Re-tweeting from a Linguistic Perspective

Aobo Wang **Tao Chen** **Min-Yen Kan**
Web IR / NLP Group (WING)
National University of Singapore
13 Computing Link, Singapore 117590
{wangaobo,taochen,kanmy}@comp.nus.edu.sg

Abstract

What makes a tweet worth sharing? We study the content of tweets to uncover linguistic tendencies of shared microblog posts (re-tweets), by examining surface linguistic features, deeper parse-based features and Twitter-specific conventions in tweet content. We show how these features correlate with a functional classification of tweets, thereby categorizing people's writing styles based on their different intentions on Twitter. We find that both linguistic features and functional classification contribute to re-tweeting. Our work shows that opinion tweets favor originality and pithiness and that update tweets favor direct statements of a tweeter's current activity. Judicious use of #hashtags also helps to encourage retweeting.

1 Introduction

Tweeting[1] is a modern phenomenon. Complementing short message texting, instant messaging, and email, tweeting is a public outlet for netizens to broadcast themselves. The short, informal nature of tweets allows users to post often and quickly react to others' posts, making Twitter an important form of close-to-real-time communication.

Perhaps as a consequence of its usability, form, and public nature, tweets are becoming an important source of data for mining emerging trends and opinion analysis. Of particular interest are retweets, tweets that share previous tweets from others. Tweets with a high retweet count can be taken as a first cut towards trend detection.

It is known that social network effects exert marked influence on re-tweeting (Wu et al., 2011; Recuero et al., 2011). But what about the content of the post? To the best of our knowledge, little is known about what properties of tweet content motivate people to share. Are there content signals that mark a tweet as important and worthy of sharing?

To answer these questions, we delve into the data, analyzing tweets to better understand posting behavior. Using a classification scheme informed by previous work, we annotate 860 tweets and propagate the labeling to a large 9M corpus (Section 2). On this corpus, we observe regularities in emoticon use, sentiment analysis, verb tense, named entities and hashtags (Section 3), that enable us to specify feature classes for re-tweet prediction. Importantly, the outcome of our analysis is that a single holistic treatment of tweets is suboptimal, and that re-tweeting is better understood with respect to the specific function of the individual tweet. These building blocks allow us to build a per-function based re-tweet predictor (Section 4) that outperforms a baseline.

2 Linguistically Motivated Tweet Classification

Before we can label tweets for more detailed classification, we must decide on a classification scheme. We first study prior work on tweet classification before setting off on creating our own classification for linguistic analysis.

This research is supported by the Singapore National Research Foundation under its International Research Centre Singapore Funding Initiative and administered by the IDM Programme Office, under grant 252-002-372-490.

[1] More generally known as microblogging, in which the post is termed a microblog.

Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012), pages 46–55,
Montréal, Canada, June 7, 2012.

Early ethnographic work on tweets manually created classification schemes based on personal, direct observation (Java et al., 2009; Kelly, 2009). Other work is more focused, aiming to use their constructed classification scheme for specific subsequent analysis (Naaman et al., 2010; Sriram et al., 2010; Ramage et al., 2010; Chen et al., 2010). All schemes included a range of 5–9 categories, and were meant to be exhaustive. They exhibit some regularity: all schemes included categories for information sharing, opinions and updates. They vary on their classification's level of detail and the intent of the classification in the subsequent analysis.

Most closely related to our work, Naaman et al. (2010) focused on distinguishing salient user activity, finding significant differences in posts about the tweeting party or about others that were reported by manually classifying tweets into nine categories, sampled from selected users. However, while their paper gave a useful classification scheme, they did not attempt to operationalize their work into an automated classifier.

Other works have pursued automated classification. Most pertinent is the work by Sriram et al. (2010), who applied a Naïve Bayes learning model with a set of 8 features (author ID, presence of shortened words, "@username" replies, opinionated words, emphasized words, currency and percentage signs and time phrases) to perform hard classification into five categories. To identify trending topics, Zubiaga et al. (2011) performed a similar classification, but at the topic level (as opposed to the individual tweet level) using aggregated language-independent features from individual tweets. Ramage et al. (2010) introduced four salient dimensions of tweets – style, status, social, substance. Individual terms and users were characterized by these dimensions, via labeled LDA, in which multiple dimensions could be applied to both types of objects.

