Predicting Discussions on the Social Semantic Web

Matthew Rowe, Sofia Angeletou, and Harith Alani

Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom
{m.c.rowe,s.angeletou,h.alani}@open.ac.uk

Abstract. Social Web platforms are quickly becoming the natural place for people to engage in discussing current events, topics, and policies. Analysing such discussions is of high value to analysts who are interested in assessing up-to-the-minute public opinion, consensus, and trends. However, we have a limited understanding of how content and user features can influence the amount of response that posts (e.g., Twitter messages) receive, and how this can impact the growth of discussion threads. Understanding these dynamics can help users to issue better posts, and enable analysts to make timely predictions on which discussion threads will evolve into active ones and which are likely to wither too quickly. In this paper we present an approach for predicting discussions on the Social Web, by (a) identifying seed posts, then (b) making predictions on the level of discussion that such posts will generate. We explore the use of post-content and user features and their subsequent effects on predictions. Our experiments produced an optimum F_1 score of 0.848 for identifying seed posts, and an average measure of 0.673 for Normalised Discounted Cumulative Gain when predicting discussion levels.

1 Introduction

The rise of the Social Web is encouraging more and more people to use these media to share opinions and ideas, and to engage in discussions about all kinds of topics and current events. As a consequence, the rate at which such discussions are growing, and new ones are initiated, is extremely high. The last few years have witnessed a growing demand for tools and techniques for searching and processing such online conversations to, for example; get a more up-to-date analysis of public opinion in certain products or brands; identify the main topics that the public is interested in at any given time; and gauge popularity of certain governmental policies and politicians. Furthermore, governments and businesses are investing more into using social media as an effective and fast approach for reaching out to the public, to draw their attention to new policies or products, and to engage them in open consultations and customer support discussions.

In spite of the above, there is a general lack of intelligent techniques for timely identification of which of the countless discussions are likely to gain more momentum than others. Such techniques can help tools and social media analysts overcome the great challenge of scale. For example, more than 7.4 million tweets

G. Antoniou et al. (Eds.): ESWC 2011, Part II, LNCS 6644, pp. 405–420, 2011.

on Wikileaks[1] were posted in just a few weeks. As early as 1997, Goldhaber [6] introduced the concept of *attention economics* as a way to stress the importance of engaging user attention in the new information era. However, there is a very limited understanding of the role that certain user-characteristics and content-features play in influencing the amount of response and attention generated on these social medias. Understanding the impact of such features can support interested parties in building more effective strategies for engaging with the public on social media.

In this work we are interested in identifying the characteristics of content posted on the Social Web that generate a high volume of attention - using the microblogging platform Twitter as our data source. In particular, we explore the attributes of posts (i.e., content and the author properties) that evolved into popular discussions and therefore received a lot of attention. We then investigate the use of such attributes for making predictions on discussion activity and their contributions. We also present a behaviour ontology, designed to model statistical features that our prediction techniques use from the Social Web in a common format. More precisely, we explore the following research questions: *Is it possible to identify discussion seed posts with high-levels of accuracy? What are the key features that describe seed posts?* And: *How can the level of discussion that a post will yield be predicted?* Investigation of these questions has lead to the following contributions in this paper:

1. *Identification of seed posts*: We present a classification-based method to identify discussion seed posts. Experiments are described using two corpora of tweets with features comprised from the posts and from the post authors. Analysis of such features identified key attributes that improve the likelihood of a post eliciting a response.
2. *Prediction of discussion activity levels*: We describe a ranking-based approach to predicting discussion activity levels, enabling our analysis to place posts in order of their expected discussion volume.

This paper is structured as follows: section 2 presents related work in the area of discussion and activity prediction on social media. Section 3 describes our ontology for modelling statistical features of users and posts. Section 4 presents our method for identifying discussion seed posts, and our experiments using two datasets of tweets. Section 5 describes our prediction method, the features used and our prediction experiments. Conclusions and future work are covered in section 6.

2 Background and Related Work

We define a discussion (conversation) as a chain of two or more posts, connected together to form a directed acyclic graph. Discussions begin with the publication of a *seed post*, following which users then post a *reply* to the seed, and then a

[1] http://www.guardian.co.uk/media/wikileaks

reply is posted in response, thus forming a *discussion chain*. Our work in this paper addresses the issue of predicting discussion activity on the Social Semantic Web. This entails the identification of discussion seeds as well as the response activity and volume of the conversation they initiate. To this end, the related literature can be studied in two partially overlapping research lines.

