2015

A Classification Based Framework to Predict Viral Threads

Hashim Sharif
Lahore University of Management Sciences, hashim@lums.edu.pk

Saad Ismail
Lahore University of Management Sciences, saad@lums.edu.pk

Shehroze Farooqi
Lahore University of Management Sciences, shehroze@lums.edu.pk

Mohammad Taha Khan
Lahore University of Management Sciences, taha.khan@lums.edu.pk

Muhammad Ali Gulzar
Lahore University of Management Sciences, aligulzar@lums.edu.pk

See next page for additional authors

Authors
Hashim Sharif, Saad Ismail, Shehroze Farooqi, Mohammad Taha Khan, Muhammad Ali Gulzar, Hasnain Lakhani, Fareed Zaffar, and Ahmed Abbasi

A CLASSIFICATION BASED FRAMEWORK TO PREDICT VIRAL THREADS

Complete Research

Hashim Sharif, LUMS, Lahore, Pakistan, hashim@lums.edu.pk
Saad Ismail, LUMS, Lahore, Pakistan, saad@lums.edu.pk
Shehroze Farooqi, LUMS, Lahore, Pakistan, shehroze@lums.edu.pk
Mohammad Taha Khan, LUMS, Lahore, Pakistan, taha.khan@lums.edu.pk
Muhammad Ali Gulzar, LUMS, Lahore, Pakistan, aligulzar@lums.edu.pk
Hasnain Lakhani, SRI International, Menlo Park, CA, USA, mhlakhani@sri.com
Fareed Zaffar, LUMS, Lahore, Pakistan, fareed.zaffar@lums.edu.pk
Ahmed Abbasi, University of Virgina, USA, ana6e@comm.virginia.edu

Abstract

Online social media allows consumers to engage with each other and to create, share, discuss and modify user-generated content in a highly interactive way. Social media platforms have therefore become critical for companies trying to gauge the pulse of consumers, help identify issues faster, receive immediate feedback on products and offering etc. An effective social media strategy therefore requires companies to mine large volumes of structured unstructured and semi-structured online textual data in order to gain insights into the underlying traits of the consumers and prevailing public opinion. These insights can provide opportunities for market research, protection of brand reputation and a mechanism to gauge user preferences in an attempt to maximize customer satisfaction and consumer-brand engagement.
In this paper, we propose and evaluate a classification based framework to predict thread lengths in online discussion forums in order to identify potential topics that may of interest to a particular online community. We identify and evaluate several key features of viral social media conversations through extensive experiments conducted on health 2.0 datasets. We also present a pharmaceutical industry based case study to illustrate how well the viral thread topics relate to real world events.

Keywords: Threads, Social Networks, Predictive Analysis.

1 Introduction

Online social media has been rapidly changing the face of brand-consumer engagement. Typical Internet users spend a lot more time on social media, with the result that a significant portion of brand engagement has moved online. Typical consumers these days often use facebook, twitter, online forums, blogs, micro-blogs etc. as their first source of information when doing research on products, reviews , diseases, drugs etc. Many such sources of information are organized as a series of discussion threads where people are able to create, share, discuss and modify user-generated content through conversational interactions. Social media platforms have therefore become critical for companies trying to gauge the pulse of consumers, help identify issues faster, receive immediate feedback on products and offering etc. Such insights are especially valuable for corporations that rely on customer centric innovation as their competitive advantage. 'Social intelligence' as this is sometimes referred to, is the process of extracting events, relationships and contextual knowledge from online social media in order to gain insights into the underlying traits of the consumers and prevailing public opinion. These insights can provide opportunities for market research, protection of brand reputation and a mechanism to gauge user preferences in an attempt to maximize customer satisfaction. Furthermore, identifying trending topics and consumer trends can help an organization provide innovative products and services.

Pharmaceutical industry is one such example where companies have successfully mined large volumes of structured unstructured and semi-structured online textual data in order to perform post marketing drug surveillance and predict adverse drug reactions (Adjeroh 2014; Fu 2012). Previous work in this area suggests that the early prediction of adverse drug reactions for instance can be essential in preventing substantial exposure and financial damage for the pharmaceutical companies (Abbasi and Adjeroh 2014). Since time-sensitive and actionable insights can potentially be gleaned from this data, there is a lot of interest among organizations to try and predict what has the potential of going viral on the web. Continuing with the pharmaceutical industry as an example, several studies have applied sentiment analysis on online medical content to predict public opinions regarding various drugs, which are indicative of future adverse drug reactions (Sharif 2014). As a result, predicting such information in advance, can serve as an early warning for organizations who can then try to remedy a situation before it blows up into a full scale disaster. The results clearly indicate that hundreds of millions of dollars can potentially be saved if such information can be detected early enough.