While the previous work provides a good overview of the genre and topic classification of tweets, their analysis of tweets have been linguistically shallow, largely confined to word identity and Twitter-specific orthography. There has been no work that examines the discoursal patterns and content regularities of tweets. Understanding microblog posts from a deeper linguistic perspective may yield insight into the latent structure of these posts, and be useful for trend prediction. This is the aim of our work.

2.1 Classification Scheme

We hypothesize that people's intentions in posting tweets determine their writing styles, and such intentions can be characterized by the content and linguistic features of tweets. To test this hypothesis, we first collect a corpus of manually annotated tweets and then analyze their regularities. In constructing our classification annotation scheme, we are informed by the literature and adopt a two-level approach. Our coarser-grained Level-1 classification generalization is an amalgam of the schemes in Naaman et al. and Sriram et al.'s work; while our finer-grained, Level-2 classification further breaks down the Update and Opinion classes, to distinguish linguistic regularities among the subclasses. The left two columns of Table 1 list the categories in our scheme, accompanied by examples.

2.2 Dataset Collection

We collected three months of public tweets (from July to September in 2011) through Twitter's streaming API[2]. Non-English tweets were removed using regular expressions, incurring occasional errors. We note that tweets containing URLs are often spam tweets or tweets from automated services (e.g., Foursquare location check-ins) (Thomas et al., 2011), and that any retweet analysis of such tweets would need to focus much more on the linked content rather than the tweet's content. We thus removed tweets containing URLs from our study. While this limits the scope of our study, we wanted to focus on the (linguistic quality of) content alone. The final dataset explicitly identifies 1,558,996 retweets (hereafter, *RT-data*) and 7,989,009 non-retweets. To perform further analysis on Twitter hashtags (i.e., "#thankyousteve"), we break them into separate words using the Microsoft Data-Driven Word-Breaking API[3]. This also benefits the classification task in terms of converting hashtags to known words.

[2] http://dev.twitter.com/docs/streaming-api

[3] http://web-ngram.research.microsoft.com/info/break.html

Table 1: Our two-level classification with example tweets.

Level-1	Level-2	Motivation	Example retweets	Corpus count (%)
Opinion	*Abstract*	Present opinions towards abstract objects.	God will lead us all to the right person for our lives. Have patience and trust him.	291 (33.8%)
	Concrete	Present opinions towards concrete objects.	i feel so bad for nolan. Cause that poor kid gets blamed for everything, and he's never even there.	99 (11.5%)
	Joke	Tell jokes for fun.	Hi. I'm a teenager & I speak 3 languages: English, Sarcasm, & Swearing (; #TeenThings	86 (10.0%)
Update	*Myself*	Update my current status.	first taping day for #growingup tomorrow! So excited. :)	168 (19.6%)
	Someone	Update others' current status.	My little sister still sleep ...	66 (7.7%)
Interaction		Seek interactions with others.	#Retweet If you're #TeamFollowBack	81 (9.4%)
Fact		Transfer information.	Learnt yesterday: Roman Empire spent 75% of GDP on infrastructure. Roads, aqueducts, etc.	23 (2.7%)
Deals		Make deals.	Everybody hurry! Get to Subway before they stop serving LIMITED TIME ONLY item 'avocados'.	29 (3.4%)
Others		Other motivations.	Ctfu Lmfao At Kevin Hart ;)	17 (2.0%)

We employed U.S.-based workers on Amazon's Mechanical Turk to annotate a random subset of the preprocessed tweets. We collected annotations for 860 tweets (520 retweets; 340 non-retweets) randomly sampled from the final dataset, paying 10 cents per block of 10 tweets labeled. Each tweet was labeled by 3 different workers who annotated using the Level-2 scheme. Gold standard labels were inferred by majority. Inter-annotator agreement via Fleiss' κ showed strong (0.79) and modest (0.43) agreement at Level-1 and Level-2, respectively.

Table 1's rightmost columns illustrate the distribution of the annotated tweets on each category. From our Level-1 classification, *Opinion*, *Update* and *Interaction*, make up the bulk of the tweets in the annotated sample set. The remaining categories of *Facts*, *Deals* and *Others* make up only 8.1% in total. We thus focus only on the three major groups.