The first line concerns the topic of identifying high quality users and content on the Social Web, understanding the factors that initiate attention towards them and contribute to their popularity. Mishne and Glance [10] juxtapose the number of comments per weblog with the number of page views and incoming links, factors which constitute the weblog's popularity - where the number of comments is strongly correlated with popularity. Hsu et.al [8] present a method to identify good quality comments on Digg stories and rank them based on user features (e.g., number of posts, age in the system, number of friendships, number of profile views and topic of activity) and content features (e.g., length, informativeness, readability and complexity). Szabo and Huberman [14] use Digg and Youtube by exploiting the number of post votes as a feature to predict views of future content. Adamic et. al [1] and Bian et.al [2] use the Yahoo! question answering service to assess the quality of questions and answers and predict the best answer chosen by the question poster, where bespoke features are used (e.g., no. of best answers per user, no. of replies per question, answer length and thread length). Ratkiewicz et. al [11], study the attention towards Wikipedia pages prior and after certain events. They quantify their models using the number of clicks on a specific page. Work by Cha et. al [5] regards attention towards a user as an indicator of influence, they study users' influence on Twitter by measuring the number of followers, number of retweets and mentions. Our work extends the current state of the art by exploring new user and content features for identifying discussion seed posts. We focus on the use of Twitter as a source for our datasets, where the role of user reputation and content quality are linked together.

The second line of work concerns the identification of conversation activity on Twitter. Although a significant amount of work exists that analyses the phenomenon of microblogging in many levels, in this work we focus on the ones that study discussion activity. Only recent work by [12] has constructed conversation models using data obtained from Twitter by employing a state transition mechanism to label dialogue acts (e.g., reference broadcast, question, reaction, comment) within a discussion chain. The most relevant work to ours, by Suh et. al [13], explores the factors which lead to a post being *retweeted*, finding that the existence of a hashtag or URL in the original post does not affect its retweeting chance, while the number of followers and followed does. Our work differs from [13] by identifying discussions conducted, as opposed to identifying information spread, and unlike existing work by [12] we extend our scope to discussions, rather than merely considering interactions of replies between 2 parties. In doing so we identify discussion seed posts and the key features that lead to starting a discussion and generating attention. To the best of our knowledge there is no existing work that can suggest and predict if and to what extent a post will be responded to.

3 Behaviour Ontology

In the context of our work - predicting discussions - we rely on statistical features of both users and posts, which we describe in greater detail in the following section. No ontologies at present allow the capturing of such information, and its description using common semantics. To fill this gap we have developed a behaviour ontology,[2] closely integrated with the Semantically Interlinked Online Communities (SIOC) ontology [3], this enables the modelling of various user activities and and related impacts on a Social Networking Site (SNS). The behaviour ontology represents posts on social networking sites, the content of those posts, the sentiment of the content and the impact which a post has had on the site. It also extends existing SIOC concepts, in particular *sioc:UserAccount*, with information about the impact the user has had on the site by capturing the number of followers and friends the user has at a given time (represented with the Data property *CollectionDate*). For the sake of brevity and space restrictions Figure 1 shows a part of the ontology that is relevant to the representation of the information used in our approach.

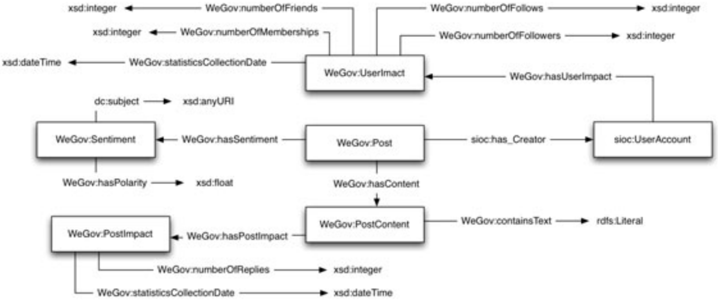


Fig. 1. The Behaviour Ontology

The two key classes presented in Figure 1 are: *PostImpact* and *UserImpact*. The former models the number of replies that a given post has generated, characterising the level of discussion that the post has yielded up until a given point in time. The latter class models the impact that the user has had within a given SNS. Capturing this information is crucial to predicting discussion activity, as according to [8,13] user reputation and standing within an online space is often a key factor in predicting whether content will generate attention or not.

4 Identifying Discussion Seeds

Identifying seed posts prior to such posts receiving a reply allows discussions to be pre-empted and tracked accordingly. We define a seed post as a post on a

[2] `http://people.kmi.open.ac.uk/rowe/ontologies/UserOnto_0.23.rdf`

given Social Web platform that will yield at least one reply - within this paper we concentrate on the use of tweets, therefore a seed post is regarded as the initial tweet that generates a reply. Features which describe seed posts can be divided into two sets: *user features* - attributes that define the user making the post; and, *content features* - attributes that are based solely on the post itself. We wish to explore the application of such features in identifying seed posts, to do this we train several machine learning classifiers and report on our findings. However we first describe the features used.

4.1 Feature Extraction

The likelihood of posts eliciting replies depends upon popularity, a highly subjective term influenced by external factors. Properties influencing popularity include user attributes - describing the reputation of the user - and attributes of a post's content - generally referred to as content features. In Table 1 we define user and content features and study their influence on the discussion "continuation".

