Searching for Quality Microblog Posts: Filtering and Ranking based on Content Analysis and Implicit Links

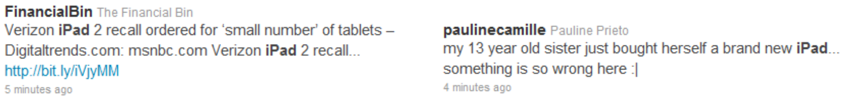
Jan Vosecky, Kenneth Wai-Ting Leung, and Wilfred Ng

Department of Computer Science and Engineering
Hong Kong University of Science and Technology, Hong Kong, China

Abstract. Today, social networking has become a popular web activity, with a large amount of information created by millions of people every day. However, the study on effective searching of such social information is still in its infancy. In this paper, we focus on Twitter, a rapidly growing microblogging platform, which provides a large amount, diversity and varying quality of content. In order to provide higher quality content (e.g. posts mentioning news, events, useful facts or well-formed opinions) when a user searches for tweets on Twitter, we propose a new method to filter and rank tweets according to their quality. In order to model the quality of tweets, we devise a new set of link-based features, in addition to content-based features. We examine the implicit links between tweets, URLs, hashtags and users, and then propose novel metrics to reflect the popularity as well as quality-based reputation of websites, hashtags and users. We then evaluate both the content-based and link-based features in terms of classification effectiveness and identify an optimal feature subset that achieves the best classification accuracy. A detailed evaluation of our filtering and ranking models shows that the optimal feature subset outperforms traditional bag-of-words representation, while requiring significantly less computational time and storage. Moreover, we demonstrate that the proposed metrics based on implicit links are effective for determining tweets' quality.

1 Introduction

In recent years, social networking and microblogging services have seen a steep rise in popularity, with users from a wide range of backgrounds contributing content in the form of short text-based messages. Microblogging services, in particular Twitter, are at the epicentre of the social media explosion, with millions of users being able to create and publish short messages, referred to as *tweets*, in real time. It is estimated that nearly 200 million tweets are generated and over 1.6 billion search queries are issued each day [1] and these figures are likely to keep rising in future. However, the work on searching tweets or similar social information is still in its infancy. Unlike traditional web search, the search results from social networking services may be mostly relevant, however may include a large proportion of low-quality and noisy messages.



FinancialBin The Financial Bin
Verizon **iPad** 2 recall ordered for 'small number' of tablets – Digitaltrends.com: msnbc.com Verizon **iPad** 2 recall...
http://bit.ly/iVjyMM
5 minutes ago

paulinecamille Pauline Prieto
my 13 year old sister just bought herself a brand new **iPad**... something is so wrong here :|
4 minutes ago

Fig. 1. Two relevant tweets returned as a result to a search for 'iPad'. The first tweet shares factual news about the product, while the second tweet only mentions about the author's family and includes an unclear subjective judgement

In this paper, we focus on Twitter, a popular social networking and microblogging platform. The social networking features include subscribing to tweets by other users, forwarding tweets from other users and explicitly addressing other users in their tweets. During recent events, such as natural disasters or political turbulences, the influence of Twitter has become even more evident.

While there has been plenty of study on the dynamics of information spreading, influence and authority in Twitter, little attention was paid to how to find good quality content in Twitter. Twitter is clearly a rich source of data, however, there are also several new challenges compared with traditional web searching.

- Very brief content: In Twitter, only 140 characters are available to convey the author's message. This poses new challenges to establish an effective set of features for filtering and ranking search results.
- Highly dynamic in scale: There is a large quantity of such postings, with nearly 200 million new tweets published each day [1]. This demands more efficient techniques for identifying high quality tweets.
- Informal language: Being user-generated content, postings often contain misspellings, abbreviations, slang expressions and the like. This makes the analysis of tweets more difficult.
- Varying quality: The level to which a tweet contains high quality information varies dramatically. [2] found that 57% of tweets are not of general interest, except to the author or the author's close friends. [3] claims that users post several types of messages, only some of which are intended to be of interest to a wider audience.

The above challenges become more obvious when searching for content in Twitter, which presents results ordered by recency of posting. The user then needs to manually pick out high quality content among potentially thousands of results. Consider a search for the term "iPad" and two different search results shown in Figure 1. The example illustrates that there may be a large difference in the level of quality of tweets returned as results to a search query.

In order to achieve more effective tweet search, we tackle the following two problems: filtering[1] and ranking tweets according to their quality. We may apply these two methods on a recent set of tweets. Basically, our approach involves the following steps, as illustrated in Figure 2. First, to capture the quality of

[1] The terms "filtering" and "classification" are used interchangeably throughout the paper.

