The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter

Chenhao Tan
Dept. of Computer Science
Cornell University
chenhao@cs.cornell.edu

Lillian Lee
Dept. of Computer Science
Cornell University
llee@cs.cornell.edu

Bo Pang
Google Inc.
bopang42@gmail.com

Abstract

Consider a person trying to spread an important message on a social network. He/she can spend hours trying to craft the message. Does it actually matter? While there has been extensive prior work looking into predicting popularity of social-media content, the effect of wording per se has rarely been studied since it is often confounded with the popularity of the author and the topic. To control for these confounding factors, we take advantage of the surprising fact that there are many pairs of tweets containing the *same* url and written by the *same* user but employing different wording. Given such pairs, we ask: which version attracts more retweets? This turns out to be a more difficult task than predicting popular topics. Still, humans can answer this question better than chance (but far from perfectly), and the computational methods we develop can do better than both an average human and a strong competing method trained on non-controlled data.

1 Introduction

How does one make a message "successful"? This question is of interest to many entities, including political parties trying to *frame* an issue (Chong and Druckman, 2007), and individuals attempting to make a point in a group meeting. In the first case, an important type of success is achieved if the national conversation adopts the rhetoric of the party; in the latter case, if other group members repeat the originating individual's point.

The massive availability of online messages, such as posts to social media, now affords researchers new means to investigate at a very large scale the factors affecting message propagation, also known as adoption, sharing, spread, or virality. According to prior research, important features include characteristics of the originating author (e.g., verified Twitter user or not, author's messages' past success rate), the author's social network (e.g., number of followers), message timing, and message content or topic (Artzi et al., 2012; Bakshy et al., 2011; Borghol et al., 2012; Guerini et al., 2011; Guerini et al., 2012; Hansen et al., 2011; Hong et al., 2011; Lakkaraju et al., 2013; Milkman and Berger, 2012; Ma et al., 2012; Petrović et al., 2011; Romero et al., 2013; Suh et al., 2010; Sun et al., 2013; Tsur and Rappoport, 2012). Indeed, it's not surprising that one of the most retweeted tweets of all time was from user BarackObama, with 40M followers, on November 6, 2012: "Four more years. [link to photo]".

Our interest in this paper is the effect of alternative message *wording*, meaning *how* the message is said, rather than what the message is about. In contrast to the identity/social/timing/topic features mentioned above, wording is one of the few factors directly under an author's control when he or she seeks to convey a **fixed** piece of content. For example, consider a speaker at the ACL business meeting who has been tasked with proposing that Paris be the next ACL location. This person cannot on the spot become ACL president, change the shape of his/her social network, wait until the next morning to speak, or campaign for Rome instead; but he/she can craft the message to be more humorous, more informative, emphasize certain aspects instead of others, and so on. In other words, we investigate whether a different choice of words affects message propagation, *controlling for user and topic*: would user BarackObama have gotten significantly more (or fewer) retweets if he had used some alternate wording to announce his re-election?

Although we cannot create a parallel universe

Table 1: Topic- and author-controlled (TAC) pairs. Topic control = inclusion of the same URL.

author	tweets	#retweets
natlsecuritycnn	t_1: FIRST ON CNN: After Petraeus scandal, Paula Broadwell looks to recapture 'normal life.' **http://t.co/qy7GGuYW**	$n_1 = 5$
	t_2: First on CNN: Broadwell photos shared with Security Clearance as she and her family fight media portrayal of her **[same URL]**	$n_2 = 29$
ABC	t_1: Workers, families take stand against Thanksgiving hours: **http://t.co/J9mQHiIEqv**	$n_1 = 46$
	t_2: Staples, Medieval Times Workers Say Opening Thanksgiving Day Crosses the Line **[same URL]**	$n_2 = 27$
cactus_music	t_1: I know at some point you've have been saved from hunger by our rolling food trucks friends. Let's help support them! **http://t.co/zg9jwA5j**	$n_1 = 2$
	t_2: Food trucks are the epitome of small independently owned LOCAL businesses! Help keep them going! Sign the petition **[same URL]**	$n_2 = 13$