Table 1. User and Content Features

	User Features	
In Degree:	Number of followers of U	#
Out Degree:	Number of users U follows	#
List Degree:	Number of lists U appears on. *Lists* group users by topic	#
Post Count:	Total number of posts the user has ever posted	#
User Age:	Number of minutes from user join date	#
Post Rate:	Posting frequency of the user	$\frac{PostCount}{UserAge}$
	Content Features	
Post length:	Length of the post in characters	#
Complexity:	Cumulative entropy of the unique words in post p λ of total word length n and pi the frequency of each word	$\frac{\sum_{i\in[1,n]} pi(\log\lambda - \log pi)}{\lambda}$
Uppercase count:	Number of uppercase words	#
Readability:	Gunning fog index using average sentence length (ASL) and the percentage of complex words (PCW).	[7] $0.4(ASL + PCW)$
Verb Count:	Number of verbs	#
Noun Count:	Number of nouns	#
Adjective Count:	Number of adjectives	#
Referral Count:	Number of @user	#
Time in the day:	Normalised time in the day measured in minutes	#
Informativeness:	Terminological novelty of the post wrt other posts The cumulative tfIdf value of each term t in post p	$\sum_{t\in p} tfidf(t,p)$
Polarity:	Cumulation of polar term weights in p (using Sentiwordnet[3] lexicon) normalised by polar terms count	$\frac{Po+Ne}{\|terms\|}$

4.2 Experiments

Experiments are intended to test the performance of different classification models in identifying seed posts. Therefore we used four classifiers: discriminative classifiers Perceptron and SVM, the generative classifier Naive Bayes and the decision-tree classifier J48. For each classifier we used three feature settings: user features, content features and user+content features.

[3] `http://sentiwordnet.isti.cnr.it/`

Datasets. For our experiments we used two datasets of tweets available on the Web: Haiti earthquake tweets[4] and the State of the Union Address tweets.[5] The former dataset contains tweets which relate to the Haiti earthquake disaster - tagged with *#haiti* - covering a varying timespan. The latter dataset contains all tweets published during the duration of president Barack Obama's State of the Union Address speech. Our goal is to predict discussion activity based on the features of a given post by first identifying seed posts, before moving on to predict the discussion level.

Within the above datasets many of the posts are not seeds, but are instead replies to previous posts, thereby featuring in the discussion chain as a node. In [13] *retweets* are considered as part of the discussion activity. In our work we identify discussions using the explicit "*in reply to*" information obtained by the Twitter API, which does not include retweets. We make this decision based on the work presented in boyd et.al [4], where an analysis of retweeting as a discussion practice is presented, arguing that message forwards adhere to different motives which do not necessarily designate a response to the initial message. Therefore, we only investigate explicit replies to messages. To gather our discussions, and our seed posts, we iteratively move up the reply chain - i.e., from reply to parent post - until we reach the seed post in the discussion. We define this process as *dataset enrichment*, and is performed by querying Twitter's REST API[6] using the *in_reply_to_id* of the parent post, and moving one-step at a time up the reply chain. This same approach has been employed successfully in work by [12] to gather a large-scale conversation dataset from Twitter.

Table 2. Statistics of the datasets used for experiments

Dataset	Users	Tweets	Seeds	Non-Seeds	Replies
Haiti	44,497	65,022	1,405	60,686	2,931
Union Address	66,300	80,272	7,228	55,169	17,875

Table 2 shows the statistics that explain our collected datasets. One can observe the difference in conversational tweets between the two corpora, where the Haiti dataset contains fewer seed posts as a percentage than the Union dataset, and therefore fewer replies. However, as we explain in a later section, this does not correlate with a higher discussion volume in the former dataset. We convert the collected datasets from their proprietary JSON formats into triples, annotated using concepts from our above behaviour ontology, this enables our features to be derived by querying our datasets using basic SPARQL queries.

Evaluation Measures. The task of identifying seed posts is a binary classification problem: *is this post a seed post or not?* We can therefore restrict our

[4] `http://infochimps.com/datasets/twitter-haiti-earthquake-data`

[5] `http://infochimps.com/datasets/tweets-during-state-of-the-union-address`

[6] `http://dev.twitter.com`

labels to one of two classes: *seed* and *non-seed*. To evaluate the performance our method we use four measures: precision, recall, f-measure and area under the Receiver Operator Curve. Precision measures the proportion of retrieved posts which were actually seed posts, recall measures the proportion of seed posts which were correctly identified and fallout measures the proportion of non-seed posts which were incorrectly classified as seed posts (i.e., *false positive rate*). We use f-measure, as defined in Equation 1 as the harmonic mean between precision and recall, setting $\beta = 1$ to weight precision and recall equally. We also plot the Receiver Operator Curve of our trained models to show graphical comparisons of performance.

$$F_\beta = \frac{(1+\beta^2) * P * R}{\beta^2 * P + R} \tag{1}$$

For our experiments we divided each dataset up into 3 sets: a training set, a validation set and a testing set using a 70/20/10 split. We trained our classification models using the training split and then applied them to the validation set, labelling the posts within this split. From these initial results we performed *model selection* by choosing the best performing model - based on maximising the F_1 score - and used this model together with the best performing features, using a ranking heuristic, to classify posts contained within our test split. We first report on the results obtained from our model selection phase, before moving onto our results from using the best model with the top-k features.