Viral trends in online social media possess two essential characteristics; volume and velocity. A forum thread, Facebook discussion or a Twitter feed that contains a lot of messages typically signifies large user involvement from the community. Hence, longer threads are suggestive of an emerging topic that seizes the attention of the online community. The same is true for online forums where often dissatisfied customers of a certain product may lead to a lot of traffic being generated for a particular thread etc. (Backstrom 2013) have also motivated the problem of thread length prediction as a means of estimating how interesting a thread is. In addition to volume, velocity or frequency of user interactions is also a key characteristic of long conversational discussions. High velocity is a direct implication of keen interest directed towards the thread topic, thus usually yields long discussions.

In this paper, we propose a classification based framework to predict thread lengths in online discussion forums in order to identify potential topics that may of interest to a particular online community. Our methodology employs diverse feature representations that are better able to correlate with the eventual thread length, thus yielding accurate prediction results. Velocity and participant features being representative of longer threads are also incorporated into the system. We evaluate our model with extensive experiments on Health and Telco related forum threads. Results reveal that the proposed thread length prediction framework outperforms the approach proposed by (Backstrom 2013). in terms of prediction accuracy. We also present results from a pharmaceutical industry based case study to illustrate how well the viral thread topics relate to the real world events. The remainder of the paper is organized as follows: Section 2 presents prior work on thread length prediction, and emphasizes the research gaps. Section 6 describes the proposed thread length prediction framework. Section 4 describes the experimental

evaluation before concluding remarks in section 5.

2 Related Work

Much of social intelligence research has focused on sentiment analysis as a quick and efficient way of providing consumer feedback to interested entities. SentiStrength, for instance, a popular stand-alone sentiment analysis tool uses a sentiment lexicon for assigning scores to negative and positive phrases in text. Phrase level scores are aggregated to determine sentence level polarities. Other approaches have focused on detecting topics and genre from a body of text. Others still have tried to discover relationships, top influencers and events from structured and semi-structured data.

(Backstrom 2013) formally introduced thread length prediction as a key problem in conversational curation. and conceptualized thread length prediction as a means of estimating the interest a conversation generates. This in turn can assist in the selection of threads that need to be brought to the attention of online users at any given point in time. Using an initial set of comments in a thread, they train a binary classifier to predict the eventual length of a thread to be above or below a threshold. Their study demonstrates strong correlation between thread-length and contextual features such as interactions between posters, comment arrival patterns, comment timestamps and text regression features etc. Significant work on evaluating quality discourse in online news comments also comes from (Diakopoulos 2011) where the authors use reader surveys etc to examine the needs and desires of news commenters and provide a description of both readers' and writers' motivation for usage of news comments. We feel the quality of our analysis of potentially viral events can be evaluated by some of the approaches adopted by this study. (Shah 2010) present a framework to evaluate and predict the quality of an answer in an online community QA setting. Quality predictions based on features derived through manual labeling of data showed a highly accurate prediction of actual ratings. The results are significant since they clearly demonstrate that contextual information such as user's profile can be critical in evaluating and predicting content quality. Similarly, (Adamic 2008) aim to categorize and cluster question-answer forum data according to several content characteristics and patterns of interactions among users. Result demonstrate that while some users focus narrowly on specific topics, others participate across many categories. Combining both user attributes and answer characteristics they predict, within a given category, whether a particular answer will be chosen as the best answer by the asker.

A similar study, by (Wanas 2008) focuses on automatically rating posts in online discussion forums. They use features from the post content and the thread structure to derive features for each posting. Using a nonlinear SVM classifier, the value of each posting is categorized either as high, medium or low. It is important to note that, the task of rating forum postings is similar to predicting thread length, particularly because a highly rated post is also likely to go viral on the web. Hence, the features extracted for rating forum postings are also relevant in the context of thread length estimation.

Our approach derives similar features of viral threads through an analysis of a large corpus of forum threads and relies on several of these content and contextual features for high predictability. Our work also highlights a few other issues that we aim to address in our approach:

- User interactions in forums and blogs are modeled on a number of different factors. Hence, in addition to the feature sets previously explored, there is a need for incorporating additional feature sets that have a logical correspondence with the growth of conversation threads.
- Conversational threads are growing body of messages, which can be conceptualized as a continuous stream of data. Previous approach such as (Wanas 2008) utilize a specified number of comments to train a binary classifier, that is used for predicting the thread length to be above or below a predefined threshold. As a consequence, the technique fails to model the growing behavior of online conversations. Furthermore, the values selected for the number of input comments and the total comments to be predicted are arbitrary and do not take into account the distribution of the thread lengths in the data.

3 Approach

The framework schematic of flow diagram for our approach is shown in Figure 1.