in which BarackObama tweeted something else[1], fortunately, a surprising characteristic of Twitter allows us to run a fairly analogous *natural experiment*: external forces serendipitously provide an environment that resembles the desired controlled setting (DiNardo, 2008). Specifically, *it turns out to be unexpectedly common for the same user to post different tweets regarding the same URL* — a good proxy for fine-grained topic[2] — within a relatively short period of time.[3] Some example pairs are shown in Table 1; we see that the paired tweets may differ dramatically, going far beyond word-for-word substitutions, so that quite interesting changes can be studied.

Looking at these examples, can one in fact tell from the wording which tweet in a topic- and author-controlled pair will be more successful? The answer may not be a priori clear. For example, for the first pair in the table, one person we asked found t_1's invocation of a "scandal" to be more attention-grabbing; but another person preferred t_2 because it is more informative about the URL's content and includes "fight media portrayal". In an Amazon Mechanical Turk (AMT) experiment (§4), we found that humans achieved an average accuracy of 61.3%: not that high, but better than chance, indicating that it is somewhat possible for humans to predict greater message spread from different deliveries of the same information.

Buoyed by the evidence of our AMT study that wording effects exist, we then performed a battery of experiments to seek generally-applicable, non-Twitter-specific features of more successful phrasings. §5.1 applies hypothesis testing (with Bonferroni correction to ameliorate issues with multiple comparisons) to investigate the utility of features like informativeness, resemblance to headlines, and conformity to the community norm in language use. §5.2 further validates our findings via prediction experiments, including on completely fresh held-out data, used only once and after an array of standard cross-validation experiments.[4] We achieved 66.5% cross-validation accuracy and 65.6% held-out accuracy with a combination of our custom features and bag-of-words. Our classifier fared significantly better than a number of baselines, including a strong classifier trained on the most- and least-retweeted tweets that was even granted access to author and timing metadata.

2 Related work

The idea of using carefully controlled experiments to study effective communication strategies dates back at least to Hovland et al. (1953). Recent studies range from examining what characteristics of *New York Times* articles correlate with high re-sharing rates (Milkman and Berger, 2012) to looking at how differences in description affect the spread of content-controlled videos or images (Borghol et al., 2012; Lakkaraju et al., 2013). Simmons et al. (2011) examined the variation of quotes from different sources to examine how textual memes mutate as people pass them along, but did not control for author. Predicting the "success" of various texts such as novels and movie quotes has been the aim of additional prior work not already mentioned in §1 (Ashok et al., 2013; Louis and Nenkova, 2013; Danescu-Niculescu-Mizil et al., 2012; Pitler and Nenkova, 2008; McIntyre and Lapata, 2009). To our knowledge, there have been no large-scale studies exploring wording effects in a both topic- and author-controlled setting. Employing such controls, we find that predicting the more effective alternative wording is much harder than the previously well-studied problem of pre-

[1] Cf. the Music Lab "multiple universes" experiment to test the randomness of popularity (Salganik et al., 2006).

[2] Although hashtags have been used as coarse-grained topic labels in prior work, for our purposes, we have no assurance that two tweets both using, say, "#Tahrir" would be attempting to express the same message but in different words. In contrast, see the same-URL examples in Table 1.

[3] Moreover, Twitter presents tweets to a reader in strict chronological order, so that there are no algorithmic-ranking effects to compensate for in determining whether readers saw a tweet. And, Twitter accumulates retweet counts for the entire retweet cascade and displays them for the original tweet at the root of the propagation tree, so we can directly use Twitter's retweet counts to compare the entire reach of the different versions.

[4] And after crossing our fingers.

dicting popular content when author or topic can freely vary.

