Automatic Scoring of Online Discussion Posts

Nayer Wanas Motaz El-Saban Heba Ashour Waleed Ammar

Cairo Microsoft Innovation Center
Smart Village, KM 28 Cairo/Alex Desert Rd.
AbouRawash, Giza, 12676, Egypt
(+202) 3536-3207

{nayerw,motazel,hebaa,i-waamma}@microsoft.com

ABSTRACT

Online discussions forums, known as forums for short, are conversational social cyberspaces constituting rich repositories of content and an important source of collaborative knowledge. However, most of this knowledge is buried inside the forum infrastructure and its extraction is both complex and difficult. The ability to automatically rate postings in online discussion forums, based on the value of their contribution, enhances the ability of users to find knowledge within this content. Several key online discussion forums have utilized collaborative intelligence to rate the value of postings made by users. However, a large percentage of posts go unattended and hence lack appropriate rating.

In this paper, we focus on automatic rating of postings in online discussion forums. A set of features derived from the posting content and the threaded discussion structure are generated for each posting. These features are grouped into five categories, namely (i) relevance, (ii) originality, (iii) forum-specific features, (iv) surface features, and (v) posting-component features. Using a non-linear SVM classifier, the value of each posting is categorized into one of three levels High, Medium, or Low. This rating represents a seed value for each posting that is leveraged in filtering forum content. Experimental results have shown promising performance on forum data.

Categories and Subject Descriptors

H.3.3 [**Information storage and retrieval**]: Filtering

I.5.2 [**Feature Evaluation and detection**]: Feature evaluation and selection

General Terms: Algorithms, Human Factors, Theory

Keywords: Content Filtering, Forums, Online Communities

1. INTRODUCTION

Online discussion forums, also known as forums, are web applications that hold user-generated content. The basic component of a forum is a threaded discussion set of posts. Threads covering the same topic are collected in a sub-forum, with a set of sub-forums collectively referred to as a forum. Forums have been around since the early days of the internet, and they form rich repositories of collaborative knowledge. Since forums are conversational social cyberspaces, the users dictate the quality of the content posted, and hence it may vary widely. Navigation through this repository to find useful information can be difficult and time consuming. Several key online discussion forums have utilized collaborative intelligence to indicate the posts that are worth attending too[9]. Most of these forums allow users to rate posts on a five point scale (1 being low and 5 being high), with some allowing for a finer resolution. These scores are used to filter online forum content based on their value to help users surf knowledge with ease.

Lampe and Resnick[9] suggested a post rating scheme that has shown to be sound. However, a good portion of the threaded discussion could pass before users identify the value of its posts. Additionally, later posts are usually overlooked by moderators. Wrongly rated posts were usually not reversed, along with the fact that the quality of the rating was highly affected by the value of the initial post. Collectively these factors play a role in the amount of knowledge being surfaced in online discussion forums.

In this paper we investigate the idea of automatically assessing online discussion posts based on their quality. The assessment tries to model how users would perceive the posts and provides a seed value for all posts. These seed values will incent users to collaboratively evaluate the quality of the posts, leading to a more refined assessment. We present a set of features through which a classifier is invoked to evaluate individual posts. These features enable automatic content filtering in online discussion forums based on their content.

2. AUTOMATIC ASSESSMENT OF ONLINE POSTS

Online discussion groups, Newsgroups, Usenet and other community generated knowledge repositories have attracted recent attention pertaining to social networks and structure within these repositories[1][8], understanding social roles and users[4][5], and computational linguistics[7]. Lui et.al.[10] evaluated automatic categorization of online discussion posts based on three category sets. The category sets included academic vs. general, seek vs. contribute, and the topic of the post. While the results were not conclusive, they suggested that the performance was sufficient for monitoring learning progress in educational online discussion forums. A document-term matrix, that employed stemming and word frequency, was used in the categorization. Topic detection has also been applied to online discussion forums[12]. Through their work, Wu and Li detected clusters of posts revolving around topics previously unknown to the forum moderators and experts. They used the participation frequency of users to detect the clusters and hence be able to appropriately label them. Fortuna et.al.[6] suggested the use of features generated by the author and user networks, developed

WICOW'08, October 30, 2008, Napa Valley, California, USA.
Copyright 2008 ACM 978-1-60558-259-7/08/10...$5.00.

through online interaction, to cluster posts. Through a trained classifier, they aimed to determine a label for each post, namely agree, disagree, insult, question, answer, off-topic or unknown. They also experimented with trying to only label posts as being a question or answer. While the first showed signs of acceptable performance using the features suggested, the latter did not demonstrate the same effect.

