two User Networks and one Thread Network, namely: (1) POSTAFTER, (2) THREADPART, and (3) COMMONAUTHORS, respectively. The networks we define build directly on work done by Fortuna et al. (2007), but the application to user-level feature extraction is novel.

Feature name	*Description*	*Type*
distribution	Mention of Linux distribution name?	Boolean
beginner	Mention "newbie" terms?	Boolean
emoticons	Presence of "smiley faces"?	Boolean
version numbers	Presence of version numbers?	Boolean
URLs	Presence of hyperlinks?	Boolean
words	Number of words in post	Integer
sentence	Number of sentences in post	Integer
question sentence	Number of questions in post	Integer
exclaim sentence	Number of exclamations in post	Integer
declarative sentence	Number of declarative sentences	Integer
other sentence	Number of other sentences	Integer

Table 3: The BALDWIN^{Post} feature set

Feature name	*Description*	*Type*
onTopic	Post's relevance to the topic of a thread	Real
overlapPrev	Post's largest overlap with a previous post	Real
overlapDist	Distance to previous overlapping post	Integer
timeliness	Ratio of time from prev post to average inter-post interval	Real
lengthiness	Ratio of post length to average post length in thread	Real
emoticons	Ratio of emoticons to sentences	Real
capitals	Ratio of capitals to sentences	Real
weblinks	Ratio of links to number of sentences	Real

Table 4: The WANAS feature set

POSTAFTER is modeled on the *reply-to* network described in Fortuna et al. (2007). Our data does not contain explicit annotation about the reply structure in a thread, so we approximate this information by the temporal relationship between posts. There exist more sophisticated approaches to the discovery of reply structure in a thread (Kim et al., 2010), and we consider integrating such methods to be an important avenue of further work.

POSTAFTER is parametrized with two values: $dist$ and $count$. Being a User Network, the nodes represent users. Two users $A1$ and $A2$ have a directed edge from $A1$ to $A2$ if and only if $A1$ submits a post to a thread that is within $dist$ posts after a post in the same thread by $A2$ on at least $count$ occasions. Note that this can occur more than once in a single thread. For our experiments, we used $dist = 1$ and $count = 3$.

THREADPART is implemented as described in Fortuna et al. (2007): nodes are again users, and each undirected edge indicates that two users have posted in the same thread on at least k occasions. Fortuna et al. (2007) set $k = 5$, but we only report on results for $k = 2$ and $k = 3$, as we found that for our dataset, the network is too sparse for higher values.

COMMONAUTHORS is also implemented as described in Fortuna et al. (2007): nodes are threads, and each undirected edge indicates that two threads have at least m users in common. We follow Fortuna et al. (2007) in setting $m = 3$.

In User Networks, the edges represent some relationship between users. From a User Network, we generate a feature vector v for each user. v is of length N, where N is the total number of nodes, or equivalently, the total number of users in the network. v has at least one feature set to 1, which corresponds to the user described by this feature vector, which we will hereafter refer to as the originator. Features representing users directly connected to the originator in the network receive a feature value of 1, and users that are second-level neighbours of the originator are set to a feature value of 0.5. All other values in v are set to 0.

For Thread Networks, edges represent relationships between threads. The method for computing a feature vector is similar to that for User Networks. The key difference is that in this instance, nodes represent threads and not users. Therefore, to describe a particular user, we consider threads that the user has posted in. We define a vector v of length T, where T is the total number of threads in the forum. Given the set S_0 of threads that the user has posted in, for each thread in S_0, we assign the value 1 to the feature in v corresponding to that thread. We then consider S_1, the set of immediate neighbours of S_0, and assign the value 1 to their corresponding features in v. Finally, we consider S_2, the immediate neighbours of S_1, and assign the value of 0.5 to their corresponding features. All other features are assigned the value 0.

6 Experimental Methodology

In all experiments, we build our classifiers using a support vector machine (`SVM`: Joachims (1998)), using `bsvm` (Hsu and Lin, 2006) with a linear kernel. For each combination of features, we evaluate it by carrying out 10-fold cross-validation. The partitioning is performed once and re-used for each pairing of learner and feature set.