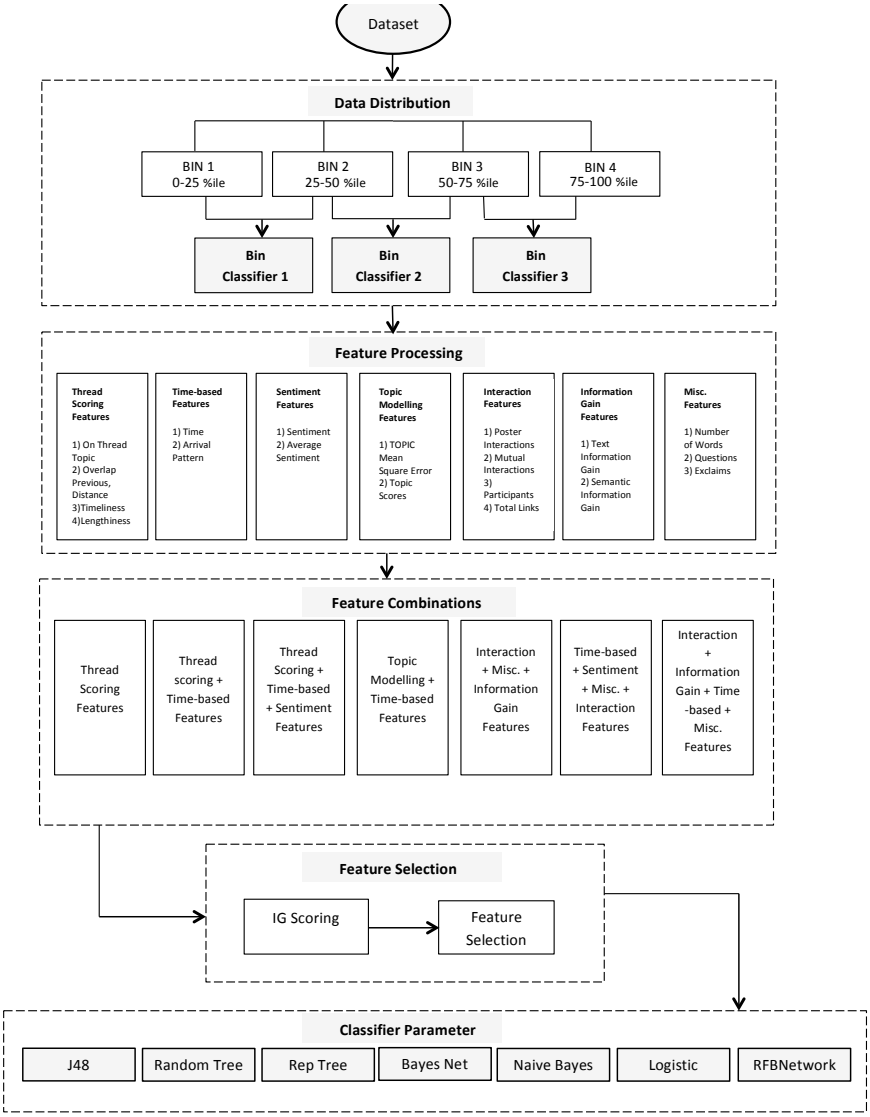

Figure 1: Schematic of thread length prediction framework

In this section we propose our thread length prediction framework that addresses the challenges highlighted in the previous section. The approach employs a number of new feature representations in addition to the features used by earlier approaches. Furthermore, it is able to adapt to the growing nature of online conversations. The feature categories used in the work of (Backstrom 2013) are broadly categorized as link features, comment arrival patterns, time-based features, text features and miscellaneous features. Though these features are able to capture important characteristics of conversation dynamics, there is a need to incorporate further parallel feature representations, so as to model behavior unexplained with the existing feature sets. These additional feature sets can be broadly categorized into four feature categories namely sentiment features, topic modeling features, semantic features and thread scoring features. Notably, the application of thread scoring features has been motivated by the work of (Wanas 2008). The description of

all the adopted thread scoring metrics will be detailed in a later subsection. Moreover, sentiment, semantic and topic modeling features are added in order to extract the underlying moods, topical features and common abstract entities that characterize conversations; hence they can potentially serve as important determiners of viral trends. It is important to mention that, the task of length prediction is positioned as a classification problem. This is particularly worth noting as predicting the thread length answers the following research question: Is this thread likely to become viral in the future?