While several researchers have attempted to automatically assess the quality of text documents[3], Weimer et.al.[11] proposed an algorithm for automatic assessment of post quality. The proposed algorithm addressed the issues pertaining to forum posts, namely being short related fragments of text. They suggested a set of features that ranged from surface features, such as Capital Word Frequency, to more linguistically intense features, such as lexical and syntactic features. A trained classifier was used to classify posts into two groups, 'bad' and 'good'. This categorization of posts is coarse and not sufficiently appropriate to seed posts' values. While the suggested forum specific features contributed most to the accuracy of the classification, many of the features captured linguistic aspects of posts. This dictated that posts used as training data observe proper use of language. This limited the initially small training set used in the experimental study further. Additionally, the assumption that posts would follow proper linguistic rules is not always true and would be dependent on the language of the forum, implying that individual classifiers would be required for forums in different languages.

Our research focuses on providing a finer level for rating posts (Low, Medium and High). It also is conscious of linguistic phenomena pertaining to online discussion forums. This is achieved through avoiding commitment to linguistic features and the generation of keywords from within the forum instead of using a predefined lexicon of terminology and jargon used in the forum. This follows from the fact that keywords used within online discussion forums reflect the understanding that this community has to specific terms, which may not be similar to the way other communities may perceive it. This makes our approach more independent of language, including community specific jargon.

3. POST SCORING METRICS

Online discussion forum posts are characterized to be generally short text fragments. Users also take significant liberties in the language and presentation styles of posts. Another important component that affects the way that users perceive online discussion forums is their order and relationship with other posts, in addition to their posting location. Collectively these factors render the accurate and exact evaluation of posts a difficult task. It also invokes a significant amount of Natural Language Processing (NLP) to fully understand and analyze the posts. However, in the scope of this work we are interested in providing a seed value for each post, through which a moderation process would rectify any misclassification. In addition, we are interested in avoiding linguistically involved approaches since users don't apply strict rules on the linguistic content of posts in online discussion forums.

To that end, we suggest a set of 22 features that are divided into five categories, namely (i) relevance (2), (ii) originality (2), (iii) forum-specific (7), (iv) surface (5), and (v) posting component features (6). In the following we will shed more light on these feature categories.

3.1 Relevance Features

Probably one of the most important aspects that affect the perception of users for a given post is its relevance. Relevance reflects the appropriateness of a post to the thread and the sub-forum it inhabits. To approximate both these aspects, two features, *OnSubForumTopic*, and *OnThreadTopic* are evaluated as follows:

3.1.1 OnSubForumTopic

In essence, OnSubForumTopic aims to capture the degree a post has remained relevant to the sub-forum it resides within. While in many contexts, a set of keywords could be formulated for prior knowledge, community dynamics within online discussion forums may shift, rendering preset keywords irrelevant. To overcome this phenomenon, keywords are generated from within the forum content as it evolves. This is achieved by generating a set of keywords that distinct each sub-forum from others existing in the same forum. These keywords are generated using a *tfidf* measure on a bag of words (BOW) combining all the words of posts in the sub-forum. These keywords represent the communal perception of important terms that distinguish the given sub-forum within online discussion forums. The top 10% of these keywords (F_N) are used to represent the knowledge of the given sub-forum. The BOW of each post in the sub-forum (P_j)is then compared against the keywords descriptive of the sub-forum to generate the *onSubForumTopic* measure for the *j*th posting. *OnSubForumTopic(Pj)* is calculated as follows :

$$OnSubForumTopic(P_j) = \frac{count(P_j \in F_N)}{|P_j|} \forall\, j = 1 \dots n$$

where n is the number of posts in the sub-forum, P_j is the set of words in the *j*th post's body and title, and F_N is the sub-forum's knowledge base.