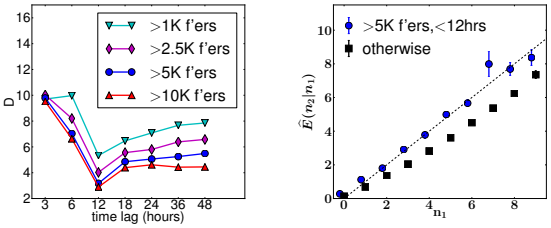
Related work regarding the features we considered is deferred to §5.1 (features description).

3 Data

Our main dataset was constructed by first gathering 1.77M topic- and author-controlled (henceforth *TAC*) tweet pairs[5] differing in more than just spacing.[6] We accomplished this by crawling timelines of 236K user ids that appear in prior work (Kwak et al., 2010; Yang and Leskovec, 2011) via the Twitter API. This crawling process also yielded 632K TAC pairs whose only difference was spacing, and an additional 558M "unpaired" tweets; as shown later in this paper, we used these extra corpora for computing language models and other auxiliary information. We applied non-obvious but important filtering — described later in this section — to control for other external factors and to reduce ambiguous cases. This brought us to a set of 11,404 pairs, with the *gold-standard* labels determined by which tweet in each pair was the one that received more retweets according to the Twitter API. We then did a second crawl to get an additional 1,770 pairs to serve as a held-out dataset. The corresponding tweet IDs are available online at `http://chenhaot.com/pages/wording-for-propagation.html`. (Twitter's terms of service prohibit sharing the actual tweets.)

Throughout, we refer to the textual content of the earlier tweet within a TAC pair as t_1, and of the later one as t_2. We denote the number of retweets received by each tweet by n_1 and n_2, respectively. We refer to the tweet with higher (lower) n_i as the "better (worse)" tweet.

Using "identical" pairs to determine how to compensate for follower-count and timing effects. In an ideal setting, differences between n_1 and n_2 would be determined solely by differences in wording. But even with a TAC pair, retweets might exhibit a temporal bias because of the chronological order of tweet presentation (t_1 might enjoy a first-mover advantage (Borghol et al., 2012) because it is the "original"; alternatively,



(a) For *identical* TAC pairs, retweet-count deviation vs. time lag between t_1 and t_2, for the author follower-counts given in the legend.

(b) Avg. n_2 vs. n_1 for identical TAC pairs, highlighting our chosen time-lag and follower thresholds. Bars: standard error. Diagonal line: $\hat{E}(n_2|n_1) = n_1$.

Figure 1: (a): The ideal case where $n_2 = n_1$ when $t_1 = t_2$ is best approximated when t_2 occurs within 12 hours of t_1 and the author has at least 10,000 or 5,000 followers. (b): in our chosen setting (blue circles), n_2 indeed tends to track n_1, whereas otherwise (black squares), there's a bias towards retweeting t_1.

t_2 might be preferred because retweeters consider t_1 to be "stale"). Also, the number of followers an author has can have complicated indirect effects on which tweets are read (space limits preclude discussion).

We use the 632K TAC pairs wherein t_1 and t_2 are *identical*[7] to check for such confounding effects: we see how much n_2 deviates from n_1 in such settings, since if wording were the only explanatory factor, the retweet rates for identical tweets ought to be equal. Figure 1(a) plots how the time lag between t_1 and t_2 and the author's follower-count affect the following deviation estimate:

$$D = \sum_{0 \leqslant n_1 < 10} |\hat{E}(n_2|n_1) - n_1|,$$

where $\hat{E}(n_2|n_1)$ is the average value of n_2 over pairs whose t_1 is retweeted n_1 times. (Note that the number of pairs whose t_1 is retweeted n_1 times decays exponentially with n_1; hence, we condition on n_1 to keep the estimate from being dominated by pairs with $n_1 = 0$, and do not consider $n_1 \geqslant 10$ because there are too few such pairs to estimate $\hat{E}(n_2|n_1)$ reliably.) Figure 1(a) shows that the setting where we (i) minimize the confounding effects of time lag and author's follower-count and (ii) maximize the amount of data to work with

[5] No data collection/processing was conducted at Google.