Table 3. Results from the classification of seed posts using varying feature sets and classification models

(a) Haiti Dataset

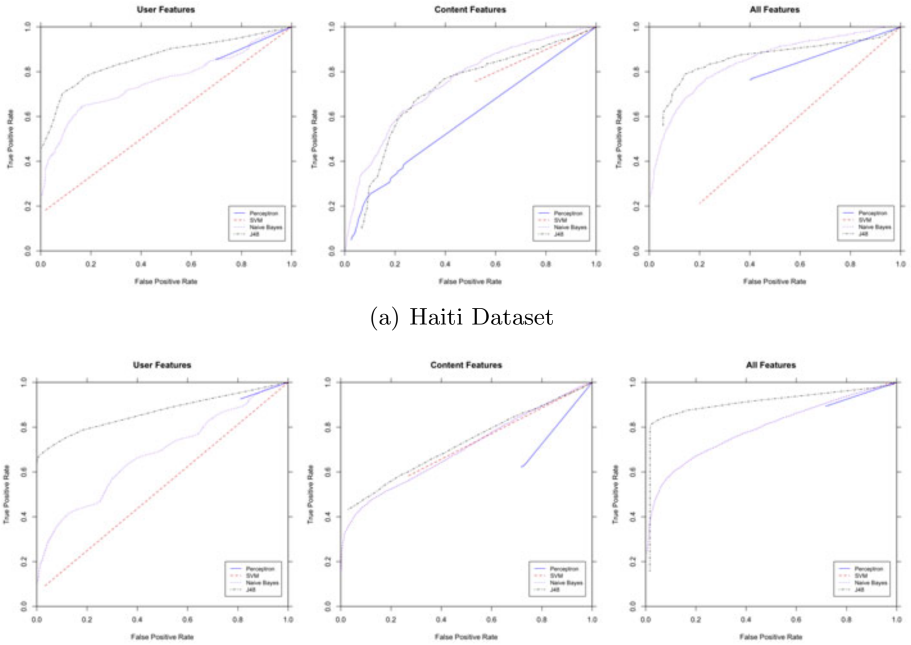
		P	R	F_1	ROC
User	Perc	0.794	0.528	0.634	0.727
	SVM	0.843	0.159	0.267	0.566
	NB	**0.948**	0.269	0.420	0.785
	J48	0.906	**0.679**	**0.776**	**0.822**
Content	Perc	**0.875**	0.077	0.142	0.606
	SVM	0.552	**0.727**	0.627	0.589
	NB	0.721	0.638	0.677	**0.769**
	J48	0.685	0.705	**0.695**	0.711
All	Perc	0.794	0.528	0.634	0.726
	SVM	0.483	**0.996**	0.651	0.502
	NB	**0.962**	0.280	0.434	**0.852**
	J48	0.824	0.775	**0.798**	0.836

(b) Union Address Dataset

		P	R	F_1	ROC
User	Perc	0.658	0.697	0.677	0.673
	SVM	0.510	**0.946**	0.663	0.512
	NB	0.844	0.086	0.157	0.707
	J48	**0.851**	0.722	**0.782**	**0.830**
Content	Perc	0.467	**0.698**	0.560	0.457
	SVM	0.650	0.589	0.618	0.638
	NB	**0.762**	0.212	0.332	0.649
	J48	0.740	0.533	**0.619**	**0.736**
All	Perc	0.630	0.762	0.690	0.672
	SVM	0.499	**0.990**	0.664	0.506
	NB	0.874	0.212	0.341	0.737
	J48	**0.890**	0.810	**0.848**	**0.877**

4.3 Results

Our findings from Table 3 demonstrate the effectiveness of using solely user features for identifying seed posts. In both the Haiti and Union Address datasets training a classification model using user features shows improved performance over the same models trained using content features. In the case of the Union dataset we are able to achieve an F_1 score of 0.782, coupled with high precision, when using the J48 decision-tree classifier - where the latter figure (precision) indicates conservative estimates using only user features. We also achieve similar high-levels of precision when using the same classifier on the Haiti dataset. The plots of the Receiver Operator Characteristic (ROC) curves in Figure 2 show similar levels of performance for each classifier over the two corpora.When using solely user features J48 is shown to dominate the ROC space, subsuming the plots from the other models. A similar behaviour is exhibited for the Naive Bayes classifier where SVM and Perceptron are each outperformed. The plots also demonstrate the poor recall levels when using only content features, where each model fails to yield the same performance as the use of only user features. However the plots show that effectiveness of combining both user and content features.