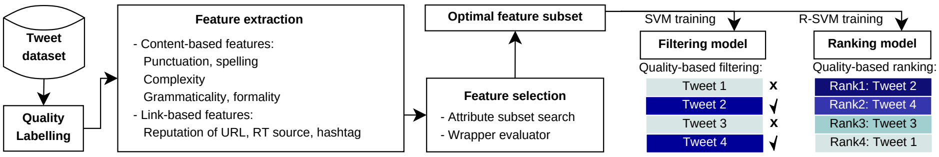


Fig. 2. Overall process of filtering and ranking tweet based on quality analysis

individual tweets, we examine each tweet with a set of content-based and link-based features. Our newly proposed link-based features leverage the implicit relationships between tweets, hashtags[2] and hyperlinks. Second, we evaluate the effectiveness of individual features and perform attribute subset selection to obtain an optimal subset of features for the classification task. Third, the optimal feature subset can be used for constructing (i) a filtering model, or (ii) a ranking model. The filtering model aims to identify high quality tweets from a given set of tweets, or to filter out low quality, noisy tweets. On the other hand, the ranking model aims to rank a set of tweets based on their quality, with high quality tweets ranked at top positions. Finally, we show that our optimal feature subset outperforms a baseline method based on TFIDF representation for the filtering task, while requiring less computational time and storage in our empirical evaluation. In addition, for the ranking task, we significantly outperform a state-of-the-art method based on hyperlink presence.

Our contributions in this paper are then as follows:

- We propose a new strategy for filtering and ranking of tweets, focusing on the quality of a tweet, in order to improve on the basic search functionality in Twitter. Our detailed evaluation shows that our strategy helps to improve the naïve recency-as-relevance approach currently used.
- We propose a novel set of link-based features in order to model the quality of a tweet, utilizing the implicit relationships between tweets, hyperlinks and users. We introduce three metrics to reflect the quality of tweets which relate to a specific URL, hashtags or a user. These features provide useful evidence to our models and boost the filtering and ranking performance.
- We examine both link-based and conventional content-based features, and evaluate their effectiveness in the modelling task. We then identify an optimal (best-performing) feature subset. To our knowledge, this is the first study with detailed analysis of features for the task of filtering and ranking of general tweets according to their quality.

The remainder of the paper is organized as follows. Section 2 discusses related work in the area of Twitter analysis. In Section 3, we present background on Twitter and define the notion of tweet quality. Section 4 discusses our filtering approach and the features of tweets. Our ranking approach is then presented in Section 5. Section 6 provides an evaluation of the filtering and ranking models, as well as of our proposed features. Finally, Section 7 concludes the paper.

[2] Hashtags are tags prefixed with a '#' symbol to indicate the topics of the tweet and enable posts related to the same topics to be quickly searched.

2 Related Work

Twitter has been an active area of research in recent years. [3, 4] provide initial insights into the usage patterns in Twitter and how communities are formed. From the content point of view, previous work largely aimed to classify tweets into a predefined set of categories, based on their purpose (such as 'news', 'events', 'opinions', etc.). [5] proposes a feature-based method for classifying tweets into 6 categories. [6] proposes the use of topic models and supervised learning to assign 4 broad topics to tweets. Our work is complementary to these works as we focus on the quality of tweets, which is a new dimension orthogonal to such predefined categories.

From the perspective of analyzing the quality of tweets, [2] is closer to our work. It provides initial insights in the classification of Tweets based on their interestingness to the reader and presents a set of potential features. However, only the presence of a hyperlink is used for classification in [2]. Our work focuses on a more generalized quality-based classification and ranking, examine a larger set of features and proposes new features which outperform the link-only approach. We also provide a detailed feature evaluation. In [7], several features are proposed to find interesting clusters of tweets for specific events. However, no experimental evaluation of the results is provided.

From the perspective of effective tweet ranking, early attempts at new algorithms to rank tweets were proposed. [8] proposes several simple methods, e.g. based on the number of followers of a user or the length of a tweet. [9] ranks tweets using non-negative matrix factorization, based on the bag-of-words representation. However, these works did not provide comprehensive evaluation or convincing empirical results. [10] ranks tweets in Twitter-like forums based on star-ratings or thumb-ratings, not taking content into account. [11] employs a learning-to-rank approach using a hybrid set of features (query-content relevance, content-based features, author features). In our work, we focus on the specific problem of analyzing the *quality* of an individual tweet. We formulate criteria of tweet quality and examine a comprehensive set of content-based and novel link-based features. Our work also provides a detailed feature evaluation for tweet filtering and ranking based on their quality.

3 Tweet Quality Analysis

3.1 Preliminaries

Twitter is a social networking and microblogging platform, in which registered users may post short messages (*tweets*) of up to 140 characters in length. These messages are published and available for search in near-real time and can be posted either using a web interface, via SMS messages or through a wide range of third-party applications. Currently, Twitter has over 300 million registered users who post over 200 million tweets and submit around 1.6 billions search queries per day [12].

The social networking features include: (1) subscribing to posts by another user (*follow*), (2) forwarding posts from other users (*re-tweet*, indicated by a "RT" prefix) and (3) explicitly addressing users in their posts (*mentions*, indicated by a "@" symbol followed by a username), thus enabling conversations and

Table 1. Examples of tweets judged by different quality criteria

Criterion	*Positive example*	*Negative example*
Well-formedness	"Lady Gaga is on the 4th place among solo artists with the most top tens in a row, only behind Janet Jackson, Madonna and Whitney Houston."	"Serena got a $2000 fine for the outburst....hahhahahaahahhaha but she told her her news yong! LOL"
Factuality	"Apple to release iOS 5 GM to assemblers during week of Sept. 23 (@thisisneil / AppleInsider)"	"so now that i have my iphone is jailbroken, what should i download on it ? i really dont know"
Navigational quality	"#Japan's prime minister promises help to city decimated by tsunami and earthquake http://dlvr.it/N2v7q" *[links to a news article]*	"This is what I call a perfect Sunday afternoon! http://bit.ly/endbUc" *[links to a family photograph]*

replies to be carried out. Within tweets, users may also include *hashtags* (tags prefixed with a '#' symbol) to indicate the topics discussed and enable posts related to the same topics to be grouped together and searched more directly.