Thread Scoring Features	
OnThreadTopic[i]	$\frac{\lvert w_i \cap w_0 \rvert}{\lvert w_i \rvert}$, w_i = set of words in comment i
OverlapPrevious[i]	$\max_{0 \le j \le i-1} (\frac{\lvert w_i \cap w_j \rvert}{\lvert w_i \rvert})$, w_i = set of words in comment i
Timeliness[i]	$\frac{t_i - t_0}{(\sum_{n=1}^{N} t_n)/N}$, t_i = time of comment i, N = number of comments
Lengthiness[i]	$\frac{w_i}{(\sum_{n=1}^{N} w_n)/N}$, w_i = words in comment i, N = number of comments
Arrival Based Features	
Time[i]	$t_i - t_0$, t_i = time of comment i
ArrivalPattern[i]	[ID_0, ID_1,..., ID_i], ID_n = unique ID of commenter n
Sentiment Features	
Sentiment[i]	S_i, sentiment of comment i
AverageSentiment[i]	$(\sum_{n=1}^{i} S_n)/i$, S_n = sentiment of comment n
Topic Modeling Features	
TopicVector[i]	TV[i] = [s_0, s_1,..., s_k] = relevance score for each one of the k topics in comment i
TopicDrift[i]	TV[i] - TV[0], TV[i] = topic vector of comment t
Links	
PosterInteractions[i]	$\lvert T_i \cap T_0 \rvert$, T_i = threads that commenter i has participated in
MutualInteractions[i]	$\sum_{n=1}^{k} (\lvert T_i \cap T_n \rvert)$, k = users that have interacted with commenter i and original poster
Participants[i]	Number of unique commenters till comment i
TotalLinks[i]	$\sum_{n=1}^{k} (\lvert T_i \cap T_n \rvert)$, k = number of unique commenters till comment i
Text Features	
TextFeatureVector[i]	[tf_0, tf_1,...tf_n] = whether one of n text information gain features is present
Semantic Features	
SemanticFeatureVector[i]	[sf_0, sf_1,...sf_n] = whether one of n semantic information gain features is present
Miscellaneous Features	
NumWords[i]	w_i, number of words in comment i
Question[i]	Q_i, true if comment i has ?, false otherwise
Exclaim[i]	E_i, true if comment i has !, false otherwise

Table 1: Feature with description

The proposed framework extracts features from the initial comments in the thread, training binary classifiers on the generated feature vectors that in turn predict the eventual thread length into one of the

pre-specified ranges. Unlike previous approaches, our framework models the dynamically expanding nature of forum threads, by incorporating multiple binary classifiers each trained on a different number of initial comments. It is important to note that the choice for the number of comments used for training X and the length to predicted Y is not arbitrary it is derived from the distribution of thread lengths in the data. Specifically, using the eventual length as the variable, the evaluation threads are equally divided into 4 bins on percentile basis. Further, the values of bin ranges serve as the X and Y in the binary classifiers, thereby providing three binary classifiers for the 4 thread bins.
Table 1 shows the feature representations applied and their corresponding textual description for each. The feature representations proposed in our framework are detailed in the following subsections. Any feature with subscript 0 refers to a feature of the original post.

3.1 Sentiment Features

Subjective words that represent emotions and moods are important indicators of sentiment orientation (Sharif 2014). Such phrases are particularly common in threaded conversations, as they are used to convey opinions regarding various products, organizations and events. It is naturally conceivable that highly opinionated statements are more likely to grasp the attention of the online community that can in turn lead to more responses from fellow forum members. Therefore mining emotion related features adds a valuable indicator of eventual conversation length. Consider for instance the following post selected from the health forum "Drugs.com".
"I will be going through hell soon, until Robert and I decide that I've reached my lucky number of 26. Took my last dose a couple of hours ago. I would appreciate any and all support to get me through this waiting period. Robert, let's use this thread to do the induction. Thanks, buddy!! And thanks to all on this site. Any helpful hints anyone?"
It can be observed that the example post has a high degree of sentiment orientation and importantly is the original post of one of the longest conversation threads in our data. In the post the user reveals herself to be in a state of uncertainty and pleads for support from the forum community. Such posts are more likely to gain a large number of responses, and hence it is important to gather these underlying emotions and moods that form the basis of sentiment directionality. A similar study by (Qiu 2011) discovers that sentiment change patterns in health community users, and investigates the factors that affect the sentiment change. Most importantly, they concluded that forum participants change their sentiment in a positive direction through online conversations with community members. This supports the fact that sentiment is indeed a good indicator of the expected responses to a post. For labeling posts with their sentiment polarity we employ the SentiStrength tool for sentiment analysis.
Figure 2 draws a relationship between the average sentiment in the initial comments and the average thread lengths. Particularly the trend for binary classifier 3 illustrates that threads with overall negative and positive sentiment are likely to expand to a large number of comments. The average sentiment is calculated accordingly:

3.2 Topic Modeling Features

The content on forums is typically topic oriented. Users create threads, in order to seek opinions, share information and start new discussions. Discussions stem from a variety of topics including drugs, diseases, products and events. Identifying the topic content of a post helps extract:

1. a topic vector for every comment defining the relevance to each topic
2. the topic drift of a comment to the original post in the thread.