3.1.2 OnThreadTopic

Since the leading post in a thread and its title are the entry point to any threaded discussion, maintaining relevance to both these components signifies that users could find information with relative ease. Therefore, *OnThreadTopic* is used to measures the relevance of a post to the discussion it is in by comparing each post's bag of words to that of the leading post according to the following equation:

$$OnThreadTopic(P_j) = \frac{count(P_j \in P_1)}{|P_j|} \forall\, j = 2 \dots n$$

The leading post of the thread is treated specially, and its *OnThreadTopic* measure follows the following equation:

$$OnThreadTopic(P_1) = \frac{count(body(P_1) \in title(P_1))}{|P_1|}$$

Where body(P_1) is the set of words in the lead post's body, and title(P_i) is the set of words in the post's title.

3.2 Originality Features

Since posts that contribute new knowledge are perceived to be of value, originality goes hand in hand with relevance in dictating the value of a given post. While originality is hard to measure, a degree of similarity is significantly easier to realize. The lack of similarity is not exactly a reflection of originality, however, it gives a good indication of the novelty presented by a given post. Two measures of originality are suggested, *OverlapPrevious* and *OverlapDistance*.

3.2.1 OverlapPrevious

This feature measures the maximum degree of overlap between the terms used in a posting and all other posts that precede it in the same thread. While the order of terms is generally perceived as important, the nature of content being short and less structured collectively reduces the importance of word order. As a result, we calculate the overlap between the words of a given post and all of its previous posts as follows:

$$Overlap(P_i, P_j) = \frac{count(P_i \in P_j)}{|P_i|} \ \forall\, i > j, j = 1 \ldots n$$

Therefore, *OverlapPrevious(P$_i$)* is evaluated as

$$OverlapPrevious(P_i) = \max_{j}(Overlap(P_i, P_j))$$

3.2.2 OverlapDistance

This feature reflects the separation distance, in terms of number of posts, between the current post and that which has been judged as the most overlapping by the *OverlapPrevious* measure. The hypothesis is that the closer the overlapping posts are, the less value a post has.

3.3 Forum-specific Features

There are a few aspects of a given post that are specific to online discussion forums, including the number of times a post is quoted and the amount of discussion a post stimulated. The features used to capture these aspects are *Referencing* and *Replies*.

3.3.1 Referencing

Quotation of text chunks from previous posts, and by subsequent posts, increases the value of a given post to a discussion. Furthermore, the method of referencing text may signify its importance. For example, utilizing fragments of text, rather than full posts, and adding comments around the fragments indicate more focused posts. Additionally, the amount of text quoted from a given post relative to its content reflects the amount of contribution present. A ratio of quoted text to the post size, normalized by the size of the original post, is used to evaluate individual chunks.

Since quotation is a direction metric, two features pairs are evaluated for a given post, namely *CountBackwardReferences* and *BackwardReferencing* on one end, and *CountForwardReferences* and *ForwardReferencing* on the other end.

3.3.1.1 CountBackwardReferences

This metric represents the number of quotation chunks in the given post that are extracted from earlier posts.

3.3.1.2 BackwardReferencing

This feature aims to quantify the value added to a given post by the quotations it contains. It is calculated according to the following equation:

$$BackwardReferencing(P_{ij}) = \sum_i \Big(\frac{size\ of\ quoted\ text}{|P_i|} \times \frac{size\ of\ quoted\ text}{|P_j|}\Big)$$

3.3.1.3 CounForwardReferences

This metric represents the number of times the post has been referenced in subsequent posts.

3.3.1.4 ForwardReferencing

This feature aims to reflect the value added by a given post to subsequent posts that quote it. It is calculated according to the following equation:

$$ForwardReferencing(P_j) = \sum_i \Big(\frac{size\ of\ quoted\ text}{|P_i|} \times \frac{size\ of\ quoted\ text}{|P_j|}\Big)$$

3.3.2 Replies

The number of replies generated by a given post is an indication of its value, either through contribution or controversy. The number of replies reflects users' interest in a given post. In case nesting of replies is allowed, the number of levels spanned by replies to the post is also an important factor.