[6] The total excludes: tweets containing multiple URLs; tweets from users posting about the same URL more than five times (since such users might be spammers); the third, fourth, or fifth version for users posting between three and five tweets for the same URL; retweets (as identified by Twitter's API or by beginning with "RT @"); non-English tweets.

[7] Identical up to spacing: Twitter prevents exact copies by the same author appearing within a short amount of time, but some authors work around this by inserting spaces.

is: when t_2 occurs within 12 hours after t_1 and the author has more than 5,000 followers. Figure 1(b) confirms that for identical TAC pairs, our chosen setting indeed results in n_2 being on average close to n_1, which corresponds to the desired setting where wording is the dominant differentiating factor.[8]

Focus on meaningful and general changes. Even after follower-count and time-lapse filtering, we still want to focus on TAC pairs that (i) exhibit significant/interesting textual changes (as exemplified in Table 1, and as opposed to typo corrections and the like), and (ii) have n_2 and n_1 sufficiently different so that we are confident in which t_i is better at attracting retweets. To take care of (i), we discarded the 50% of pairs whose similarity was above the median, where similarity was tf-based cosine.[9] For (ii), we sorted the remaining pairs by $n_2 - n_1$ and retained only the top and bottom 5%.[10] Moreover, to ensure that we do not overfit to the idiosyncrasies of particular authors, we cap the number of pairs contributed by each author to 50 before we deal with (ii).

4 Human accuracy on TAC pairs

We first ran a pilot study on Amazon Mechanical Turk (AMT) to determine whether humans can identify, based on wording differences alone, which of two topic- and author- controlled tweets is spread more widely. Each of our 5 AMT tasks involved a disjoint set of 20 randomly-sampled TAC pairs (with t_1 and t_2 randomly reordered); subjects indicated "which tweet would other people be more likely to retweet?", provided a short justification for their binary response, and clicked a checkbox if they found that their choice was a "close call". We received 39 judgments per pair in aggregate from 106 subjects total (9 people completed all 5 tasks). The subjects' justifications were of very high quality, convincing us that they all did the task in good faith[11]. Two examples for the third TAC pair in Table 1 were: "[t_1 makes] the cause relate-able to some people, therefore showing more of an appeal as to why should they click the link and support" and, expressing the opposite view, "I like [t_2] more because [t_1] starts out with a generalization that doesn't affect me and try to make me look like I had that experience before".

If we view the set of 3900 binary judgments for our 100-TAC-pair sample as constituting independent responses, then the accuracy for this set is 62.4% (rising to 63.8% if we exclude the 587 judgments deemed "close calls"). However, if we evaluate the accuracy of the *majority* response among the 39 judgments per pair, the number rises to 73%. The accuracy of the majority response generally increases with the dominance of the majority, going above 90% when at least 80% of the judgments agree (although less than a third of the pairs satisfied this criterion).

Alternatively, we can consider the average accuracy of the 106 subjects: 61.3%, which is better than chance but far from 100%. (Variance was high: one subject achieved 85% accuracy out of 20 pairs, but eight scored below 50%.) This result is noticeably lower than the 73.8%-81.2% reported by Petrović et al. (2011), who ran a similar experiment involving two subjects and 202 tweet pairs, but where the pairs were *not* topic- or author-controlled.[12]

We conclude that even though propagation prediction becomes more challenging when topic and author controls are applied, humans can still to some degree tell which wording attracts more retweets. Interested readers can try this out themselves at `http://chenhaot.com/retweetedmore/quiz`.