2.3 Labeled LDA Classification

Given the labeled data, we first observed that tweets in different classes have different content and language usage patterns. For example, tweets belonging to *Opinion* display more of an argumentative nature, exhibiting a higher use of second person pronouns (e.g., "you", "your"), modal verbs (e.g., "can", "could", "will", "must"), and particular adverbs (e.g., "almost", "nearly") than the other two groups. These observations lead us to employ the classifier that make use of words' co-occurrence feature to categorize tweets.

Hence, we adopt Labeled LDA, which extends Latent Dirichlet Allocation (LDA) (Blei et al., 2003) by incorporating supervision at the document level (here, tweet-level), enabling explicit models of text content associated with linguistic features. In adopting this methodology, we follow (Ramage et al., 2009) previous work on tweet classification. Features are encoded as special tokens to not overlap the tokens from the tweet content.

Tweets arguing in one style tend to share similar linguistic features. For example in Table 1, *Update* talks about ongoing events using present tense; and *Opinion* uses conjunctions to compose and connect ideas. To discover how people talk differently across genres of tweets, we extract five sets of linguistic features from each tweet, namely *Tense*[4], *Discourse Relations*[5], *Hashtags*, *Named Entities*[6], and *Interaction Lexical Patterns*[7].

We use default parameter settings for Labeled LDA. All the combinations of features were tested to find the best performing feature set. Table 2 quantifies the contribution of each feature and demonstrate the result from the best combination, as measured by Weighted Average F-Measure (WAFM). Compared to the performance of using baseline feature set using tweet content alone, the use of linguistic features improve the performance accordingly, with the exception of the use of named entities which reduced performance slightly, and hence was removed from the final classifier's feature set.

[4]Using the OpenNLP toolkit.

[5]Using (Lin et al., 2010)'s parser.

[6]Using the UW Twitter NLP tools (Ritter et al., 2011).

[7]Defined as Boolean matches to the following regular expressions: *"RT @[username]..."*, *"...via @[username]..."*, *"Retweeting @[username]..."*,*"Follow me if..."*, *"retweet @[username]..."*, *"...RT if..."* and *"Retweet if..."*

Scheme	C	CI	CT	CD	CH	CE	CITDH
Level-1	.625	.642	.635	.637	.629	.611	.670
Level-2	.413	.422	.427	.432	.415	.409	.451

Table 2: Weighted average F-measure results for the labeled LDA classification. Legend: C: tweet context; I: *Interaction*; T: *Tense*; D: *Discourse Relations*; H: *Hashtags*; E: *Named Entities*.

Require: Training set L**; Test collection** C**; Evaluation set** E**; Iteration count** I

function incrementalTraining(L,C,E,)
  $M \leftarrow$ labeledLDATraining(L)
  $e \leftarrow$ evaluate(M,E)
  for $c_i \epsilon C$ and $i < I$ do
    $r_i \leftarrow$ predictLabel(c_i,M)
    $r_{selected} \leftarrow$pickItemsWithHighConfidence(r_i);
    $L' \leftarrow$ add($r_{selected}$) into L
    $M' \leftarrow$ retrainLDAModel(L')
    $e' \leftarrow$ evaluate(M',E)
    if e' is better than e then $M \leftarrow m'$; $e \leftarrow e'$;
    else return M
    $i \leftarrow i + 1$
    keepLog(e')
  return M

Figure 1: Pseudocode for incremental training.

2.4 Automated Classification

Starting with the best performing model trained on the *Level-1* schema (the CITDH feature set), we automatically classified the remaining tweets, using the incremental training algorithm described in Figure 1. The 860 annotated tweets were randomly split into a training set L and evaluation set E with a 5:1 ratio. The 9M unannotated tweets form the test collection C. c_i is assigned by randomly selecting 1000 tweets from C. I is computed as the size of C divided by the size of c_i. Note that retraining becomes more expensive as the dataset L' grows. Thus, we greedily generate a locally-optimal model, which completes after 6 iterations.