By default, the user's profile and tweets are publicly accessible, unless restricted to the user's *followers*. Data available on Twitter is also accessible via Twitter's REST API.

3.2 Goal Definition: Defining Tweet Quality

In this section, we focus on the notion of *tweet quality* more closely and set out our goals for modelling and assessing the quality of tweets. Based on their purpose, messages on Twitter have been found to fall into several categories, such as conversational, information sharing, news reporting, etc. [3]. Instead of focusing on a specific type or category of tweets, we aim to establish criteria for judging the quality of tweets in general. Therefore, we define our notion of an 'interesting' tweet along the following 3 criteria:

- *Well-formedness.* Well-written, grammatically correct and understandable tweets are preferred over tweets containing heavy slang, uncomprehensible language or excessive punctuation.
- *Factuality.* News, events, announcements and other facts of general interest are preferred over tweets with an unclear message, private conversations and generic personal feelings, which typically do not convey useful information.
- *Navigational quality.* A tweet that links to reputable external resources (e.g. news articles, reports, or other online materials) may provide further information to the reader. However, not all links may be of general interest (e.g. links to photo sharing websites, used for sharing personal photos). Therefore, it is important to distinguish what type of website a tweet refers to.

Examples of tweets judged along these criteria are shown in Table 1. In real scenarios, tweets may exhibit more than one of the criteria (e.g. news-oriented tweets are typically factual and provide a link to the full news article). In fact, these 3 criteria allow for flexibility when judging different types of tweets[3].

[3] For example, when searching for tweets reviewing a movie, 'well-formed' tweets would be preferred over those containing excessive slang or strong language. Or, when

In order to assess tweets according to the quality criteria, we follow a process described in subsequent sections. In particular, we extract features from tweets (Section 4.2) to capture various characteristics, as inspired by the 3 criteria described in this section. These features then form a basis of our filtering and ranking models.

4 Quality-Based Tweet Filtering

4.1 Classification Method

Since our work focuses on the tweet-specific feature extraction and evaluation, rather than on the classification algorithm itself, we utilize standard classification tools. Due to its wide-spread adoption and proven effectiveness in text mining tasks, we use *Support Vector Machines* (SVM) for the classification task.

4.2 Characterizing Tweets with Features

To gain deeper insight into which factors most influence the quality of a tweet, we extract a number of features from every tweet. The features can be broadly divided into *content-based* and *link-based* features. *Content-based* features may be used to identify low-quality tweets which contain many spelling mistakes and use punctuation excessively. These features correspond to the *well-formedness* criteria in Section 3.2. Next, features based on the complexity and formality of the language correspond to the *factuality* criteria. *Link-based* features include the presence of hyperlinks, hashtags or mentions of other users. We also propose a set of novel metrics to obtain reputation scores for URL domains, users and hashtags. These features, in particular the URL domain reputation, addresses the *navigational quality* criteria.

Punctuation and Spelling Features:

Excessive punctuation. We measure any abnormalities in punctuation with features, such as the number of exclamation marks, number of question marks and the maximum number of repeated characters.

Capitalization. Another kind of abnormality is content written in all-capitalized letters. We capture the presence of all-capitalized words and the largest number of consecutive words in capital letters.

Spelling. We extract the number of correctly spelled words and the percentage of words found in a dictionary. The dictionary used in this task is provided by the Stanford Natural Language Processing lab.

Syntactic and semantic complexity:

Syntactic complexity. We measure the absolute length of the tweet, average word length, maximum word length and the percentage of stopwords. We also determine whether specific symbols, such as emoticons are present. The presence of numbers and measure symbols ($, %) is also extracted, and would apply to tweets that mention specific monetary or statistical data.

Tweet uniqueness. On a higher level, we measure the uniqueness of a tweet relative to other tweets by the same author. This feature is based on the traditional TFIDF approach in information retrieval. We may view a tweet t_j as

searching for tweets about ‘iPhone’, tweets linking to news articles about ‘iPhone’ would be preferred over tweets linking to private photos taken with an ‘iPhone’.

a set of terms $t_j = \{w_1, \ldots, w_n\}$. The uniqueness of a tweet t_j is then defined as $uniq(t_j) = \sum_{w_i \in t_j} tf_{i,t_j} \times idf_i$, where tf_{i,t_j} is the frequency of term i in tweet j and idf_i is the inverse document frequency of term i. More specifically, $tf_{i,t_j} = \frac{n_{i,t_j}}{\sum_k n_{k,t_j}}$ where n_{i,t_j} is the number of occurrences of term i in tweet j. The inverse document frequency of term i is defined as $idf_i = \log \frac{|T_u|}{|\{t_k : w_i \in t_k\}|+1}$ where $|T_u|$ is the total number of tweets from user u and the denominator indicates the number of tweets containing term i.