3.4 Surface Features

Surface features reflect the way a user presents a given post, irrespective of the content. The amount of care a given user gives to a post affects the way readers perceive its value. The more readers find the post easier to read, the greater the value they associate to the post. Three metrics are used to assess surface features, namely *Timeliness*, *Lengthiness*, and *Formatting Quality*,

3.4.1 Timeliness

This feature is a reflection of how fast a user presents his contribution. The rate of replies is dictated by the community, and falling within their norm would increase the probability of posts being viewed. To reflect this aspect, timeliness is calculated as follows:

$$Timeliness(P_j) = \frac{time\ difference\ between\ P_j\ and\ P_{j-1}}{Average\ inter - posting\ time\ in\ thread}$$

3.4.2 Lengthiness

Similar to *Timeliness*, this measure is associated with the length of a post, measured by word count. A post conforming to the length of a posting the community accepts as normal reflects value. Hence, the length of a give post is normalized by the mean length of posts in a give thread as follows:

$$Lengthiness(P_j) = \frac{|P_j|}{Average\ length\ of\ postings\ in\ thread}$$

3.4.3 Formatting Quality

Aspects involved in post formatting affect the perception of users of its value. The excessive use of punctuation marks, emoticons and consecutive capital letters generally reduces the level of professionalism of posts, consequently undermining their value. These three aspects are reflected using three features, namely *FormatPunctuation*, *FormatEmoticons*, and *FormatCapitals*. Collectively, they comprise the metrics of formatting quality.

3.4.3.1 FormatPunctuation

The hypothesis behind this feature is that extensive use of creative punctuation affects the perceptions of the post by users. For that reason, *FormatPunctuation(P$_j$)* is calculated as follows:

$$FormatPunctuation(P_j) = \frac{number\ of\ chunks\ of\ consecutive\ punctuations\ in\ posting\ j}{number\ of\ sentences\ in\ posting\ j}$$

3.4.3.2 FormatEmoticons

The hypothesis behind this feature is that extensive use of emoticons in a given post conveys a level of emotion that affects the perceptions of the post by users. For that reason, *FormatEmoticons(P$_j$)* is calculated as follows:

$$FormatEmoticons(P_j) = \frac{number\ of\ emoticons\ in\ posting\ j}{number\ of\ sentences\ in\ posting\ j}$$

The set of emoticons considered is the set of 76 emoticons presented in the Windows Live Messenger program.

3.4.3.3 FormatCapitals

The hypothesis behind this feature is that extensive use of consecutive capital letters gives the post a tone that might affect its perceptions by users. For that reason, *FormatCapitals(P$_j$)* is calculated as follows:

$$FormatCapitals(P_j) = \frac{number\ of\ chunks\ of\ consecutive\ capitals\ in\ posting\ j}{number\ of\ sentences\ in\ posting\ j}$$

3.5 Posting Component Features

A post may contain important syntactic and web elements such as questions and web-links respectively. While most dialogues on online discussion forums revolve around questions, web-links add value and credibility to posts, through soliciting the value of the referenced content. These forum elements are captured by two metrics, *WebLinks*, and *Questioning*.

3.5.1 Weblinks

The presence of appropriate web-links generally adds value to posts. This value is composed of three factors, (i) the relevance of the web-link, (ii) the presentation of the web-link, and (iii) the added value contributed by the user to explain the value of this web-link. These three factors comprise a set of two metrics to assess the value of web-links present in the post, namely *Weblinking* and *WeblinkQuality*.

3.5.1.1 Weblinking

This feature represents the effort aspects of how the user presented the webl-inks in his post. It is calculated as follows:

$$Weblinking(P_j) = \frac{\sum_{All\ Weblinks} number\ of\ sentences\ with\ weblinks\ in\ post\ j}{number\ of\ sentences\ in\ post\ j} \times WeblinkFormat$$

where

$$WeblinkFormat = \begin{cases} 1 & if\ URL\ is\ inserted \\ 0.5 & if\ hyperlinked\ text \end{cases}$$

3.5.1.2 WeblinkQuality

Since the user is presenting the web-link in his post as an additional resource, its content should be relevant to the general content of the sub-forum the post is in. Therefore, this feature measures the similarity between the words in the webpage linked to and the sub-forum the post is in. This is captured as follows:

$$OnForumTopic(P_j) = \sum_{\forall\ weblinks} \frac{count(WebPage\ words\ \in F_N)}{|WebPage\ words|}$$

where F$_N$ is the sub-forum's knowledge base i.e. its representative set of keywords.