Specifically, topic drift is a very good indicator of estimating how focused a conversation is. A high topic drift is often suggestive of irrelevant content, or a possible diversion from the intended purpose of the conversation. On the contrary, it is also commonly observed that often within a forum discussion,

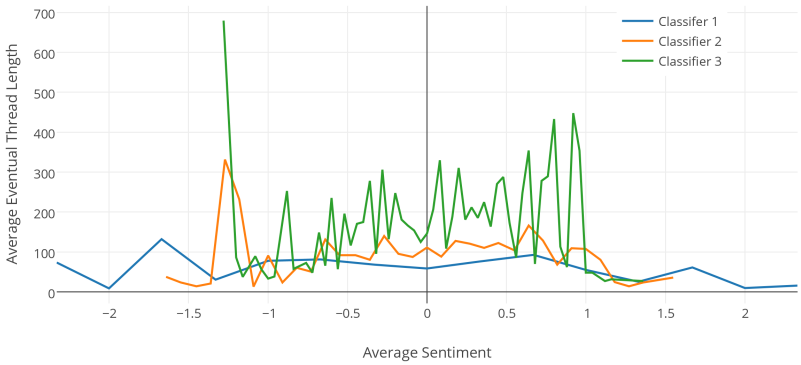

Figure 2: Average thread length per average thread sentiment (negative values relate to negative sentiment and correspondingly for positive values)

a sub-discussion spawns that is able to solicit engagement from the user community. The relationship between the eventual thread lengths and the average topic drift across threads is demonstrated in Figure 3. The trend observed in the figure is rather interesting; initially with the increase in average topic drift among the comments, the eventual thread length increases, however, if the topic drift is very high, the threads are expected to be very short. This supports the hypothesis that irrelevant comments with a high topic drift are likely to prevent the growth of the thread.

There is a visible increase in topic drift with the increasing number of responses. This validates the hypothesis that as more and more users get engaged in a thread topic, a topic drift is likely.

The topic scores are extracted using the topic modeling software developed by (Blei 2003). For a set of K topics, the tool returns a measure of relatedness to each topic. Thus a topic vector is a K-tuple of topic scores, and the topic drift is a vector difference of a comment's topic vector to the original post's topic vector.

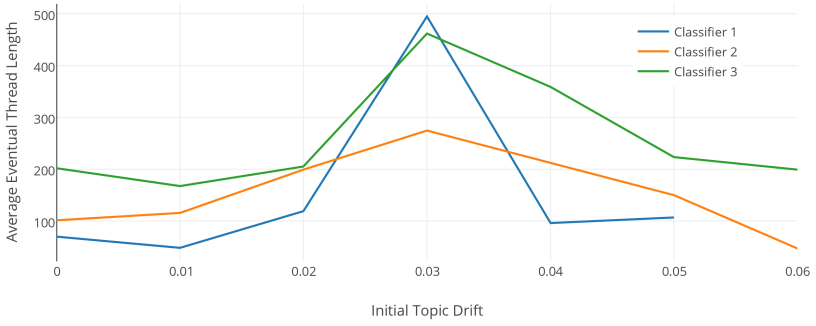

Figure 3: Average thread lengths per initial topic drift

3.3 Semantic Features

Previous work on blog content volume prediction and thread length prediction have employed the original post terms as features. Notably, (Backstrom 2013) also included the top 50 text regression model. Similarly, the proposed framework also adds the top 50 post terms with the highest information gain as a feature representation. In addition to textual features we also incorporate semantic features; whereas semantic features refer to logical abstract entities that group together a number of semantically similar words. In this regard, the WordNet (Miller, 1995) lexical database serves as a useful resource for mapping words to their respective semantic classes. The hypernym hierarchy in WordNet is established on the basis of a "type of" relationship. The framework uses the hypernym tree, for labeling semantically similar words with a common semantic class. The common semantic classes are able to group common entities that in turn serve as a valuable feature representation.

3.4 Thread Scoring Features

A bulk of content is posted in online discussion forums, much of which is irrelevant, uninteresting and unlikely to gain attention. Evaluation of online posts in terms of their value of contribution to the community can help users find knowledge within forum content. Notably, the work of (Wanas 2008) on automatic scoring of online discussions is relevant to the proposed problem of measuring user engagement and number of responses to forum threads. Moreover, experiments have demonstrated improved prediction performance with adding post scoring features as a feature representation. Hence, the proposed approach utilizes the thread scoring metrics as features for length prediction. The features are broadly classified as:

1. Relevance features that measure the relatedness of a thread post to the content of the leading post in the thread, and the forum in general
2. Originality features that quantify the overlap of terms of a post with preceding posts in the same thread
3. A set of forum specific features. These features include backward references, forward references, etc
4. Surface features that include the relative time elapsed between comments, and the lengthiness of the relative content length of a post
5. Posting component features that represent important syntactic and web elements, for example Weblinks

The description of each thread scoring features specific to each category is show in Table 1. The next subsections include descriptions for some of the important thread scoring features; ones that demonstrate a strong correlation with the eventual thread length.

3.4.1 OnThreadTopic

A discussion with high relevance to the original post signifies a topic where users were able to find the required information (Wanas 2008). The OnThreadTopic feature is used in order to gauge the relatedness to the discussion. The OnThreadTopic feature is computed by comparing a post's bag of words to the original post. It is worth noting that OnThreadTopic computes the relevance over the bag of words, it does not take into account the topical content, therefore, we need to incorporate the topic drift feature.
The graph with the averaged OnThreadTopic scores across the initial comments, and the corresponding thread lengths is shown in Figure 4. The figure presents a rather interesting trend, as the thread length reaches its peak value, when the OnThreadTopic is minimal. The trend is different to the one observed with Topic Modeling features, particularly as it employs a mere bag of words comparison.