Grammaticality:

Parts-of-speech. We analyze parts-of-speech (PoS) within the tweet (such as nouns, verbs, adjectives, etc.). We use a PoS tagger to tag each word within the tweet with its corresponding PoS. We also check whether first-person parts-of-speech are present.

From existing metrics to measure the complexity and formality of written text, we chose the “formality score” from [13], which is based on the amount of different PoS that occurs in a text. The score is typically used to estimate the difficulty of understanding longer pieces of text, such as articles or books. The formality score[4] is defined as:

$$F = ((nounfrequency + adjectivefreq. + prepositionfreq. + articlefreq. - \\ pronounfreq. - verbfreq. - adverbfreq. - interjectionfreq. + \lambda)/2) \tag{1}$$

Presence of names. We identify proper names within the tweet as words with a single initial capital letter. We also determine the maximum number of consecutive proper names in the tweet.

Next, we identify named entities in the tweet using a Named Entity Recognition (NER) tagger. The tagger labels words or word groups which are likely to refer to names of places, persons or organizations.

Link-based features:

Out-link features. We extract whether the tweet contains a hyperlink, if it is a re-tweet (indicated by the “RT” prefix), the number of ‘@username’ mentions within tweet and the number of ‘#hashtags’.

Reputation features. One observation made about tweets that contain links is that tweets which link to specific web sites, such as news portals, generally contain higher quality information than tweets which link to domains such as social networking or picture sharing web sites. We generalize this problem to any URL domain and propose a feature that captures the reputation of the domain, based on the quality of tweets that point to that domain. A URL domain should have a high reputation score if (1) many tweets link to the domain, and (2) the tweets are of good quality. Conversely, if many low-quality tweets link to a domain, its reputation score should be low.

We then extend this concept to hashtags and re-tweeted users. The re-tweet based reputation of a user captures the quality of re-tweets originally posted by that user. The intuition is that a user should have a high reputation score if his

[4] The formality score was originally designed for longer pieces of text, with $\lambda = 100$. We adapt the value of λ in order to match the restricted length of tweet messages, $\lambda = 10$.

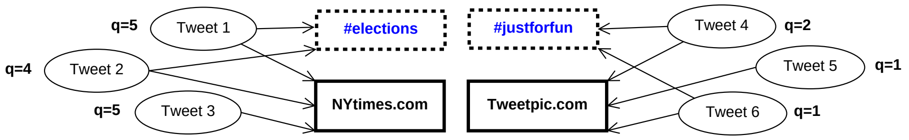


Fig. 3. Illustration of tweets containing links to two URL domains and two hashtags, including the tweets' quality scores (q)

or her tweets have been (1) re-tweeted often and also (2) are of good quality. For hashtags, the reputation score is based on the quality of tweets containing a specific hashtag. Figure 3 shows an example of two groups of tweets that link to two websites, with some tweets also containing hashtags. Our focus is on what can be generalized about the websites and hashtags from the fact that Tweets 1-3 have a higher quality than Tweets 4-6.

URL domain reputation: As mentioned earlier, the purpose of the URL domain reputation feature is to capture the quality of tweets that link to a specific URL domain[5].

To calculate the URL reputation, we firstly define an average domain quality measure. For a set $T_d = \{t \in T : t \mapsto d\}$ consisting of tweets that link to domain d, the Average Domain Quality ($AvgDQ$) is given by:

$$AvgDQ(d) = \frac{1}{|T_d|} \sum_{t \in T_d} q_t, \tag{2}$$

where q_t denotes the quality score of tweet t, $q_t \in [-1, +1]$.

We then use this measure to define the Domain Reputation Score (DRS) as:

$$DRS(d) = AvgDQ(d) \times \log(|T_d|), \tag{3}$$

where $AvgDQ(d) \in [-1, +1]$.

Intuitively, DRS formalizes the idea that the reputation score is increasing with more high-quality tweets linking to d and vice versa. To illustrate the reputation scores obtained using this approach, we calculate DRS for all URL domains in our dataset. Table 2 lists domains with the highest and the lowest DRS, the $AvgDQ$, and the number of tweets linking to the domain (Inlinks).

RT source reputation: Similarly to URL domain reputation, we leverage the quality of re-tweets that originate from a specific user in order to obtain the source user's reputation.

The RT Source Reputation Score (RRS) is given by:
$RRS(u) = \left[\frac{1}{|RT_u|} \sum_{t \in RT_u} q_t\right] \times \log(|RT_u|)$, where RT_u is the set of re-tweets originally posted by user u and $q_t \in [-1, +1]$ is the quality score of tweet t.

Hashtag reputation: Hashtag reputation leverages the quality of tweets related to a particular hashtag. The Hashtag Reputation Score (HRS) is calcu-

[5] The process of extracting the feature requires two pre-processing steps. First, we need to translate shortened URL links to their original destinations. Links posted in microblogs are commonly shortened to save space in the post, resulting in the need to retrieve the real destination of each link. Second, we group each tweet containing a link to the respective first-order domain of the URL link.