3.5.2 Questioning

Questions, and subsequently their answers, are one of the major components of online discussion forums. In order to capture their value, we include a set of three question related features in our feature set. The first feature is the number of questions in a post. The presence of question in a post is signaled based on a set of templates both from the surface form of the sentences (e.g. question mark and Wh-questions) and the part of speech (POS) tags of words surrounding a Wh-question word (e.g. which). The idea behind including POS tags is to improve the precision of question detection to overcome situations like: "I mean what you heard". The second and third question features aim at capturing the intuition that if a question A asked in a post P$_j$ is quite similar to a previously asked question in the forum, then the value of post P$_j$ is not increased by much. Hence, we include a distance measure between the question in the current post and questions previously encountered in the forum. In our implementation, we index all the questions posted in the forum and we use a search engine to compute a question distance measure. To allow for multiple questions in the same post, we use the average and standard deviation of distance measures over the set of questions in a given post as our second and third question features.

4. EXPERIMENTAL SETUP

The goal of this work is to provide a seed rating for posts in online discussion forums. The seed rating of posts is based on a three level scale, namely low, medium or high. To train a classifier to determine the post rating, manually labeled data must be used. Several key online discussion forums have used collaborative intelligence to rate postings. These labels are used to train a classifier.

4.1 Dataset

The dataset used is composed of discussion threads from the Slashdot online discussion forum[1]. 200 threads with a maximum of 200 posts each were selected from the 14 sub-forums on Slashdot. A total of 120,000 posts were scraped from the discussion forum. Posts on Slashdot are rated on a scale from -1 to 5, where irrelevant posts are rated -1 and high quality posts are rated 5. The default rating for a registered user is 1 and for an unregistered user is 0. In order to ensure that posts used in training were moderated, posts rated as 0 were removed, unless they were from a registered user. Posts with a rating 1 from a registered user were also removed. In addition to the rating, posts are tagged. Posts with tags "funny", "troll", "flamboyant" and those with no tag were removed, because these tags distorted the classification process. Hence, the final dataset was composed of 20,008 rated posts, which were clustered into three groups, namely low, medium, and high, according to their value. Since the default rating of a post, depending on the user karma, could range between 0 and 2, posts rated as-1,0,1 or 2 were clustered as low.

[1] http://www.slashdot.org

Posts rated as 3 were considered medium. While posts rated 4 or 5 were clustered as high since this reflects that they have definitely been manually rated more than twice.

4.2 Classifier Training

A non-linear Support Vector Machine (SVM) classifier was trained using *LibSVM*[2], using and RBF kernel, to test the effectiveness of the features used. It is worth mentioning that it is possible that we could treat this as a ranking problem, yet the performance observed using the LambdaMART (multiple additive regression trees) algorithm[12] was sub-optimal compared to classification. The SVM classifier parameters were optimized for best performance on an independent test set to C=100000, and γ= 0.00005. Five-fold cross validation was used on balanced data to evaluate the classifier performance. The performance was evaluated based on the accuracy and F1-measure.

The overall average accuracy and F1-measure of applying the classifier on the data set was 49.5% and 48.9% respectively. This level of accuracy is acceptable to provide a seed rating for posts, where moderation would subsequently apply. It was observed that the performance on the posts rated “High” was significantly better than those rated as “Medium” and “Low” (Table 1). This is mostly due to the incremental nature of the moderation policy implemented in Slashdot[9], where each post could be boosted or demoted by only 1 point at a time, and the large volume of posts being published to the forum imply that posts with medium or low rating may be of value but have gone unattended. Additionally, posts with high rating means several moderators have attended to it and have increased its value accordingly. However, since rating is inherently subjective, it is hard to attain accurate distinction between these levels. Another notable element in these results is the 1-step error. While the overall accuracy is around the 50% mark, the 1-step error is almost 90%. That is to say that, misclassification of posts usually implies moving from one level to that immediately adjacent to it.

Table 1: F1-measure on the three rating levels High, Medium and Low

	High	Medium	Low
F1-Measure	0.61	0.42	0.46

4.3 Parametric Evaluation

While a more detailed and involved parametric study is still due on the metrics suggested, we experimented with some preliminary experimentation to determine the role of the different metrics. Table 2 outlines the relative accuracy and F1-measure contributed by each metric category individually. It is worth noting that the forum-specific features provide the most significant contribution to the overall accuracy (Achieves almost 97% of the accuracy of all metrics combined). These features measure the relationship between a given post and other posts within the forum. The results also signify that the relevance and posting-component metrics have the lowest performance. This is mostly due to their dependence on analyzing post text, without linguistic tools to maintain language independence. While it is expected that the performance could improve if language tools are deployed, it would imply addition requirements on users to observe proper language in their posts. Such limitations might limit forum usage in general.