3.4.2 Timeliness Features

The timeliness feature is a metric for quantifying the rate of replies. Specifically, it is the inter-arrival time between consecutive posts, normalized over the average inter-arrival time across posts in the same thread.

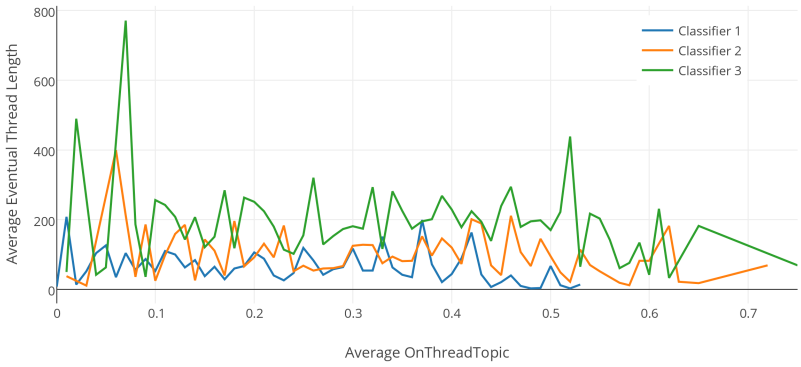

Figure 4: Average thread length per average OnThreadTopic

Figure 5 shows a plot of average timeliness scores for the initial comments in each classifier, with the corresponding eventual thread lengths. There is a visible trend that demonstrates that the timeliness of the comments in the thread is indirectly proportional to the thread lengths. Thereby, the timeliness feature serves as a good indicator of thread length.

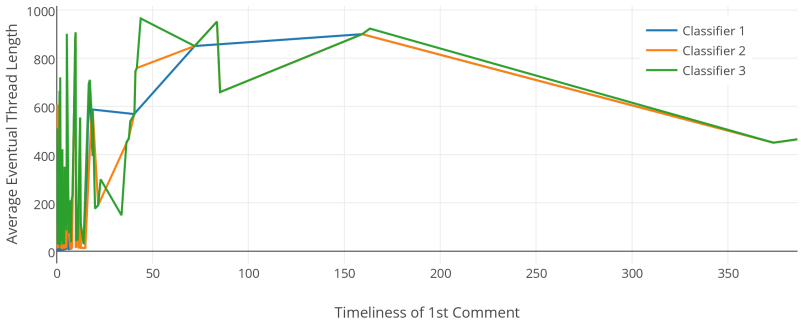

Figure 5: Average thread length per timeliness of 1st comment

4 Experiments

For our experimental evaluation we include a Telco dataset with 2487 derived from the telecommunications forum "Digitalhome" and a Health dataset with 3166 threads derived from the medical forum "Drugs.com". The feature processing phase, generates 7 parallel feature representations against each comment in each thread; the representations are used retaining the best performing model. Such a feature ensemble based approach is particularly useful as it assists the classifier in the extraction of improved discriminatory patterns. However, the plethora of features generated across all comments adversely affected classification performance. Hence, we applied feature selection using information gain to select only a subset of the highest performing features. Moreover, we employed a number of classifiers provided with the Weka toolchain including J48, RandomTree, REPTree, BayesNet, NaiveBayes, Logistic and RBFNetwork.

4.1 Results

The performance of our proposed method versus Backstrom's et al approach with Telco dataset is illustrated in Table 2 and Table 3, respectively. Results for Health dataset are illustrated in Table 4 and Table 5. The value X denotes the number of comments used for training and Y represents the length to be predicted. In addition to overall accuracy, class level precisions, F-measures, recalls and area under curves are also used as evaluation metrics. The experimental results demonstrate that our approach significantly outperforms the earlier approach with respect to accuracy, class level precisions and recalls. Importantly, the recall and precision statistics are balanced across the class. Overall the adopted approach demonstrates an accuracy improvement of 10% on the Telco dataset. While we can see an improvement of 9% on the Health dataset. The improved accuracies, precisions and recalls are particularly important as they represent better prediction of top trending results.

Range	Accuracy	F-Meas+	F-Meas-	Recall+	Recall-	Precision+	Precision-	ROC+	ROC-
3-10	71.16	70.83	71.47	70.05	72.27	71.64	70.70	71.16	71.16
11-24	70.8	70.68	70.92	70.4	71.2	70.97	70.63	70.84	70.84
25-97	78.03	78.90	77.09	82.13	73.93	75.91	80.54	77.86	77.86

Table 2: Experimental results from Backstrom et al. framework on Telco dataset

Range	Accuracy	F-Meas+	F-Meas-	Recall+	Recall-	Precision+	Precision-	ROC+	ROC-
3-10	87.80	88.26	87.29	91.76	83.84	85.02	91.05	93.91	93.91
11-24	83.44	83.51	83.37	83.84	83.04	83.17	83.71	83.11	83.11
25-97	86.72	86.63	86.81	86.07	87.38	87.21	86.25	88.63	87.97