From the result of automatically labeled dataset, we see that the *Opinion* dominates the collection in count (44.6%), followed by *Interaction* (28.4%) and *Update* (20.5%). This result partially agrees with the manual classification results in Naaman et al. (2010), but differs in their *Information Sharing* category, which is broken down here as *Facts*, *Deals* and *Others*. We believe the discrepancies are due to the differences between the two datasets used. Their retweets were sampled from selected users who are active participants, and did not include tweets from organizations, marketers and dealers; in our case, the tweets are generally sampled without constraints.

3 Analysis of Linguistic Features

We now dissect retweets using the 1.5M *RT-data* defined in Section 2.2. We do this from a linguistic perspective, based on observations on the values and correlations among the features used for the automatic classification.

3.1 Emoticons and Sentiment

Emoticons such as smilies – `:)` – and frownies – `:(` – and their typographical variants, are prevalent in tweets. Looking at the distribution of emoticons, we find that 2.88% of retweets contain smilies and 0.26% contain frownies. In other words, smileys are used more often than frownies.

To give an overall picture of how sentiment is distributed among retweets, we employed the *Twitter Sentiment* Web API service (Go et al., 2009) to obtain polarity. Figure 2 shows that while neutral tweets dominate in all three classes, there are more negative tweets in the *Interaction* than in the other two. Such negative interactive comments usually find their use in sharing negative experiences in a dialogue or with their followers. *"Yeah I hate talking IN my phone. RT @Jadon Don't you guys hate talking in the phone"* is a representative example.

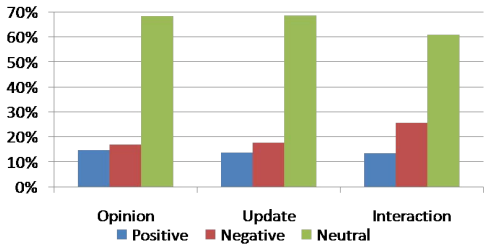


Figure 2: Sentiment distribution of retweets.

Previous works have leveraged emoticons to automatically build corpora for the sentiment detection task, through labeling tweets with smilies (frownies) as true positive (negative) instances (Read, 2005; Alexander and Patrick, 2010; Cui et al., 2011), and training statistical classification models on the result. We wish to verify the veracity of this hypothesis. Do emoticons actually reflect sentiment in

Table 3: Manual sentiment annotation results and confusion matrix. Bolded numbers highlight the error caused by neutral posts.

	Positive	Neutral	Negative
Retweets with smilies	55 (27.5%)	**140 (70%)**	5 (2.5%)
Retweets with frownies	9 (4.5%)	**118 (59%)**	73(36.5%)
Predicted Positive	43	**30**	0
Predicted Neutral	11	206	12
Predicted Negative	7	**29**	62

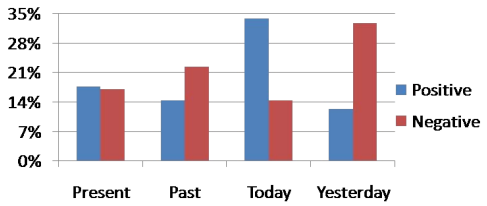


Figure 3: Tenses (l) and specific times (r) and their sentiment.

retweets? To answer the question, we randomly subselected 200 retweets with smilies and another 200 with frownies from *RT-data*, and then manually labeled their sentiment class after removing the emoticons. Table 3's top half shows the result.

While our experiment is only indicative, neutral posts are still clearly the majority, as indicated by bold numbers. Simply labeling the sentiment based on emoticons may mistake neutral posts for emotional ones, thus introducing noise into training data. *"Fishers people have no idea how lawrence kids are, guess they do now :)"* is such an example.

To demonstrate this effect, we evaluated Go et al. (2009)'s API on our annotated corpus. We present the confusion matrix in bottom half of Table 3. A common error is in mistaking neutral tweets as positive or negative ones, as indicated by the bold numbers. Given that the detector is trained on the corpus, in which neutral tweets with smiles (frownies) are labeled as positive (negative) ones, the detector may prefer to label neutral tweets as sentiment-bearing. This observation leads us to believe that more careful use of emoticons could improve sentiment prediction for tweets and microblog posts.