5 Experiments

We now investigate computationally what wording features correspond to messages achieving a broader reach. We start (§5.1) by introducing a set of generally-applicable and (mostly) non-Twitter-specific features to capture our intuitions about what might be better ways to phrase a message. We then use hypothesis testing (§5.1) to evaluate the importance of each feature for message prop-

[8] We also computed the Pearson correlation between n_1 and n_2, even though it can be dominated by pairs with smaller n_1. The correlation is 0.853 for "$> 5K$ f'ers, <12hrs", clearly higher than the 0.305 correlation for "otherwise".

[9] Idf weighting was not employed because changes to frequent words are of potential interest. Urls, hashtags, @-mentions and numbers were normalized to [url], [hashtag], [at], and [num] before computing similarity.

[10] For our data, this meant $n_2 - n_1 \geqslant 10$ or $\leqslant -15$. Cf. our median number of retweets: 30.

[11] We also note that the feedback we got was quite positive, including: "...It's fun to make choices between close tweets and use our subjective opinion. Thanks and best of luck with your research" and "This was very interesting and really made me think about how I word my own tweets. Great job on this survey!". We only had to exclude one person (not counted among the 106 subjects), doing so because he or she gave the same uninformative justification for all pairs.

[12] The accuracy range stems from whether author's social features were supplied and which subject was considered.

Table 2: Notational conventions for tables in §5.1.

One-sided paired t-test for feature efficacy	
↑↑↑↑: p<1e-20	↓↓↓↓: p>1-1e-20
↑↑↑ : p<0.001	↓↓↓ : p>0.999
↑↑ : p<0.01	↓↓ : p>0.99
↑ : p<0.05	↓ : p>0.95
∗: passes our Bonferroni correction	

One-sided binomial test for feature increase (Do authors prefer to 'raise' the feature in t_2?)

YES : t_2 has a higher feature score than t_1, $\alpha = .05$
NO : t_2 has a lower feature score than t_1, $\alpha = .05$
(x%): $\%(f_2 > f_1)$, if sig. larger or smaller than 50%

agation and the extent to which authors employ it, followed by experiments on a prediction task (§5.2) to further examine the utility of these features.

5.1 Features: efficacy and author preference

What kind of phrasing helps message propagation? Does it work to explicitly ask people to share the message? Is it better to be short and concise or long and informative? We define an array of features to capture these and other messaging aspects. We then examine (i) how effective each feature is for attracting more retweets; and (ii) whether authors prefer applying a given feature when issuing a second version of a tweet.

First, for each feature, we use a one-sided paired t-test to test whether, on our 11K TAC pairs, our score function for that feature is larger in the better tweet versions than in the worse tweet versions, for significance levels $\alpha = .05, .01, .001,$ 1e-20. Given that we did 39 tests in total, there is a risk of obtaining false positives due to multiple testing (Dunn, 1961; Benjamini and Hochberg, 1995). To account for this, we also report significance results for the conservatively Bonferroni-corrected ("BC") significance level α = 0.05/39=1.28e-3.

Second, we examine author preference for applying a feature. We do so because one (but by no means the only) reason authors post t_2 after having already advertised the same URL in t_1 is that these authors were dissatisfied with the amount of attention t_1 got; in such cases, the changes may have been specifically intended to attract more retweets. We measure author preference for a feature by the percentage of our TAC pairs[13] where t_2 has more "occurrences" of the feature than t_1, which we denote by "$\%(f_2 > f_1)$". We use the one-sided binomial test to see whether $\%(f_2 > f_1)$ is significantly larger (or smaller) than 50%.

[13] For our preference experiments, we added in pairs where $n_2 - n_1$ was not in the top or bottom 5% (cf. §3, meaningful changes), since to measure author preference it's not necessary that the retweet counts differ significantly.