Table 2: Relative Accuracy and F1-measure for each metric category

Metric Category	Relative Accuracy	Relative F1-measure
Relevance	64.46%	53.94%
Originality	89.17%	70.20%
Forum-Specific	**96.97%**	**96.25%**
Surface	76.98%	71.85%
Posting-Component	65.20%	44.72%

It is interesting to observe that while originality was the second most contributing metric on its own, the combination of forum-specific and surface features was the best pair (Table 3). It is interesting that these two features assess posts at face value and don’t look at the posting content.

Table 3: Relative accuracy and F1-measure for metrics pairs (R: Relevance, O: Originality, F: Forum-Specific, S: Surface, P: Posting-Component)

R	O	F	S	P	Rel. Accuracy	Rel. F1-measure
✓	✓				87.39%	78.09%
✓		✓			95.02%	94.89%
✓			✓		79.55%	74.80%
✓				✓	67.91%	52.45%
	✓	✓			94.91%	94.67%
	✓		✓		79.29%	77.11%
	✓			✓	88.51%	75.31%
		✓	✓		**98.82%**	**98.53%**
		✓		✓	97.71%	97.31%
			✓	✓	77.52%	73.43%

Table 4 shows the results of all remaining combinations of metrics. It is observed that the metric combinations including the forum-specific and surface features are the best performers. It is worth noting that the relevance and posting-component metrics require an overhead to evaluate keywords, and resulting improvement on accuracy and F1-measure may not be considered significant.

5. ONLINE DISCUSSION CONTENT FILTERING

The ability to find knowledge within online discussion forums is an important task. Currently, most online forums depend on human ratings through a variety of collaborative intelligence schemes to filter and order content. However, among the major drawbacks of these schemes is the fact that many posts go unattended or inappropriately rated. An automatic post assessment approach could provide an alternative to this manual rating.

The rating provided through training is considered a seed value for the posting. Users are allowed to moderate the value of a posting appropriately, however, postings that have not been moderated will get a better opportunity to be visible, and hence the potential knowledge they contain would be used.

In addition, through the course of evaluating relevance, a set of key words are generated to describe each sub-forum. These keywords could be used as browsing elements in posts in a given sub-forum. This is in contrast to having a pre-defined set of keywords, which the community might not be using to express their ideas.

Another application of automatic scoring is the accumulation of user credibility. Analysis of user participation in and contribution in online discussion forums represents a means to assess their value to this community. It can also serve as a means to direct participation and improve performance and engagement of users in these discussions. This is equivalent to the user karma, or credibility, with an online discussion community.

Table 4: Relative accuracy and F1-measure for metric combinations (R: Relevance, O: Originality, F: Forum-Specific, S: Surface, P: Posting-Component)

R	O	F	S	P	Rel. Accuracy	Rel. F1-measure
✓	✓	✓	✓	✓	100.00%	100.00%
✓		✓	✓		99.64%	99.41%
	✓	✓	✓	✓	99.64%	99.57%
✓	✓	✓	✓		99.61%	99.43%
	✓	✓	✓		**99.36%**	**99.10%**
		✓	✓	✓	99.36%	99.20%
✓		✓	✓	✓	99.23%	99.12%
✓	✓	✓			95.90%	95.64%
✓	✓	✓		✓	95.75%	95.73%
	✓	✓		✓	95.50%	95.48%
✓		✓		✓	95.24%	95.30%
✓	✓			✓	88.73%	78.47%
✓	✓		✓	✓	86.30%	83.06%
✓	✓		✓		86.19%	82.6%
✓			✓	✓	80.90%	77.49%
	✓		✓	✓	80.76%	78.92%

6. CONCLUSION AND DISCUSSION

In this paper, we present a set of metrics to assess the value of a post in an online discussion forum. These metrics are language independent and were used to train a non-linear SVM classifier to rate posts in one of three groups (High, Medium or Low). The classifier achieved an accuracy of 50%, and a 1-step error of 90%. This accuracy is sufficient to provide a seed rating to a given post, a rating which users are allowed to alter through moderation. This will allow posts better exposure to the online discussion forum readers. Experiments have demonstrated that structural features of posts were significantly more important than those utilizing text analysis, allowing for a language independent approach to be adopted. The poor performance of the relevance and originality features were mostly due to the use of only individual word frequencies, without any computational linguistic elements. Stop-word removal, POS tagging and phrase extraction are technologies that potentially help improve accuracy. Applying these technologies would however, make this approach language dependent and hence limit its generality.