(a) Haiti Dataset

(b) Union Address Dataset

Fig. 2. ROC Curves for Classification Models with differing Feature Sets

Experiments identify the J48 classifier as being our best performing model yielding optimum F_1 scores and by analysing the induced decision tree we observe the affects of individual features. Extremes of post polarity are found to be good indicators of seed posts, while posts which fall within this mid-polarity range are likely to be objective. One reason for this could be that the posts which elicit an emotional response are more likely to generate a reply. Analysis of the time of day identifies 4pm to midnight and 3pm to midnight as being associated with seed posts for the Haiti and Union Address datasets respectively.

Top-*k* Feature Selection. Thus far we have only analysed the use of features grouped together prompting questions: *which features are more important than others?* and *what features are good indicators of a seed post?* To gauge the importance of features in identifying seed posts we rank our features by their Information Gain Ratio (IGR) with respect to seed posts. Our rankings in Table 4 indicate that the number of lists that a user is featured in appears in the first position for both the Haiti and Union Address datasets, and the in-degree of the user also features towards the top of each ranking. Such features increase

Table 4. Features ranked by Information Gain Ratio wrt Seed Post class label. The feature name is paired within its IG in brackets.

Rank	Haiti	Union Address
1	user-list-degree (0.275)	user-list-degree (0.319)
2	user-in-degree (0.221)	content-time-in-day (0.152)
3	content-informativeness (0.154)	user-in-degree (0.133)
4	user-num-posts (0.111)	user-num-posts (0.104)
5	content-time-in-day (0.089)	user-post-rate (0.075)
6	user-post-rate (0.075)	user-out-degree (0.056)
7	content-polarity (0.064)	content-referral-count (0.030)
8	user-out-degree (0.040)	user-age (0.015)
9	content-referral-count (0.038)	content-polarity (0.015)
10	content-length (0.020)	content-length (0.010)
11	content-readability (0.018)	content-complexity (0.004)
12	user-age (0.015)	content-noun-count (0.002)
13	content-uppercase-count (0.012)	content-readability (0.001)
14	content-noun-count (0.010)	content-verb-count (0.001)
15	content-adj-count (0.005)	content-adj-count (0.0)
16	content-complexity (0.0)	content-informativeness (0.0)
17	content-verb-count (0.0)	content-uppercase-count (0.0)

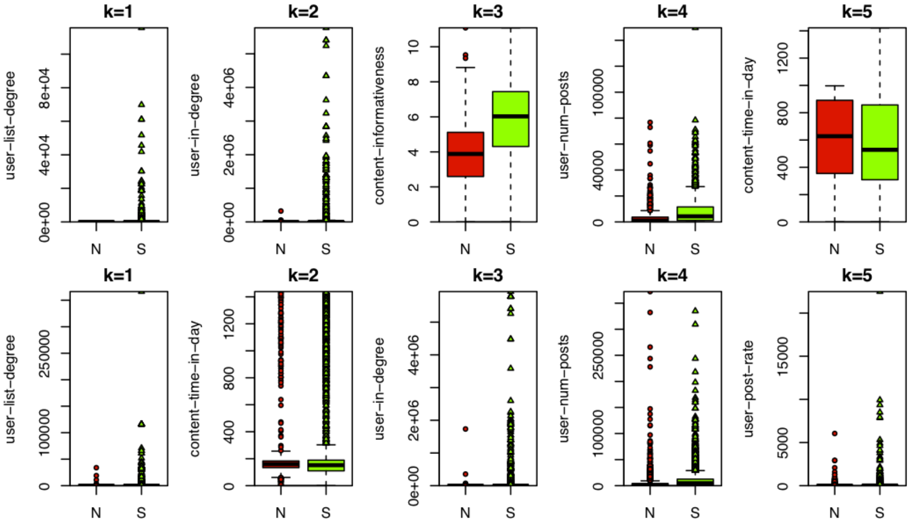
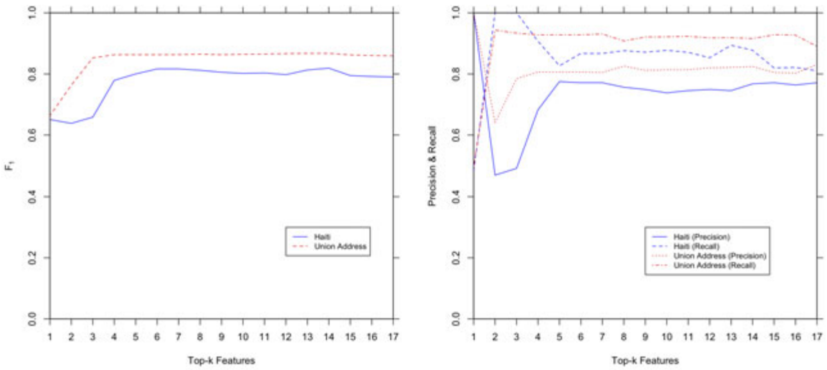


Fig. 3. Contributions of top-5 features to identifying Non-seeds (N) and Seeds (S). Upper plots are for the Haiti dataset and the lower plots are for the Union Address dataset.

the broadcast capability of the user, where any post made by the user is read by a large audience, increasing the likelihood of yielding a response. To gauge the similarity between the rankings we measured the Pearson Correlation Coefficient, which we found to be 0.674 indicating a good correlation between the two lists and their respective ranks.