Table 3: Explicit requests for sharing (where only occurrences POS-tagged as verbs count, according to the Gimpel et al. (2011) tagger).

	effective?	author-preferred?
rt	↑↑↑↑ *	——
retweet	↑↑↑↑ *	YES (59%)
spread	↑↑↑ *	YES (56%)
please	↑↑↑ *	——
pls	↑	——
plz	↑↑	——

Table 4: Informativeness.

	effective?	author-preferred?
length (chars)	↑↑↑↑ *	YES (54%)
verb	↑↑↑↑ *	YES (56%)
noun	↑↑↑↑ *	——
adjective	↑↑↑ *	YES (51%)
adverb	↑↑↑ *	YES (55%)
proper noun	↑↑↑ *	NO (45%)
number	↑↑↑↑ *	NO (48%)
hashtag	↑	——
@-mention	↓↓↓ *	YES (53%)

Not surprisingly, it helps to ask people to share. (See Table 3; the notation for all tables is explained in Table 2.) The basic sanity check we performed here was to take as features the number of occurrences of the verbs 'rt', 'retweet', 'please', 'spread', 'pls', and 'plz' to capture explicit requests (e.g. "please retweet").

Informativeness helps. (Table 4) Messages that are more informative have increased *social exchange value* (Homans, 1958), and so may be more worth propagating. One crude approximation of informativeness is length, and we see that length helps.[14] In contrast, Simmons et al. (2011) found that shorter versions of memes are more likely to be popular. The difference may result from TAC-pair changes being more drastic than the variations that memes undergo.

A more refined informativeness measure is counts of the parts of speech that correspond to content. Our POS results, gathered using a Twitter-specific tagger (Gimpel et al., 2011), echo those of Ashok et al. (2013) who looked at predict-

[14] Of course, simply inserting garbage isn't going to lead to more retweets, but adding more information generally involves longer text.

Table 5: Conformity to the community and one's own past, measured via scores assigned by various language models.

	effective?	author-preferred?
twitter unigram	↑↑↑ *	YES (54%)
twitter bigram	↑↑↑ *	YES (52%)
personal unigram	↑↑↑ *	YES (52%)
personal bigram	———	NO (48%)

ing the success of books. The diminished effect of hashtag inclusion with respect to what has been reported previously (Suh et al., 2010; Petrović et al., 2011) presumably stems from our topic and author controls.

Be like the community, and be true to yourself (in the words you pick, but not necessarily in how you combine them). (Table 5) Although distinctive messages may attract attention, messages that conform to expectations might be more easily accepted and therefore shared. Prior work has explored this tension: Lakkaraju et al. (2013), in a content-controlled study, found that the more upvoted Reddit image titles balance novelty and familiarity; Danescu-Niculescu-Mizil et al. (2012) (henceforth DCKL'12) showed that the memorability of movie quotes corresponds to higher lexical distinctiveness but lower POS distinctiveness; and Sun et al. (2013) observed that deviating from one's own past language patterns correlates with more retweets.

Keeping in mind that the authors in our data have at least 5000 followers[15], we consider two types of language-conformity constraints an author might try to satisfy: to be similar to what is normal in the Twitter community, and to be similar to what his or her followers expect. We measure a tweet's similarity to expectations by its score according to the relevant language model, $\frac{1}{|T|}\sum_{x\in T}\log(p(x))$, where T refers to either all the unigrams (unigram model) or all and only bigrams (bigram model).[16] We trained a Twitter-community language model from our 558M unpaired tweets, and personal language models from each author's tweet history.

Imitate headlines. (Table 6) News headlines are often intentionally written to be both informative and attention-getting, so we introduce the idea of scoring by a language model built from New York Times headlines.[17]

Table 6: LM-based resemblance to headlines.

	effective?	author-preferred?
headline unigram	↑↑	YES (53%)
headline bigram	↑↑↑↑ *	YES (52%)

Table 7: Retweet score.

	effective?	author-preferred?
rt score	↑↑ *	NO (49%)
verb rt score	↑↑↑↑ *	——
noun rt score	↑↑↑ *	——
adjective rt score	↑	YES (50%)
adverb rt score	↑	YES (51%)
proper noun rt score	———	NO (48%)