Our experiments were performed using `hydrat` (Lui and Baldwin, 2009a), an open-source framework for comparing classification systems. `hydrat` provides facilities for managing and combining feature sets, setting up cross-validation tasks and automatically computing corresponding results. Features were extracted from the forum data using `forum_features`,[2] a Python module implementing a data model for forum data.

We evaluate our classifiers using microaveraged F-score ($\mathcal{F}_\mu$), reflecting the average performance *per-document*. As our classes are ordinal (representing quintiles of users), we additionally present results based on mean absolute error (MAE). MAE is the average absolute distance of the predicted ($Pred$) ordinal value from the gold-standard (G) value. It is a reflection of how far off the mark the average prediction is, with an MAE of 0 indicating perfect classifier performance.

As a baseline, we use a simple majority-class (`ZeroR`) classifier. A benchmark classifier is constructed based on a BOW feature set, as is the standard in text categorization. To derive an upper bound for the task, we perform leave-one-out cross-validation over our annotations, and calculate the mean F-score and MAE between each annotator and the combination of the remaining annotators.

When comparing a result to a baseline or a benchmark value, we also compute the p-value for a two-tailed paired t-test. In line with standard practice, we interpret $p < 0.05$ as statistically significant.

7 Results

First, we present results for each of the feature sets in isolation over the four user characteristics (Table 5). In each case, we present the results for the majority class (`ZeroR`) baseline and the bag-of-words (BOW) benchmark in the first two rows. Statistically-significant improvements over `ZeroR` (including BOW) are suffixed with "*", and statistically-significant improvements over BOW are suffixed with "$^+$". The best overall result for a given task achieved across all combinations of feature sets is presented in **boldface**, and is achieved for a single feature set

[2] `http://github.com/saffsd/forum_features`

Attribute	*Feature Set*	$\mathcal{F}_\mu$	*MAE*
CLARITY	ZeroR	0.020	2.040
	BOW	0.120	1.620*
	WANAS$_{\text{AGG}}$	0.100	1.760
	BALDWIN$^{Post}_{\text{AGG}}$	0.120	1.860
	THREADPART$_2$	0.240*	1.540*
	THREADPART$_3$	**0.260**$^{*+}$	**1.360***
	POSTAFTER	0.220*	1.440*
PROFICIENCY	ZeroR	0.000	1.980
	BOW	**0.240***	**1.380***
	WANAS$_{\text{AGG}}$	0.000	2.080
	BALDWIN$^{Post}_{\text{AGG}}$	0.060	1.740
	THREADPART$_2$	0.180*	1.820
	THREADPART$_3$	0.120*	1.700
	POSTAFTER	0.200*	1.900
POSITIVITY	ZeroR	0.040	1.880
	BOW	0.140	1.660
	WANAS$_{\text{AGG}}$	0.120	1.680
	BALDWIN$^{Post}_{\text{AGG}}$	0.120	1.580
	THREADPART$_2$	0.180*	1.720
	THREADPART$_3$	0.120	1.760
	POSTAFTER	**0.220** *	**1.340***
EFFORT	ZeroR	0.000	1.760
	BOW	**0.320***	**1.240***
	WANAS$_{\text{AGG}}$	0.100	1.600
	BALDWIN$^{Post}_{\text{AGG}}$	0.300*	1.420
	THREADPART$_2$	0.180*	1.700
	THREADPART$_3$	0.140*	1.700*
	POSTAFTER	0.100*	1.900

Table 5: Results for individual feature sets.

in the case of CLARITY and POSITIVITY, both using User Network feature sets.

The benchmark results (BOW) are considerably more impressive than the ZeroR baseline. For CLARITY, THREADPART$_3$ achieves the best result for the task, beating the BOW at a level of statistical significance for $\mathcal{F}_\mu$. Recall that THREADPART$_2$ was based on a graph of co-participation in threads, suggesting that knowledge of which users co-post to threads is informative in predicting how clear their posts are on average. In other words, there are clusters of users who co-predict their respective post clarity.