3.2 Verb Tense

We analyze the tense of the verbs in retweets, using a simplified inventory of tenses. We assign two tenses to verbs: past and present. Tense is assigned per-sentence; tweets that consist of multiple sentences may be assigned multiple tenses. Based on our statistics, one notable finding is that *Update* has a higher proportion of past tense use (33.70%) than *Opinion* (14.9%) and *Interaction* (24.2%). This validates that updates often report past events and verb tense is a more crucial feature for *Updates*.

Building on the previous section, we ask ourselves whether sentiment is correlated with verb tense use. Interestingly, the results are not uniform. Figure 3 shows our analysis of positive and negative (omitting neutral) sentiments as they co-occur with verb tense in our corpus. It shows that people tend to view the past negatively (e.g., *"I dont regret my past, I just regret the times I spent with the wrong people"*), whereas emotions towards current event do not have any obvious tendency. A case in point is in the use of "*today*" and "*yesterday*" as time markers related to present and past use. Figure 3 shows the number of tweets exhibiting these two words and their sentiment. The results are quite marked: tweets may be used to complain about past events, but look optimistically about things happening now.

3.3 Named Entities

To study the diversity of named entities (NEs) in retweets, we used UW Twitter NLP Tools (Ritter et al., 2011) to extract NEs from *RT-data*. 15.9% of retweets contain at least one NE, indicating that NEs do play a large role in retweets.

So what types of NEs do people mention in their tweets? From each of our primary Level-1 classes, we selected the top 100 correctly recognized NEs, in descending order of frequency. We then standardized variants (i.e. "*fb*" as a variant of "*Facebook*"), and manually categorized them against the 10-class schema defined by Ritter et al. (2011).

Table 4: The distribution of top 100 named entities[8].

Class	Opinion	Update	Interaction
PERSON	41.2%	44.7%	38.8%
GEO-LOC	7.8%	28.9%	25.4%
COMPANY	15.7%	6.6%	10.4%
PRODUCT	5.9%	5.3%	6.0%
SPORTS-TEAM	2.0%	5.3%	1.5%
MOVIE	7.8%	5.3%	7.5%
TV-SHOW	3.9%	0.0%	3.0%
OTHER	15.7%	3.9%	7.5%

Table 4 displays the distribution of the different classes of NEs, by frequency. People's names represent the largest portion in each class, of which the majority are celebrities. Geographical locations – either countries or cities – make up the second largest class for *Update* and *Interaction*, accounting for 28.9% and 25.4%, respectively, whereas they take only 7.8% of *Opinion*. A possible reason is that people prefer to broadcast about events (with locations mentioned) or discuss them through *Update* and *Interaction* classes, respectively. *"California, I'm coming home."* is a typical example.

3.4 Hashtags

Previous work (Cunha et al., 2011) showed that popular hashtags do share common characteristics, such as being short and simple. We want to push more in our analysis of this phenomenon. We organize our hashtag analysis around three questions: (a) Do people have any positional preference for embedding hashtags? (b) Are there any patterns to how people form hashtags? and (c) Is there any relationship between such patterns and their placement?

To answer these questions, as shown in Table 5, we extracted the hashtags from *RT-data* and categorized them by the position of their appearance (at the beginning, middle, or end) of tweet. 69.1% of hashtags occur at the end, 27.0% are embedded in the middle, and 8.9% occur at the beginning. In Figure 4, we plot the frequency and length (in characters) of the hashtags with respect to their position, which shows that the three placement choices lead to different distributions. Beginning hashtags (hereafter, beginners) tend to peak around a length of 11, while middlers peaked at around 7. Enders feature a bimodal distribution, favoring short (3) or longer (11+) lengths. We found these length distributions are artifacts of how people generate and (functionally) use the hashtags.

Beginners are usually created by concatenating the preceding words of a tweet, therefore, the common patterns are subject+verb (e.g.,"*#IConfess*"), subject+verb+object (e.g., "*#ucanthaveme*"), and similar variants. Middlers, often acting as a syntactic constituent in a sentence, are usually used

[8]The other two classes, *facility* and *band*, are not found in the top 100 NEs.

Table 5: Hashtags and example tweets.