Table 2. URL domains with the highest and lowest Domain Reputation Score (*DRS*)

10 Domains with Highest DRS				**10 Domains with Lowest DRS**			
Domain	*Inlinks*	*AvgDQ*	*DRS*	*Domain*	*Inlinks*	*AvgDQ*	*DRS*
gallup.com	99	0.96	1.92	tweetphoto.com	126	-0.86	-1.80
mashable.com	101	0.76	1.53	twitpic.com	140	-0.80	-1.72
hrw.org	58	0.86	1.52	twitlonger.com	58	-0.93	-1.64
shoppingblog.com	47	0.87	1.46	lockerz.com	54	-0.81	-1.41
redcross.org	30	0.80	1.18	yfrog.com	93	-0.70	-1.38
intuit.com	61	0.57	1.02	laurenconrad.com	33	-0.88	-1.33
good.is	31	0.68	1.01	celebuzz.com	19	-1.00	-1.28
usa.gov	30	0.67	0.98	myloc.me	24	-0.83	-1.15
thegatesnotes.com	24	0.67	0.92	instagr.am	54	-0.63	-1.09
reuters.com	8	1.00	0.90	formspring.me	20	-0.80	-1.04

lated as: $HRS(h) = \left[\frac{1}{|T_h|}\sum_{t\in T_h} q_t\right] \times \log(|T_h|)$, where T_h is the set of tweets including hashtag h and $q_t \in [-1, +1]$ is the quality score of tweet t.

Timestamp: We use two features based on the timestamp of the tweet. The timestamp is discretized by hour of the day, as well as day of the week.

5 Ranking Tweets by quality

One of the drawbacks of the filtering method proposed in Section 4 is that it may not always be possible to clearly determine which class a particular tweet belongs to. This is true especially for tweets close to the classification boundary. Intuitively, such tweets could be labelled as being "average quality" or "neutral". While multiple classes of quality could be introduced, their exact meaning would be hard to define or interpret. A more intuitive solution might be to assign a continuous-valued score to a tweet (given by a regression model), or to produce a ranking for a set of tweets.

The goal of our ranking approach is to order a set of tweets based on their relative quality. More specifically, the ranking is based on the quality when considering each pair of tweets in the dataset. Our aim is to find a function $\mathcal{F}$ which, given two tweets t_1 and t_2, would output an ordered pair $\mathcal{F}(t_1, t_2) = (t_1 \succ t_2)$ iff $q_{t_1} > q_{t_2}$. In this way, given a set of tweets, we can produce an ordered sequence based on their quality.

5.1 Ranking Method

Our general approach proceeds in three phases: (1) tweets matching a query (based on string matching) are retrieved, (2) features of the tweets are extracted (as presented in Section 4.2) and (3) the query-tweet pairs, together with the quality scores of the tweets, are passed as input to a Learning-to-rank algorithm.

We adopt Rank SVM [14] to construct our ranking model, which is a simple and widely used Learning-to-rank technique. It takes pair-wise relationships between queries and tweets with their corresponding quality labels to learn a ranking model. Given an input set of unordered instances, the model will then output a sequence of instances ordered by their relative quality.

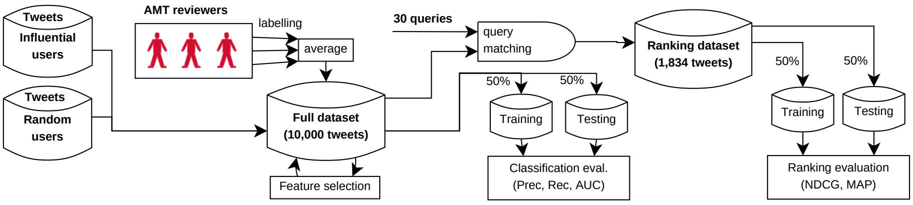


Fig. 4. Evaluation flow diagram

5.2 Features for Ranking

Similarly to our filtering approach, the criteria for ranking are based on the quality of a tweet. For this reason, we adopt the same set of features as presented in Section 4.2 to describe the content-based and implicit link-based characteristics of tweets for our ranking model.

6 Experimental Evaluation

In this Section, we describe our evaluation dataset and present the results of our filtering and ranking methods, with a particular focus on feature importance. We illustrate the overall evaluation flow in Figure 4.

6.1 Dataset

The dataset used in our experiments consists of 10,000 tweets from 100 Twitter users, with 100 recent tweets from each user. The dataset is collected from two different user classes, namely *general users* and *influential users*. The first user class contains 50 randomly selected users from a Twitter dataset provided by [15]. These users represent members of the general public. The second user class contains 50 influential users, selected from a popular website[6] that lists influential Twitter users in various categories. The 50 users are randomly selected from 5 different categories (technology, business, politics, celebrities and activism) to avoid any topical bias and their tweets were crawled using Twitter's REST API.

Training data labelling Due to a lack of a publicly available Twitter dataset with quality judgments, we manually build an evaluation dataset. To obtain the quality labels for our Twitter dataset, we utilize the Amazon Mechanical Turk[7] crowdsourcing service. The collected tweets are presented in a random order to reviewers, who are asked to assign a 1-5 rating to each tweet. Rating "1" represents a low-quality tweet, while rating "5" represents a high-quality tweet. To increase the objectivity of labelling and avoid bias from any individual reviewer, the ratings are collected and averaged from three different reviewers.