Another aspect that has contributed to the level of accuracy achieved is the moderation process of the online discussion forum used (Slashdot). The forum uses an incremental moderation process, where each post could be boosted or demoted by only 1 point at a time. In addition, we were unable to assess the number of judges and inter-judging agreement on posts. This would reflect in the quality of the labeling used and hence would reflect on the overall accuracy of classification. Nonetheless, the results were suitable to provide a seed rating for posting through which filtering was possible. In addition, keywords generated to evaluate relevance of a post were useful in filtering posts based on textual content they contain.

7. REFERENCES

[1] Borgs, C., Chayes, J., Mahdian, M., and Saberi, A., 2004. Exploring the Community Structure of Newsgroups, In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Seattle, WA, USA, August 22-25, 2004) KDD'04, ACM Press, New York, NY, 783-787

[2] Chang, C., and Lin, C. 2001. *LibSVM*: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[3] Dikli, S., 2006. An Overview of Automatic Scoring of Essays. The Journal of Technology, Learning, and Assessment, Vol 5(1) August 2006, 3-35.

[4] Fiore, A., Teirman, S., and Smith, M., 2002. Observed behavior and perceived value of authors in usenet newsgroups: bridging the gap. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Minneapolis, MN, USA, April 20-25, 2002). CHI '02. ACM Press, New York, NY, 323-330.

[5] Fisher, D., Smith, M., and Welser, H., 2006. "You Are Who You Talk To: Detecting Roles in Usenet Newsgroups. In Proceedings of the 39th Hawaii International Conference on System Sciences (Kauai, HI, USA, January 4-7, 2006) Track 3, HICSS-39, IEEE Press, New Jersey, NJ, 59b.

[6] Fortuna, B., Rodrigues, E., and Milic-Frayling, N. 2007. In Proceedings of the Conference on Information and Knowledge Management (Lisboan, Portugal, November 6-8, 2007). CIKM '07. ACM Press, New York, NY, 585-588

[7] Glance, N., Hurst, M., Nigam, K., Siegler, M., Stockton, R., and Tomokiyo, T. 2005. Deriving Marketing Intelligence from Online Discussion. In Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining (Chicargo, IL, USA, August 21-24, 2005). KDD'05. ACM Press, New York, NY, 419-428

[8] Gómez, V., Kaltenbrunner, A., and López, V. 2008. Statistical Analysis of the Social Network and Discussion Threads in Slashdot. In Proceedings of the 17th International World Wide Web Conference (Beijing, China, April 21-25, 2008). WWW2008. ACM Press, New York, NY, 645-654

[9] Lampe, C. and Resnick, P. 2004. Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Vienna, Austria, April 24-29, 2004). CHI '04. ACM Press, New York, NY, 543-550

[10] Lui, A., Li, S., and Choy, S. 2007. An Evaluation of Automatic Text Categorization in Online Discussion Analysis. In Proceedings of the Seventh IEEE International Conference on Advanced Learning Technologies (Niigata, Japan, July 18-20, 2007) ICALT 2007, IEEE Computer Society Press, New Jersey, NJ, 205-209

[11] Weimer, M., Gurevych, I., and Mühlhäuser, M. 2007. Automatically Assessing the Post Quality in Online Discussions on Software. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (Prague, Czech Republic, June 23-30, 2007). ACL2007 Volume P07-2, 125-128.

[12] Wu, Q., Burges, C. Svore, K, and Gao, J, 2008, Ranking, Boosting, and Model Adaptation, Technical Report, MSR-TR-2008-109, Microsoft Corporation, Redmond, WA, August 2008.

[13] Wu, Z., and Li, C., 2007. Topic Detection in Online Discussion using Non-Negative Matrix Factorization. In Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology- Workshops (Silicon Valley, CA, USA, November 2-5, 2007) WI-IATW 2007, IEEE Computer Society Press, New Jersey, NJ, 272-275