Position	Tweets
Beginning	**#ihateitwhen** random people poke you on facebook
Middle	I just saw the **#Dodgers** listed on Craig's List.
End	Success is nothing without someone you love to share it with. **#TLT**
	Goodmorning Tweethearts....wishing u all blessed and productive day! **#ToyaTuesday**

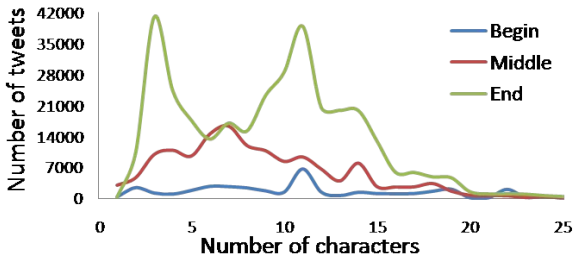


Figure 4: Length distribution of sampled hashtags.

to highlight tweet keywords, which are single-word nouns (e.g.,"*#Scorpio*" and "*#Dodgers*"). Enders provide additional information for the tweets. A popular ender pattern is Twitter slang that have been used enough to merit their own Twitter acronym, such as "*#TFB*" (Team Follow Back), and "*#TLT*" (Thrifty Living Tips). Another popular form is concatenating multiple words, indicating the time ("*#ToyaTuesday*"), the category ("*#Tweetyquote*") or the location ("*#MeAtSchool*"). Knowing such hashtag usage can aid downstream applications such as hashtag suggestion and tweet search.

3.5 Discourse Relations

In full text, textual units such as sentences and clauses work together to transmit information and give the discourse its argumentive structure. How important is discourse in the microblog genre, given its length limitation? To attempt an answer to this question, we utilized the end-to-end discourse parser proposed by Lin et al. (2010) to extract PDTB-styled discourse relations (Prasad et al., 2008) from *RT-data*. Figure 5 shows the proportion of the five most frequent relations. 68.0% of retweets had at least one discourse relation – per class, this was 55.2% of *Opinion*, 44.7% of *Interaction*, and 21.6% of *Update*. Within *Opinions*, we find that negative opinions are often expressed using a *Synchrony* relation (i.e., negative tweet: "*I hate when I get an itch at a*

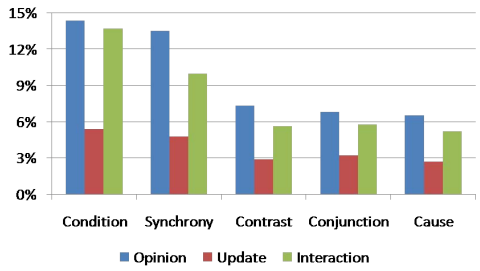


Figure 5: The distribution of five selected discourse relations.

place where my hand can't reach."), while positive and neutral opinions prefer *Condition* relations (i.e., positive tweet: "*If I have a girlfriend :) I will tell her beautiful everyday.*").

3.6 Sentence Similarity

"*On Twitter people follow those they wish they knew. On Facebook people follow those they used to know.*"

We round out our analysis by examining the sentence structure of retweets. Sometimes it is not what you say but on how you say it. This adage is especially relevant to the *Opinion* class, where we observed that the craftiness of a saying influences its "retweetability". This can be reflected in tweets having parallel syntactic structure, which can be captured by sentence similarity within a tweet, as illustrated in the quote/tweet above. We employ the *Syntactic Tree Matching* model proposed by Wang et al. (2009) on tweets to compute this value. This method computes tree similarity using a weighted version of tree kernels over the syntactic parse trees of input sentences. When we set the similarity threshold to 0.2 (determined by observation), 723 retweets are extracted from the *Opinion* class of which over 500 (70%) are among the top 5% most retweeted posts (by count). Examining this set reveals that they are more polarized (22.6% positive, 23.2% negative) than the average *Opinion* (14.7% and 16.9%, respectively).

4 Predicting Retweets

Given the diversity in function which we have illustrated in our linguistic analyses in the previous sections, we argue that whether a tweet is shared with others is best understood by modeling each function (Level-1) class independently. We validate this claim here, by showing how independently building classification models for the *Opinion*, *Update* and *Interaction* classes outperforms an agglomerated retweet predictor.