After labelling the Twitter dataset, we analyse the distribution of tweet quality in the dataset (Figure 5). Apart from the overall distribution, we also extract quality distributions for the two different user classes, general users (5,000 tweets) and influential users (5,000 tweets). Furthermore, we also analyze the distributions for re-tweets (1,941 tweets) and reply-tweets (2,676 tweets). Based

[6] http://www.listorious.com

[7] https://www.mturk.com

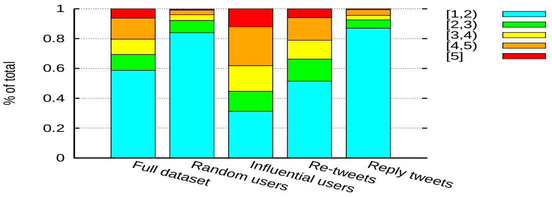


Fig. 5. Distribution of tweet quality as average quality ratings assigned by reviewers (higher rating implies higher quality)

on the results in Figure 5, we observe that the proportion of high-quality tweets from influential users is considerably larger than those from random users. We also observe that influential users may sometimes post low-quality tweets, thus the proposed filtering and ranking methods are also useful for tweets written by influential authors. Interestingly, we find that re-tweets are only slightly better than general tweets in terms of quality, indicating that even re-tweets may include low-quality or noisy messages. Finally, we observe that reply tweets have mostly low quality, most likely due to their conversational nature.

6.2 Filtering Evaluation

Evaluation Methodology To evaluate our filtering method, we use 50% of randomly selected tweets from our labeled dataset as the training set and the remaining 50% as the test set. Since the labeled ratings are in the range of $[1, 5]$, they are converted to binary labels based on the mean value 3 ($label \leq 3$ meaning 'low-quality', $label > 3$ meaning 'high-quality'). The SVM classifiers are trained using the features as discussed in Section 6.2. In additional to the proposed features, we also extract n-grams up to length 5 ($1 \leq n \leq 5$) from the tweets and use a TFIDF representation as an opponent in the comparison. In the evaluation, we use standard precision and recall with respect to each of the binary labels. We also present the Area under the ROC Curve (AUC) as an overall performance metric in the comparsion.

Feature Selection. To study the importance of each feature for the classification task, we first calculate Information Gain (IG) of each feature with respect to the class label. Table 4 lists all features sorted by their IG values. We observe that the top two link-based features ($IG = 0.374$ and 0.287) significantly outperform other features ($IG \leq 0.130$) in terms of IG, showing that they are the two most important features in the classification. Moreover, language complexity and named entities are also very useful in the classification with a high IG.

The next goal is to identify the optimal subset of features for the SVM classification model. For that purpose, we employ Greedy Forward attribute subset search and a Wrapper evaluator [16] as the feature selection algorithm. Greedy Forward attribute search starts with an empty set of features and greedily adds new features which contribute most to the classification. The search finishes once the newly added feature would no longer affect the performance of the classification. To pick the optimal subset of features, a Wrapper evaluator is used. Wrappers are used to measure classification performance based on a particular subset of features. In our experiments, SVM is used as the learning scheme and

Table 3. Feature subset selected by greedy attribute subset search and SVM-wrapper

Domain reputation	RT source reputation	No. named entities
Formality	Tweet uniqueness	% correct. spelled words
Max. no. repeat. letters	Contains numbers	No. capitalized words
No. hash-tags	No. exclam. marks	Avg. word length
Contains first-person	Is re-tweet	Is reply-tweet

Table 4. Importance of individual features based on Information Gain (IG)

IG	*Feature*	*IG*	*Feature*	*IG*	*Feature*
0.374	Domain reputation	0.078	Day in the week	0.014	Contains a measure
0.287	Contains link	0.071	Is reply-tweet	0.011	No. hashtags
0.130	Formality score	0.060	Avg. word length	0.008	No. of mentions
0.127	Num. proper names	0.042	% of correct spell.	0.007	Contains emoticons
0.113	Max. proper names	0.041	Hour of the day	0.007	Contains nums
0.111	Tweet length	0.041	Hashtag reputation	0.005	No. quest. marks
0.089	No. named entities	0.034	RT source reput.	0.003	No. capital. words
0.087	% of stopwords	0.023	Max. repeated chars.	0.001	Is re-tweet
0.083	Max. word length	0.023	Uniqueness score	0.000	Max. capital. words
0.081	Has first-person	0.019	No. excl. marks		

the performance of each feature subset is evaluated using 2-fold cross-validation on the training dataset. The optimal 15 features are shown in Table 3.