The top-most ranks from each dataset are dominated by user features including the list-degree, in-degree, num-of-posts and post-rate. Such features describe a user's reputation, where higher values are associated with seed posts. Figure 3 shows the contributions of each of the top-5 features to class decisions in the training set, where the list-degree and in-degree of the user are seen to correlate heavily with seed posts. Using these rankings our next experiment explored the effects of training a classification model using *only* the top-k features, observing the affects of iteratively increasing k and the impact upon performance. We selected the J48 classifier for training - based on its optimum performance during the model selection phase - and trained the classifier using the training split from each dataset and only the top-k features based on our observed rankings. The model was then applied to the held out test split of 10%, thereby ensuring independence from our previous experiment.



(a) F_1 scores

(b) Precision and Recall

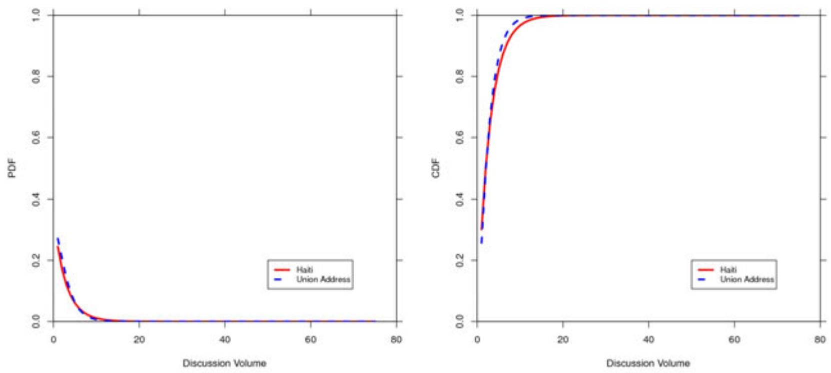
Fig. 4. Classification using J48 and top-k features ranked by Information Gain Ratio

Figure 4 shows the results from our experiments, where at lower levels of k we observe similar levels of performance - particularly when only the highest ranked feature is used (i.e., list-degree). As we increase k, including more features within our classification model, we observe improvements in F_1 scores for both datasets. The lower ranks shown in Table 4 are dominated by content features. As we include the lower-ranked features, our plots show a slight decrease in performance for the Haiti dataset due to the low IGR scores yielded for such features. For both datasets we yield 1 for precision when each model is trained using just the list-degree of the user, although J48 induces different cutoff points for the two datasets when judging the level of this nominal value. Using this feature provides a conservative estimate of seed posts, highlighted by the correlation in our training split in Figure 3. Such findings are consistent with work by [14] which found a *rich get richer* phenomena, where popular users - with a large audience of listeners/followers/observers - would yield feedback, and in doing so increase their in-degree: as the followers of their followers observed replies.

The results from our experiments have empirically demonstrated the affects of using user features in comparison with content features for identifying seed posts. Ranking features by their Information Gain Ratio also explains the importance of features, where we observe the effectiveness of including user features in identifying seed posts. However the role of content features also plays an important part, where the inclusion of the correct time of day to make a post impacts upon the likelihood of the post yielding a reply and starting a discussion.

5 Predicting Discussion Activity Levels

Identifying seed posts and the features which are commonly found within such posts, enables policy makers, and regular web users alike, to improve their posts or role in a community to ensure a reply and start a discussion. Identifying seed posts before they receive a reply also enables the tracking of key posts which are likely to yield a discussion - rather than monitoring a wide range of posts, some of which will lead to a discussion, while others may not. A natural progression of the identification of seed posts is it to predict the level of discussion that a seed post will generate. In doing so we can identify the optimum features that a given post, and its user, must have in order to maximise impact, generate a response and lead to a lengthy discussion. Predicting an exact number of replies that a given post will generate is not a straightforward task. Analysis of the training splits in our tweet datasets (Haiti and Union Address) identifies a large range in the discussion volumes and the distribution of these levels: in the Haiti dataset we found that the discussion volume ranged from 1 reply through to 74, and for the Union Address dataset we found similar levels: 1 post to 75.



(a) Probability distribution functions (b) Cumulative distribution functions

Fig. 5. Distribution functions of the Haiti and Union Address datasets

More in-depth analysis of the data is shown in Figure 5(a) and Figure 5(b), displaying the probability distributions and cumulative distributions respectively. For each dataset we used maximum likelihood estimation to optimise parameters for various distribution models, and selected the best fitting model using the Kolmogorov-Smirnov goodness-of-fit test against the training splits from our datasets. For the Haiti dataset we fitted the Gamma distribution - found to be a good fit at $\alpha = 0.1$, and for the Union dataset we fitted the Chi-squared distribution - however this was found to provide the minimum deviation from the data without satisfying any of the goodness of fit tests. The distributions convey, for both fitted datasets, that the probability mass is concentrated towards the head of the distribution where the volume of the discussion is at its lowest. The likelihood of a given post generating many replies - where many can be gauged as the mean number of replies within the discussion volume distribution - tends to 0 as the volume increases. Such density levels render the application of standard prediction error measures such as Relative Absolute Error inapplicable, given that the mean of the volumes would be used as the random estimator for accuracy measurement. A solution to this problem is instead to assess whether one post will generate a larger discussion than another, thereby producing a ranking, similar to the method used in [8].