Previous research have found that features representing the author's profile (e.g., number of followers), tweet metadata (time interval between initial posting and current checkpoint, previously retweeted) and Twitter-specific features (URL presence) weight heavily in predicting retweets (Suh et al., 2010; Peng et al., 2011; Hong et al., 2011). In contrast, our study is strictly about the content and thus asks the question whether retweeting can be predicted from the content alone.

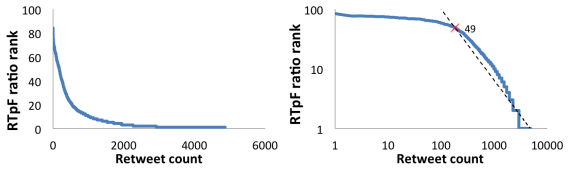
Before we do so, we call attention to a caveat about retweet prediction that we feel is important and unaccounted for in previous work: the actual probability of retweet is heavily dependent on how many people view the tweet. Twitter tracks the follower count of the tweet's author, which we feel is the best approximation of this. Thus we do not perform retweet count prediction, but instead cast our task as:

Given the content of a tweet, perform a multi-class classification that predicts its range of retweet per follower (RTpF) ratio.

4.1 Experiment and Results

We first examine RTpF distribution over the 9M tweets in the dataset. Figure 6 plots RTpF rank against retweet count on both normal and log-log scales. While the normal scale seems to show a typical exponential curve, the log-log scale reveals a clear inflection point that corresponds to an RTpF of 0.1. We use this inflection point to break the predicted RTpF values into three ordinal classes: no retweets ("N", RTpF = 0), low ("L", RTpF < 0.1), and high ("H", RTpF ≥ 0.1).

We use 10-fold cross validation logistic regression in Weka3 (Hall et al., 2009) to learn prediction models. The regression models use both binary presence-of feature classes (quotation; past, present tense; 16 types of discourse relations; 10 NE types; 3 hashtag positions) as well as normalized numeric features (tweet length, hashtag count, sentence similarity, 3 sentiment polarity strengths). Note that the models reported here do not factor the content (lexi-



(a) Normal scale (b) Logarithmic scale

Figure 6: Retweet per follower (RTpF) ratio rank versus retweet count. The highlighted point shows the boundary between classes H and L.

Class	$\bar{F}_1$	Salient Features	Feature Weight
Opinion	0.57	Sentence Similarity	10.34
		Conjunction	-21.09
		Quotation	-19.2
Update	0.54	Sentence Similarity	-2.81
		Past	-5.2
		Present	1.3
Interaction	0.53	Sentence Similarity	-55.33
		Hashtag Count	5.34
All w/ L-1 class	0.52	Sentence Similarity	9.8
All w/o L-1 class	0.42	Hashtag Count	22.03

Table 6: Logistic regression results. Salient features also shown with their respective weight, where a +ve value denotes a +ve contribution to retweet volume.

cal items) directly, but represent content through the lens of the feature classes given.

We build individual regression models for the three major Level-1 classes, and aggregate models that predict RTpF for all three classes. The two aggregate models differ in that one is informed of the Level-1 class of the tweets, while the other is not. We report average F-measure in Table 6 over the three RTpF classes ("N", "L" and "H"). Adding the Level-1 classification improves the RTpF prediction result by 10% in terms of average F_1. This results validate our hypothesis – we see that building separate logistic models for each class improves classification results uniformly for all three classes.

4.2 Remarks

We make a few conjectures based on our observations, in concluding our work:

1. Getting your *Opinion* retweeted is easier when your readership feels a sense of originality, pithiness and wittiness in your post. "*If you obey all the rules, you miss all the fun - Katharine Hepburn*" exemplifies these factors at conflict: while being witty in exhibiting parallel syntactic structure (high sentence similarity), it has a low RTpF. Perhaps followers are unsurprised when they find such beautiful words are not originally the poster's. Tweets having complex conjoined components and multiple clauses also exhibit a negative RTpF tendency – find a short and simple way of getting your message across.

2. *Update* tweets show the least bias towards any particular feature, exhibiting little weight towards any one convention. Update tweets prefer simple tenses, eschewing perfect and progressive variants. Perhaps followers are more curious about what you are doing now but not what you have done.