We can see that $\frac{1}{3}$ of the optimal features are linked-based, showing that the proposed link-based features (corresponding to the *navigational quality* criteria introduced in Section 3.2) contribute most to the classification, a conclusion also derived from the classification results (Section 6.2). Also, language formality and complexity features (corresponding to the *factuality* criteria) are strong indicators, as high quality tweets tend to use more formal language, named entities, etc. Some of the features, however, do not provide as useful characterization. We observe that spelling and punctuation of the tweet are not particularly strong indicators, which may be due to the informal language (e.g., short forms and abbreviations) commonly used in Twitter. Also, the number of '@username' mentions is not a strong indicator. We observe that while some tweets mentioning many users generally have lower quality (e.g. private conversations between a group of users), many high-quality tweets also mention users, such as names of public figures or organizations. Based on IG, the 'Is re-tweet' feature is not a strong indicator of high-quality tweets, aligning with our observation in Section 6.1 that even re-tweets may contain low-quality or noisy messages. In contrast, 'RT source reputation' proves to be a clearly stronger indicator, leveraging the reputation of re-tweeted users.

An overview of feature sets used for experiments is presented in Table 5.

Filtering Results We evaluate the accuracy of a SVM classifier on different feature sets for the filtering of high-quality, as well as low-quality tweets. The results are presented in Table 6.

According to the AUC results in Table 6, 'Subset.SVM' performs the best among all the feature sets, achieving the highest recall in high-quality filtering,

Table 5. Description of feature sets used in experiments

Feature set	*#Ftr's*	*Description*
Text (TFIDF)	3322	TFIDF represent. of term n-grams up to length 5. Baseline.
Link only	1	Single feature - presence of a hyperlink. Method used in [2].
C1.Spell	6	Punctuation and spelling features
C2.Comp	8	Syntactic and semantic complexity features
C3.Gram	5	Grammaticality features
C4.Links	8	Link-based features
C5.Time	2	Timestamp features
Subset.Cont	19	All content-based features (C1 - C3).
Subset.Reput	3	Reputation score features (DRS, RRS, HRS)
Subset.SVM	15	Features selected by greedy attribute selection (see Table 3)
All features	29	All content and link-based features
All ftr's + Text	3351	All content, link and TF-IDF features

Table 6. Precision (P), Recall (R) and Area Under the ROC Curve (AUC) results for the task of finding high-quality and low-quality tweets using different feature sets

	High-Quality		*Low-Quality*				*High-Qual.*		*Low-Qual.*		
Features	*P*	*R*	*P*	*R*	*AUC*	*Features*	*P*	*R*	*P*	*R*	*AUC*
Text (TFIDF)	**0.862**	0.665	0.885	0.96	0.813	Link only	0.798	0.702	0.894	0.934	0.818
Subset.Cont	0.721	0.61	0.863	0.913	0.762	C1.Spell	0.5	0.004	0.73	0.999	0.501
Subset.Reput	0.812	0.746	0.909	0.936	**0.841**	C2.Comp	0.628	0.165	0.757	0.964	0.564
Subset.SVM	0.715	**0.758**	**0.912**	0.936	**0.847**	C3.Gram	0.648	0.472	0.822	0.905	0.688
All features	0.815	0.66	0.882	0.944	0.802	C4.Links	**0.82**	0.74	0.907	0.94	**0.84**
All ftr's+text	0.739	**0.775**	**0.915**	0.899	0.837	C5.Time	0	0	0.729	1	0.5

also achieving the highest precision in low-quality filtering. Furthermore, the link-based features ('C4.Links') also archive high AUC ($AUC = 0.84$), especially the subset that contains only reputation-based features ('Subset.Reput', $AUC = 0.841$). The two sets ('C4.Links' with 8 features only, and 'Subset.Reputation' with 3 features only) outperform the 'TFIDF' method (with 3322 features). Link-based features are also useful in filtering out high-quality tweets: among the 5 feature categories (C1 - C5), 'C4.Links' yields the best precision for high-quality tweets. However, it does not yield the best recall, because we find that quite a large portion of tweets in out dataset do not contain hyperlinks (68.9%) or hashtags (85.9%), and thus 'C4.Links' features cannot be directly applied to them. Finally, 'Subset.Cont' yields relatively high precision on low-quality tweets, showing that Content-based features are fairly useful in filtering out low-quality tweets.

We observe that 'TFIDF' yields high precision for high-quality tweets, because it employs a large number of features in the classification. However, a comparison of the training time and storage space requirements (shown in Figure 6) reveals that 'TFIDF' consumes the largest amount of training time and space due to the large number of features (i.e., 3322 features). Overall, the optimal feature subset 'Subset.SVM' not only yields better overall results, but also requires less training time and space compared to the 'TFIDF' representation.

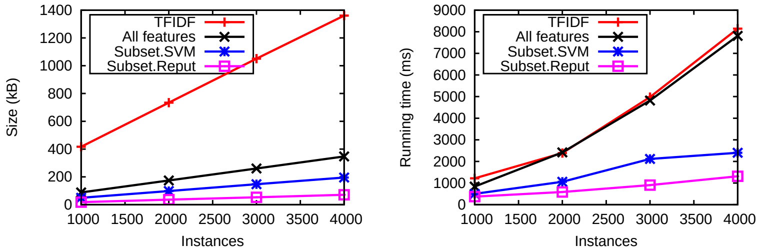


Fig. 6. Storage cost (left) and training time cost (right) of different feature sets

Table 7. Ranking accuracy in terms of NDCG@N and Mean Average Precision (MAP)