To predict the discussion activity level we use a Support Vector Regression model trained using the three distinct feature set combinations that we introduced earlier in this paper: user features, content features and user+content features. Using the predicted values for each post we can then form a ranking, which is comparable to a ground truth ranking within our data. This provides *discussion activity levels*, where posts are ordered by their expected volume. This approach also enables contextual predictions where a post can be compared with existing posts that have produced lengthy debates.

5.1 Experiments

Datasets. For our experiments we used the same datasets as in the previous section: tweets collected from the Haiti crisis and the Union Address speech. We maintain the same splits as before - training/validation/testing with a 70/20/10 split - but without using the test split. Instead we train the regression models using the seed posts in the training split and then test the prediction accuracy using the seed posts in the validation split - seed posts in the validation set are identified using the J48 classifier trained using both user+content features. Table 5 describes our datasets for clarification.

Table 5. Discussion volumes and distributions of our datasets

Dataset	Train Size	Test Size	Test Vol Mean	Test Vol SD
Haiti	980	210	1.664	3.017
Union Address	5,067	1,161	1.761	2.342

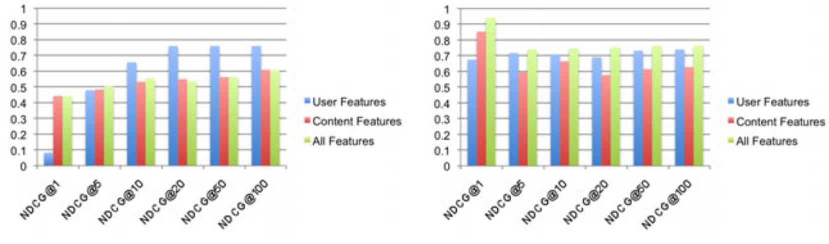
Evaluation Measures. To assess the accuracy of learnt regression models we compare the predicted rank from each model against the actual rank within our datasets - given that we have measured the discussion volume when collecting the data. Our intuition is that certain posts will attract a larger discussion volume than others, where the preference between posts based on such volume measures will enable a comparison of the ranking against a ground truth ranking based on the actual discussion volumes. To compare rankings we use the Normalised Discounted Cumulative Gain measure [9] for the top-k ranked elements, defined as $NDCG_k = DCG_k/iDCG_k$. This divides the Discounted Cumulative Gain (DCG) derived from the predicted rank against the actual rank defined as iDCG above. DCG is an empirical measure that is tailored towards rewarding rankings where the top-most elements in the ground truth are found within the same position in the predicted rank. This is motivated by web search scenarios where end users wish to find the most important documents on the first page of search results. We have a similar motivation given that we wish to identify those seed posts which yield the largest discussions and should therefore appear at the top of our ranking. We formalise DCG as:

$$DCG_k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(1+i)} \tag{2}$$

In order to define rel_i we use the same approach as [8]: $rel_i = N - rank_i + 1$, where $rank_i$ is the ground truth rank of the element at index i from the predicted ranking. Therefore when dividing the predicted rank by the actual rank, we get a normalised value ranging between 0 and 1, where 1 defines the predicted rank as being equivalent to the actual rank. To provide a range of measures we calculated $NDCG@k$ for 6 different values where $k = \{1, 5, 10, 20, 50, 100\}$, thereby assessing the accuracy of our rankings over different portions of the top-k posts. We learnt a Support Vector Regression model for each dataset using the same feature sets as our earlier identification task: user features, content features and user+content features.

5.2 Results

Figure 6 shows the ranking accuracy that we achieve using a Support Vector Regression model for prediction over the two datasets, where we observe differing performance levels achieved using different feature set combinations. For the Haiti dataset the user features play a greater role in predicting discussion activity levels for larger values of k. For the Union Address dataset user features also outperform content features as k increases. In each case we note that content features do not provide as accurate predictions as the use of solely user features. Such findings are consistent with experiments described in [8] which found that user features yielded improved ranking of comments posted on Stories from Digg in comparison with merely content features.



(a) Haiti (b) Union Address

Fig. 6. Predicted seed post ranking using Support Vector Regression compared with ground truth ranking using $NCDG@k$

Following on from our initial rank predictions we identify the user features as being important predictors of discussion activity levels. By performing analysis of the learnt regression model over the training split we can analyse the coefficients induced by the model - Table 6 present the coefficients learnt from the user features. Although different coefficients are learnt for each dataset, for major features with greater weights the signs remain the same. In the case of the list-degree of the user, which yielded high IGR during classification, there is a similar positive association with the discussion volume - the same is also true for the in-degree and the out-degree of the user. This indicates that a constant increase in the combination of a user's in-degree, out-degree, and list-degree will lead to increased discussion volumes. Out-degree plays an important role by enabling the seed post author to see posted responses - given that the larger the out-degree the greater the reception of information from other users.