3. Sentence similarity negatively affects retweeting among *Interaction* tweets. This implies that people prefer direct sounds to well-designed proverbs in the daily interaction, which is mostly in the form of question answering or voting.

4. Globally, the presence and count of hashtags is correlated with retweeting, but this effect is greatly lessened when Level-1 class features are used. This further validates the importance of our functional classification of tweets.

5 Conclusion

People tweet for different reasons. Understanding the function of the tweet is interesting in its own right, but also useful in predicting whether it will be shared with others. We construct a two-level classification informed by prior work and have annotated a corpus of 860 tweets.

Employing Labeled LDA, we propagated our annotations to a large 9M tweet corpus and investigated the linguistic characteristics of the 1.5M retweets. We created a model to predict the level of retweeting per follower given a tweet's content.

Finally, to further encourage investigation on these topics, we have made the annotated corpus and the two tools described in this paper – the functional classifier and the retweet predictor – available to the public to test and benchmark against[9].

In future work, we plan to combine the content analysis from this study with known social, time and linked URL features to see whether content features can improve a holistic model of retweeting.

[9] http://wing.comp.nus.edu.sg/tweets/

References

Pak Alexander and Paroubek Patrick. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).

David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:2003.

Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed H. Chi. 2010. Short and tweet: Experiments on recommending content from information streams. In *CHI 2010*.

Anqi Cui, Min Zhang, Yiqun Liu, and Shaoping Ma. 2011. Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. In *AIRS*, volume 7097 of *Lecture Notes in Computer Science*, pages 238–249. Springer.

Evandro Cunha, Gabriel Magno, Giovanni Comarela, Virgilio Almeida, Marcos André Gonçalves, and Fabricio Benevenuto. 2011. Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 58–65, Portland, Oregon, June. ACL.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classication using distant supervision. Technical report, Stanford University.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Liangjie Hong, Ovidiu Dan, and Brian D. Davison. 2011. Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 57–58, New York, NY, USA. ACM.

Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2009. Why we twitter: An analysis of a microblogging community. In *Proceedings of WebKDD/SNA-KDD 2007*, volume 5439 of *LNCS*, pages 118–138.

Ryan Kelly. 2009. Twitter study august 2009: Twitter study reveals interesting results about usage. `http://www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf`, August.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. A pdtb-styled end-to-end discourse parser. Technical report, School of Computing, National University of Singapore.

Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. 2010. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, pages 189–192, New York, NY, USA. ACM.

Huan-Kai Peng, Jiang Zhu, Dongzhen Piao, Rong Yan, and Ying Zhang. 2011. Retweet modeling using conditional random fields. In *ICDM 2011 Workshop on Data Mining Technologies for Computational Collective Intelligence*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of LREC*.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA. ACL.

Daniel Ramage, Susan Dumais, and Dan Liebling. 2010. Characterizing microblogs with topic models. *International AAAI Conference on Weblogs and Social Media*, 5(4):130–137.

Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, pages 43–48, Stroudsburg, PA, USA. ACL.

Raquel Recuero, Ricardo Araujo, and Gabriela Zago. 2011. How does social capital affect retweets? In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 841–842, New York, NY, USA. ACM.

Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, SOCIALCOM '10, pages 177–184, Washington, DC, USA. IEEE Computer Society.

Kurt Thomas, Chris Grier, Justin Ma, Vern Paxson, and Dawn Song. 2011. Design and evaluation of a real-time url spam filtering service. In *Proceedings of the 2011 IEEE Symposium on Security and Privacy*, SP '11, pages 447–462, Washington, DC, USA. IEEE Computer Society.

Kai Wang, Zhaoyan Ming, and Tat-Seng Chua. 2009. A syntactic tree matching approach to finding similar questions in community-based qa services. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 187–194, New York, NY, USA. ACM.

Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Who says what to whom on twitter. In *Proceedings of the World Wide Web Conference*.

Arkaitz Zubiaga, Damiano Spina, Víctor Fresno, and Raquel Martínez. 2011. Classifying trending topics: a typology of conversation triggers on twitter. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 2461–2464, New York, NY, USA. ACM.