For POSITIVITY, POSTAFTER beats the BOW benchmark, but not at a level of statistical significance in this case. POSTAFTER may work in capturing POSITIVITY due to sets of antagonistic users who respond to each other's posts negatively (e.g. commonly engage in flame wars), or to cooperative users who engage in a mutually-supportive dialogue, each building positively on the previous poster's comments.

For both CLARITY and POSITIVITY, the aforementioned individual feature sets achieve the best overall results in our experiments, i.e. combining these feature sets with BOW or other feature sets did not improve the results. In both cases, the MAE is around 1.3.

For PROFICIENCY and EFFORT, the BOW $\mathcal{F}_\mu$ results were notably higher, to the degree that none of the feature sets in isolation were able to better it. As a result, we looked to the combination of up to three feature sets, and present in Table 6 the best-achieved results with two or three feature sets for PROFICIENCY and EFFORT. In both cases, it is the combination of the BOW feature set with one of the User Network feature sets and one of the post-level feature sets that produces the best result, illustrating the complementary nature of the three basic feature set types. Results for the BOW feature set in isolation, along with results for BOW with each of the two feature sets in the best-performing method, are presented to illustrate the relative effect of each. In the case of PROFICIENCY, THREADPART$_2$ and BALDWIN$^{Post}_{\text{AGG}}$ both lead to increased $\mathcal{F}_\mu$ when combined with BOW, as compared to the simple feature set (but only the combination of all three is significantly better than simple BOW). That is, PROFICIENCY appears to be the most multi-faceted of the four user classification attributes, in being best captured through the combination of lexical choice, macro post-level features, and network-based analysis of thread co-participation. With the network-based features, we suggest this is largely a negative effect, in that "hacks" and "newbies" are characterised by a *lack* of thread co-participation.

With EFFORT, BOW achieves by far its highest $\mathcal{F}_\mu$ across all four classification tasks, and the combination with THREADPART$_3$ and WANAS$_{\text{AGG}}$ barely surpasses it, at a level which is not statistically significant.

That the best results are achieved in all four classification tasks with network-based features (possibly in combination with other feature sets) is telling, and underlines the potential of network analysis for user classification. The aggregate post-level feature sets BALDWIN$^{Post}_{\text{AGG}}$ and WANAS$_{\text{AGG}}$ are less effective, but bear in mind that they were not tailored specifically for the user classification task, so it is a positive result that they have an impact when aggregated over user-level structure, and suggests that further work in customizing the per-post feature set will yield further improvements on this task.

Finally, we turn to analysis of inter-annotator agreement for the four user classification subtasks,

Attribute	*Feature Sets Present*	$\mathcal{F}_\mu$	*MAE*
PROFICIENCY	BOW	0.240	1.380
	BOW $\oplus$ THREADPART$_2$	0.260	1.280
	BOW $\oplus$ BALDWIN$^{Post}_{\text{AGG}}$	0.320	1.200
	BOW $\oplus$ THREADPART$_2$ $\oplus$ BALDWIN$^{Post}_{\text{AGG}}$	**0.360**$^+$	**1.080**
EFFORT	BOW	0.320	1.240
	BOW $\oplus$ THREADPART$_3$	0.320	1.240
	BOW $\oplus$ WANAS$_{\text{AGG}}$	0.300	1.280
	BOW $\oplus$ THREADPART$_3$ $\oplus$ WANAS$_{\text{AGG}}$	**0.340**	**1.220**

Table 6: Results for augmented feature sets

Attribute	BOW	*Best*	*MIA*	p_{BoW}	p_{Best}
CLARITY	0.120	0.260	0.240	**0.049**	0.723
PROF	0.240	0.360	0.395	**0.009**	0.427
POS	0.140	0.220	0.335	**0.011**	0.126
EFFORT	0.320	0.340	0.410	0.108	0.193