Table 3: Experimental results from our framework on Telco dataset

Range	Accuracy	F-Meas+	F-Meas-	Recall+	Recall-	Precision+	Precision-	ROC+	ROC-
3-5	77.58	79.33	75.51	86.03	69.13	73.60	83.19	82.17	82.17
6-22	64.55	67.20	61.44	72.63	56.47	62.53	67.35	71.08	71.08
23-48	60.07	60.85	59.27	62.05	58.09	59.69	60.49	61.26	61.26

Table 4: Experimental results from Backstrom et al. framework on Drugs dataset

Range	Accuracy	F-Meas+	F-Meas-	Recall+	Recall-	Precision+	Precision-	ROC+	ROC-
3-5	86.60	86.93	86.25	89.16	84.05	84.82	88.57	92.18	92.18
6-22	74.91	77.25	72.03	85.20	64.61	70.65	81.37	81.91	81.91
23-48	67.45	67.15	67.74	66.55	68.35	67.77	67.14	72.98	72.98

Table 5: Experimental results from our framework on Drugs dataset

4.2 Case Study

The results presented in the previous section have demonstrated improved thread length prediction of our approach in comparison to previous methods. As previously mentioned, the motivation of thread length prediction on social media content, is the identification of trending topics. In the context of medical social media this is particularly useful as it allows for gaining insights into the popular drugs, and the user experiences relating to those drugs (Sharif et al, 2014). The user experiences can in turn serve as important indicators for future adverse drug reactions.

In order to identify adverse drug reactions using thread length predictions and correlate them with actual events we conducted a set of experiments on forum datasets from "Drugs.com". Further, we use a custom

list of 4000 drugs, and 5800 reactions. For our analysis, we only include threads with lengths predicted to exceed a pre-specified length. The pre-specified length is kept to a value that is 75 percentile with respect to the distribution of threads. The total number of threads employed in our analysis are 1558. Specifically, we only include the longer threads, as they are representative of the prevailing public opinion about certain discussion topics. For labeling a thread as discussing a certain drug, we place a minimum drug count threshold of 3. Furthermore, we require that the drug co-occurs with at least 3 reactions/diseases in each thread. This condition has been imposed so that we only include threads, where the users express the reactions/diseases they experience on usage of the drug. Moreover, we employ Point-Wise Mutual Information (PMI) to gauge the association of each drug with each disease/reaction, thereby ranking their correlation. This is particularly useful as it not only allows us to identify possible adverse drugs, but also their reactions.

For the purpose of this study, we use the adverse drug reactions list published by the FDA as the gold standard. Among the list of 119 drugs (including impacted drugs), the health dataset contained a total of 38 drugs. From this list of 38 drugs, our framework was successful in identifying 22 drugs as having possible adverse reactions. Furthermore, as many as 21 drugs were predicted before the FDA prediction. Table 6 shows the detected drugs, along with detection dates and associated reactions. It can be observed that in many cases the detected reaction list for the drugs, consists of closely related reactions for instance fear, depression, anxiety, irritability and headaches. Most importantly, in some cases we have also been able to correctly predict the reaction caused by the drug; for instance the drug Zocor we are able to identify the reactions: pain, rhabdomyolysis and discomfort.

Drug Name	Threads	Detection Date	FDA Detection Date	Diseases/Reactions Discussed
Escitalpram	6	17/11/2006	14/12/2011	Fatty liver, hypertrophy, diabetes
Sertraline	10	14/09/2004	14/12/2011	Sexual dysfunction, suicidal thoughs
Celexa	71	17/11/2006	14/12/2011	Fear, emotional changes, hostility
Paxil	83	21/12/2005	14/12/2011	Suicide, depression, anxiety
Percodan	15	19/10/2005	09/01/2012	Weight, gain, allergy
Tylenol	483	15/09/2004	22/12/2011	Pain, cramps, migraine, headaches, stress
Zocor	4	13/08/2004	08/06/2011	Pain, rhabdomyloysis, discomfort
Lexapro	116	17/11/2006	14/12/2011	Anxiety, depression, death, suicidal ideation
Opana	65	23/11/2010	09/01/2012	Death, abdominal pain, pain, constipation
Atorvastatin	2	08/07/2004	28/02/2012	Abdominal pain, allergic reaction
Prozac	107	10/05/2005	14/12/2011	Depression, death, agitation
Fluvoxamine	7	14/09/2004	14/12/2011	Lung cancer, bipolar disorder, depression
Acetaminophen	102	03/02/2005	22/12/2011	Liver damage, depression, fever
Luvox	18	14/09/2004	14/12/2011	Constipation, pcp, diarrhea
Citalopram	14	14/09/2004	24/08/2011	Adenoidectomy, infection
Zofran	13	08/08/2012	15/09/2011	Withdrawal symptoms, vomiting, euphoria
Excedrin	118	26/12/2007	09/01/2012	Migraine, headaches, chills
Codeine	143	15/09/2004	15/08/2012	Pain, nightmare, discomfort, anxiety
Percocet	443	15/09/2004	09/01/2012	Pain, depression, injury
Lipitor	7	13/08/2004	28/02/2012	Back pain, chest pain, muscle pain
Zoloft	103	05/11/2004	14/12/2011	Anxiety, irritability, withdrawal symptoms
Adderall	79	05/03/2005	30/05/2012	Cancer, pain, headache