Table 6. Coefficients of user features learnt using Support Vector Regression over the two Datasets. Coefficients are rounded to 4 dp.

	user-num-posts	user-out-degree	user-in-degree	user-list-degree	user-age	user-post-rate
Haiti	-0.0019	+0.001	+0.0016	+0.0046	+0.0001	+0.0001
Union	-0.0025	+0.0114	+0.0025	+0.0154	-0.0003	-0.0002

6 Conclusions

The abundance of discussions carried out on the Social Web hinders the tracking of debates and opinion, while some discussions may form lengthy debate others may simply die out. Effective monitoring of high activity discussions can be solved by predicting which posts will start a discussion and their subsequent discussion activity levels. In this paper we have explored three research questions, the first of which asked *Is it possible to identify discussion seed posts with high-levels of accuracy?* We have presented a method to identify discussion seed posts achieving an optimum F_1 score of 0.848 for experiments over one dataset. Experiments with both content and user features demonstrated the importance of the user's reputation in eliciting a response. We sought further analyses of individual features - exploring *What are the key features that describe seed posts?* - and identified the importance of users' list-degree and in-degree: the former measuring the number of subscription channels that a given user has been added to and the latter defining the number of people subscribing to the user.

Following identification of seed posts, we then investigated: *How can the level of discussion that a post will yield be predicted?* Implementing a Support Vector Regression model produced a ranking of seed posts ordered by expected discussion activity from high to low, achieving an average measure of 0.673 for Normalised Discounted Cumulative Gain using user features. Creation of such rankings allows seed posts to be compared against other posts already generating lengthy discussions. Furthermore, through our use of regression models we are able to alter the predicted feature, enabling the time of day to publish a post to be predicted in order to maximise response.

The prediction techniques that we have outlined within this paper are designed to work over data obtained from disparate data sources. Our future work will explore the adaptation of induced models across domains and platforms and look to leverage overriding patterns for later application. Such a crossover however is not possible without a common understanding of information across such sources, therefore in this paper we have presented a behaviour ontology - built as an extension to SIOC - which allows the necessary features to be captured using common semantics from disparate domains and platforms. We will also explore the notion of *user-dependent post topic entropy*, where the specialisation of a user is captured through a topic distribution. Our intuition is that incurring followers is dependent on certain topics, where the publication of posts that cite such topics are more likely to elicit a response.

Acknowledgements

This work is funded by the EC-FP7 projects WeGov (grant number 248512) and Robust (grant number 257859).

References

1. Adamic, L.A., Zhang, J., Bakshy, E., Ackerman, M.S.: Knowledge sharing and Yahoo Answers: Everyone knows something. In: Proceedings of WWW 2008, pp. 665–674. ACM, New York (2008)
2. Bian, J., Liu, Y., Zhou, D., Agichtein, E., Zha, H.: Learning to Recognize Reliable Users and Content in Social Media with Coupled Mutual Reinforcement. In: 18th International World Wide Web Conference (WWW 2009) (April 2009)
3. Bojars, U., Breslin, J.G., Peristeras, V., Tummarello, G., Decker, S.: Interlinking the social web with semantics. IEEE Intelligent Systems 23, 29–40 (2008)
4. Boyd, D., Golder, S., Lotan, G.: Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In: Hawaii International Conference on System Sciences, pp. 1–10 (2010)

5. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring User Influence in Twitter: The Million Follower Fallacy. In: Fourth International AAAI Conference on Weblogs and Social Media (May 2010)
6. Goldhaber, M.H.: The Attention Economy and the Net. First Monday 2(4) (1997)
7. Gunning, R.: The Technique of Clear Writing. McGraw-Hill, New York (1952)
8. Hsu, C.-F., Khabiri, E., Caverlee, J.: Ranking Comments on the Social Web. In: International Conference on Computational Science and Engineering, CSE 2009, vol. 4 (August 2009)
9. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. 20, 422–446 (2002)
10. Mishne, G., Glance, N.: Leave a Reply: An Analysis of Weblog Comments. In: Third annual workshop on the Weblogging ecosystem (2006)
11. Ratkiewicz, J., Menczer, F., Fortunato, S., Flammini, A., Vespignani, A.: Characterizing and modeling the dynamics of online popularity. Physical Review Letters (May 2010)
12. Ritter, A., Cherry, C., Dolan, B.: Unsupervised Modeling of Twitter Conversations. In: Proc. HLT-NAACL 2010 (2010)
13. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In: Proceedings of the IEEE Second International Conference on Social Computing (SocialCom), pp. 177–184 (August 2010)
14. Szabo, G., Huberman, B.A.: Predicting the popularity of online content. ACM Commun. 53(8), 80–88 (2010)