Table 7: BOW benchmark, best result and mean inter-annotator (MIA) $\mathcal{F}_\mu$ over each user attribute

to gauge the quality of the results achieved by our best classifiers in each case. In Table 7, we reproduce the BOW and best $\mathcal{F}_\mu$ results from Tables 5 and 6, and additionally present the mean inter-annotator (MIA) $\mathcal{F}_\mu$ based on leave-one-out cross-validation. We additionally present the p-value for the two-tailed paired t-test for each of BOW–MIA and best–MIA. In addition to being able to compare the $\mathcal{F}_\mu$ values directly, we can observe that for CLARITY, PROF(ICIENCY) and POS(ITIVITY), the best-performing classifier is both significantly better than the BOW benchmark (and `ZeroR` baseline), and statistically indistinguishable from the upper bound figure. In the case of EFFORT, there is no significant difference between BOW and the upper bound, so it would highly unlikely that we could achieve a significant improvement over BOW for any of our classifiers.

In summary, we were able to consistently exceed the majority class baseline on this task using user-level features, attaining results that were competitive with those utilising a state-of-the-art bag-of-words benchmark. We found that in most cases our results exceeded the benchmark to a high degree of statistical significance, with network-based features featuring prominently for all classification subtasks.

8 Further Work

Given that the intention of this work is to enhance information access over web forum data, the next step we intend to take is to apply our trained classifiers to a larger corpus of web forum data, and assess the impact of the predictions in a task-based evaluation. Examples of such tasks include predicting perceived post quality (Weimer and Gurevych, 2007) and identifying troubleshooting-oriented threads (Baldwin et al., 2007). We also note that there is limited room for progress given our current interpretation of the inter-annotator agreement. We intend to further analyze the annotations. In particular, since each annotator annotated each user five times, we intend to study the interaction between the number of context posts and the ratings given by the annotator.

9 Conclusion

In this work, we introduced a novel user classification task over web user forums. We prepared an annotated dataset relevant to the task, which we will release to the research community.

We extracted user-level features over aggregations of user posts, as well as via anaylsis of social networks in a web forum. We investigated each feature set we defined in isolation as well as in combination with the benchmark feature sets. We have shown that these user-level features can consistently outperform a majority-class baseline over a user classification task.

We succeeded in showing that user-level features have empirical utility in user classification, and we expect that the use of these features will generalize well to tasks over other aspects of threaded discourse, for example in profiling users or in ranking threads for information retrieval.

Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

Rakesh Agrawal, Sridhar Rajagopalan, Ramakrishnan Srikant, and Yirong Xu. 2003. Mining newsgroups using networks arising from social behavior. In *Proceedings of the Twelfth International World Wide Web Conference (WWW'03)*, pages 529–535, Budapest, Hungary.

Timothy Baldwin, David Martinez, and Richard Baron Penman. 2007. Automatic Thread Classification for Linux User Forum Information Access. In *Proceedings of the Twelfth Australasian Document Computing Symposium (ADCS 2007)*, pages 72–79, Melbourne, Australia.

Timothy Baldwin, David Martinez, Richard B Penman, Su Nam Kim, Marco Lui, Li Wang, and Andrew Mackinlay. 2010. Intelligent Linux Information Access by Data Mining : the ILIAD Project. In *Proceedings of the NAACL 2010 Workshop on Computational Linguistics in a World of Social Media: #SocialMedia*, pages 15–16, Los Angeles, USA.

Krisztian Balog, Leif Azzopardi, and Maarten de Rijke. 2006. Formal models for expert finding in enterprise corpora. In *Proceedings of 29th International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 43–50.

Sergei Brin and Larry Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117.

Vitor Carvalho, Wen Wu, and William Cohen. 2007. Discovering leadership roles in email workgroups. In *Proceedings of the 4th Conference on Email and Anti-Spam (CEAS 2007)*, Mountain View, USA.