	NDCG@N						NDCG@N				
Features	*1*	*2*	*5*	*10*	*MAP*	*Features*	*1*	*2*	*5*	*10*	*MAP*
Link only	0.067	0.111	0.22	0.324	0.398	C1.Spell	0.511	0.456	0.466	0.51	0.5
Subset.Cont	0.622	0.644	0.653	0.651	0.55	C2.Comp	0.8	0.756	0.639	0.603	0.536
Subset.Reput	0.822	**0.777**	**0.777**	0.764	**0.661**	C3.Gram	0.622	0.6	0.612	0.6	0.513
Subset.SVM	**0.867**	0.767	**0.778**	**0.769**	0.653	C4.Links	0.733	0.656	0.687	0.711	0.639
All features	0.733	0.733	0.763	0.753	0.637	C5.Time	0.156	0.267	0.282	0.346	0.377

6.3 Ranking Evaluation

Experiment Methodology To evaluate our ranking method, we prepare 30 single-word test queries in the ranking evaluation. The queries are randomly selected from 5 different categories: News, Politics, Technology, Business and Entertainment, to avoid any topical bias. For each query, a set of labeled tweets containing the query term is retrieved. A total of 1,834 tweets are retrieved for the 30 test queries. We divide the 1,834 tweets into two sets, one set for the training and the other set for the Rank SVM testing. Basically, the tweets retrieved from 15 test queries are used for the training, while the tweets from the remaining 15 test queries are used for the testing. In the ranking evaluation, we use Normalized Discounted Cumulative Gain ($NDCG$) and Mean Average Precision (MAP), which are standard metrics for evaluating ranking accuracy.

Ranking results We evaluate the ranking model with different features as shown in Table 7. We observe that 'Subset.Reput' (the set of reputation-based features) and 'Subset.SVM' achieve the overall best results ($MAP = 0.661$ and 0.653), significantly outperforming the 'Link only' feature used in [2]. Furthermore, among the 5 sets of features (C1 - C5) proposed in Section 4.2, link-based features achieve the best results. This aligns with our observations in Section 6.2 that link-based features (especially the three reputation-based features) are useful for identifying high-quality tweets.

7 Conclusion

In this paper, we study the problem of finding high quality content in Twitter. We formulate the criteria of quality tweets and tackle the filtering and tweet ranking problems. The quality of a tweet is modelled using a set of features based on the tweet's content, as well as links to websites, hashtags and users.

Our proposed link-based features are able to boost the filtering and ranking performance, indicating that the implicit "reputation" of a web domain, hashag or re-tweeted user is highly useful in the filtering and ranking tasks. In our experiments, the optimal feature subset that includes link-based features achieves the best overall classification and ranking accuracy.

Although we focus on Twitter in this work, the results are potentially useful in the contexts of other social networks and microblogging services. For future work, we plan to consider different types of queries in Twitter (e.g. hot topic queries, movie reviews, highly factual seeking queries) and study the importance of tweet features for filtering and ranking in these different scenarios.

References

1. TwitterEngineering, "200 million tweets per day." [Online]. Available: http://blog.twitter.com/2011/06/200-million-tweets-per-day.html
2. O. Alonso, C. Carson, D. Gerster, X. Ji, and S. Nabar, "Detecting Uninteresting Content in Text Streams," in *Proc. of SIGIR CSE Workshop*, 2010.
3. A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *Proc. of WebKDD/SNA-KDD*, 2007.
4. D. Zhao and M. B. Rosson, "How and why people twitter: the role that microblogging plays in informal communication at work," in *Proc. of GROUP*, 2009.
5. B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in *Proc. of SIGIR Conference*, 2010.
6. D. Ramage, S. T. Dumais, and D. J. Liebling, "Characterizing microblogs with topic models," in *Proc. of ICWSM Conference*, 2010.
7. H. Lauw, A. Ntoulas, and K. Kenthapadi, "Estimating the quality of postings in the real-time web," in *Proc. of SSM Conference*, 2010.
8. R. Nagmoti, A. Teredesai, and M. De Cock, "Ranking approaches for microblog search," in *Proc. of WI-IAT Conference*, 2010.
9. M. Trifan and D. Ionescu, "A new search method for ranking short text messages using semantic features and cluster coherence," in *Proc. of ICCC-CONTI*, 2010.
10. A. D. Sarma, A. D. Sarma, S. Gollapudi, and R. Panigrahy, "Ranking mechanisms in twitter-like forums," in *Proc. of WSDM Conference*, 2010.
11. Y. Duan, L. Jiang, T. Qin, M. Zhou, and H.-Y. Shum, "An empirical study on learning to rank of tweets," in *Proc. of COLING Conference*, 2010.
12. J. V. Grove, "Twitter attempts to personalize 1.6 billion search queries per day." [Online]. Available: http://mashable.com/2011/06/01/twitter-search-queries/
13. S. Lahiri, P. Mitra, and X. Lu, "Informality judgment at sentence level and experiments with formality score," in *Proc. of CICLing Conference*, 2011.
14. T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. of ACM SIGKDD Conference*, 2002.
15. Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proc. of CIKM Conference*, 2010.
16. R. Kohavi and G. H. John, "Wrappers for feature subset selection," *ARTIFICIAL INTELLIGENCE*, 1997.