Table 6: Prediction of Drug side-effects

5 Conclusion

In this study, we propose a classification framework for the prediction of thread lengths in online forums. Our framework leverages from feature representations, extracting underlying attributes associated with that thread. Longer thread lengths are usually associated with viral topics on the Internet. We run extensive experiments on Telco and Health forum data-sets. We also present a case study where we perform an

analysis on the adverse effects of different drugs. Our analysis focused on predicting if we could find out if and when adverse effects were arising from the consumption of drugs. Our predictions corroborate with the data provided by FDA, and in fact, our framework found these anomalies earlier than FDA had documented them.

6 Acknowledgements

This work was funded in part by the following grant from the U.S. National Science Foundation: IIS-1236970.

References

Backstrom, L., Kleinberg, J., and Lee, L., and C. Danescu-Niculescu-Mizil (2013). "Characterizing and Curating Conversation Threads: Expansion, Focus, Volume, Re-entry." In: *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM '13)*, pp. 13-22

Blei, D.M., Andrew, Y.N., Michael, I.J., and J. Lafferty (2003). "Latent Dirichlet Allocation." *The Journal of Machine Learning Research (2003)*, vol. 3, pp. 993-1022

Hassan, A., Abbasi, A., and D. Zeng (2013). "Twitter Sentiment Analysis: A Bootstrap Ensemble Framework." *2013 International Conference on Social Computing (SocialCom '13)*, pp. 357-364

Honey, C., and S.C. Herring (2009). "Beyond Microblogging: Conversation and Collaboration via Twitter." *42nd Hawaii International Conference on System Sciences (HICSS '09)*, pp. 1-10

Miller, G.A (1995). "WordNet: A Lexical Database for English." *Communications of the ACM*, vol. 38, no. 11, pp. 39-41

Nicholas Diakopoulos and Mor Naaman. (2011). "Towards quality discourse in online news comments." *In Proceedings of the ACM 2011 conference on Computer supported cooperative work (CSCW '11)*. ACM, New York, NY, USA, 133-142.

Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. (2008). "Knowledge sharing and yahoo answers: everyone knows something."" *In Proceedings of the 17th international conference on World Wide Web (WWW '08)*. ACM, New York, NY, USA, 665-674.

Qiu, B., Zhao, K., Mitra, P., Wu, D., Caragea, C., Yen, J., Greer, G.E., and K. Portier (2011). "Get Online Support, Feel Better–Sentiment Analysis and Dynamics in an Online Cancer Survivor Community." *IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom '11)*, pp. 274-281

Chirag Shah and Jefferey Pomerantz. 2010. "Evaluating and predicting answer quality in community QA."" *In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '10)*. ACM, New York, NY, USA, 411-418.

Sharif, H., Abbasi, A., Zaffar, F., and D. Zimbra (2014). "Detecting Adverse Drug Reactions using a Sentiment Classification Framework." *In the 6th ASE International Conference on Social Computing, Stanford*, pp. 27-31

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and A. Kappas (2010). "Sentiment Strength Detection in Short Informal Text." *Journal of the American Society for Information Science and Technology*, 61: 2544-2558

Tsagkias, M., Weerkamp, W., and M.D. Rijke (2009). "Predicting the Volume of Comments on Online News Stories." In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*, pp. 1765-1768

Wanas, N., El-Saban, M., Ashour, H., and W. Ammar (2008). "Automatic Scoring of Online Discussion Posts." In: *Proceedings of the 2nd ACM Workshop on Information Credibility on the Web (WICOW '08)*, pp. 19-26

Yano, T., and N.A. Smith (2010). "What's Worthy of Comment? Content and Comment Volume in Political Blogs." In: *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM '10)*

Abbasi, A. and Adjeroh, D. "Social Media Analytics for Smart Health," *Proceedings of the IEEE Intelligent Systems, 29(2), 2014*, pp. 60-64.

Adjeroh, D., Beal, R., Abbasi, A., Zheng, W., Abate, M., and Ross, A. "Signal Fusion for Social Media Analysis of Adverse Drug Events," *Proceedings of the IEEE Intelligent Systems, 29(2), 2014*, pp. 74-80.

Fu, T., Abbasi, A., Zeng, D., and Chen, H. "Sentimental Spidering: Leveraging Opinion Information in Focused Crawlers," *ACM Transactions on Information Systems, 30(4), 2012*, no. 24.