Anurat Chapanond, Mukkai S. Krishnamoorthy, and Bülent Yener. 2005. Graph theoretic and spectral analysis of Enron email data. *Computational and Mathematical Organization Theory*, 11(3):265–281.

Gao Cong, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng Sun. 2008. Finding question-answer pairs from online forums. In *Proceedings of 31st International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*, pages 467–474, Singapore.

Nick Craswell, Arjen P. de Vries, and Ian Soboroff. 2005. Overview of the TREC-2005 Enterprise track. In *Proceedings of the 14th Text REtrieval Conference (TREC 2005)*, Gaithersburg, USA.

Petros Drineas, Mukkai S. Krishnamoorthy, Michael D. Sofka, and Bülent Yener. 2004. Studying e-mail graphs for intelligence monitoring and analysis in the absence of semantic information. In *of the IEEE International Conference on Intelligence and Security Informatics*, pages 297–306, Tucson, USA.

Jonathan L. Elsas and Jaime G. Carbonell. 2009. It pays to be picky: An evaluation of thread retrieval in online forums. In *Proceedings of 32nd International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*, pages 714–715, Boston, USA.

Hui Fang and ChengXiang Zhai. 2007. Probabilistic models for expert finding. In *Proceedings of the 29th European Conference on Information Retrieval (ECIR'07)*, pages 418–430, Rome, Italy.

Blaz Fortuna, Eduarda Mendes Rodrigues, and Natasa Milic-Frayling. 2007. Improving the classification of newsgroup messages through social network analysis. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management (CIKM '07)*, pages 877–880, Lisboa, Portugal.

Chih-Wei Hsu and Chih-Jen Lin. 2006. BSVM-2.06. `http://www.csie.ntu.edu.tw/cjlin/bsvm/`. Retrieved on 15/09/2009.

Thorsten Joachims. 1998. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, pages 137–142, Chemnitz, Germany.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Su Nam Kim, Li Wang, and Timothy Baldwin. 2010. Tagging and Linking Web Forum Posts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 192–202, Uppsala, Sweden.

Marco Lui and Timothy Baldwin. 2009a. hydrat. `http://hydrat.googlecode.com`. Retrieved on 15/09/2009.

Marco Lui and Timothy Baldwin. 2009b. You Are What You Post: User-level Features in Threaded Discourse. In *Proceedings of the 14th Australasian Document Computing Symposium*, pages 98–105, Sydney, Australia.

Marco Lui. 2009. *Impact of user characteristics on online forum classification tasks*. Honours thesis, The University of Melbourne, Australia.

Robert Malouf and Tony Mullen. 2008. Taking sides: Graph-based user classification for informal online political discourse. *Internet Research*, 18(2):177–190.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Mark E.J. Newman and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical Review E*, 69. Article Number 26113.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Jangwon Seo, W. Bruce Croft, and David A. Smith. 2009. Online community search using thread structure. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 1907–1910, Hong Kong, China.

Ian Soboroff, Arjen P. de Vries, and Nick Craswell. 2006. Overview of the TREC-2006 Enterprise track. In *Proceedings of the 15th Text REtrieval Conference (TREC 2006)*, Gaithersburg, USA.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 327–335, Sydney, Australia.

Nayer Wanas, Motaz El-Saban, Heba Ashour, and Waleed Ammar. 2008. Automatic scoring of online discussion posts. In *Proceedings of the 2nd ACM Workshop on Information Credibility on the web (WICOW '08)*, Napa Valley, USA.

Markus Weimer and Iryna Gurevych. 2007. Predicting the perceived quality of web forum posts. In *Proceedings of the 2007 Conference on Recent Advances in Natural Language Processing (RANLP-07)*, Borovets, Bulgaria.

Markus Weimer, Iryna Gurevych, and Max Mühlhäuser. 2007. Automatically assessing the post quality in online discussions on software. In *Proceedings of the 45th Annual Meeting of the ACL: Interactive Poster and Demonstration Sessions*, pages 125–128, Prague, Czech Republic.