Use words associated with (non-paired) retweeted tweets. (Table 7) We expect that provocative or sensationalistic tweets are likely to make people react. We found it difficult to model provocativeness directly. As a rough approximation, we check whether the changes in t_2 with respect to t_1 (which share the same topic and author) involve words or parts-of-speech that are associated with high retweet rate in a very large separate sample of unpaired tweets (retweets and replies discarded). Specifically, for each word w that appears more than 10 times, we compute the probability that tweets containing w are retweeted more than once, denoted by $rs(w)$. We define the *rt score* of a tweet as $max_{w\in T}rs(w)$, where T is all the words in the tweet, and the *rt score* of a particular POS tag z in a tweet as $max_{w\in T\&\mathrm{tag}(w)=z}rs(w)$.

Include positive and/or negative words. (Table 8) Prior work has found that including positive or negative sentiment increases message propagation (Milkman and Berger, 2012; Godes et al., 2005; Heath et al., 2001; Hansen et al., 2011). We measured the occurrence of positive and negative words as determined by the connotation lexicon of Feng et al. (2013) (better coverage than LIWC). Measuring the occurrence of both *simultaneously* was inspired by Riloff et al. (2013).

Refer to other people (but not your audience). (Table 9) First-person has been found useful for success before, but in the different domains of scientific abstracts (Guerini et al., 2012) and books (Ashok et al., 2013).

[15] This is not an artificial restriction on our set of authors; a large follower count means (in principle) that our results draw on a large sample of decisions whether to retweet or not.

[16] The tokens [at], [hashtag], [url] were ignored in the unigram-model case to prevent their undue influence, but retained in the bigram model to capture longer-range usage ("combination") patterns.

[17] To test whether the results stem from similarity to *news* rather than headlines per se, we constructed a NYT-text LM, which proved less effective. We also tried using Gawker headlines (often said to be attention-getting) but pilot studies revealed insufficient vocabulary overlap with our TAC pairs.

Table 8: Sentiment (contrast is measured by presence of both positive and negative sentiments).

	effective?	author-preferred?
positive	↑↑↑ *	——
negative	↑↑↑ *	——
contrast	↑↑↑ *	——

Table 9: Pronouns.

	effective?	author-preferred?
1st person singular	——	YES (51%)
1st person plural	——	YES (52%)
2nd person	——	YES (57%)
3rd person singular	↑↑	YES (55%)
3rd person plural	↑	YES (58%)

Generality helps. (Table 10) DCKL'12 posited that movie quotes are more shared in the culture when they are general enough to be used in multiple contexts. We hence measured the presence of indefinite articles vs. definite articles.

The easier to read, the better. (Table 11) We measure readability by using Flesch reading ease (Flesch, 1948) and Flesch-Kincaid grade level (Kincaid et al., 1975), though they are not designed for short texts. We use negative grade level so that a larger value indicates easier texts to read.

Final question: Do authors prefer to do what is effective? Recall that we use binomial tests to determine author preference for applying a feature more in t_2. Our preference statistics show that author preferences in many cases are aligned with feature efficacy. But there are several notable exceptions: for example, authors tend to increase the use of @-mentions and 2nd person pronouns even though they are ineffective. On the other hand, they did not increase the use of effective ones like proper nouns and numbers; nor did they tend to increase their rate of sentiment-bearing words. Bearing in mind that changes in t_2 may not always be intended as an effort to improve t_1, it is still interesting to observe that there are some contrasts between feature efficacy and author preferences.

5.2 Predicting the "better" wording

Here, we further examine the collective efficacy of the features introduced in §5.1 via their performance on a binary prediction task: given a TAC pair (t_1, t_2), did t_2 receive more retweets?

Our approach. We group the features introduced in §5.1 into 16 lexicon-based features (Table 3, 8, 9, 10), 9 informativeness features (Table 4), 6 language model features (Table 5, 6), 6 rt score features (Table 7), and 2 readability features (Table 11). We refer to all 39 of them together as *custom* features. We also consider tagged bag-of-words ("BOW") features, which includes all the unigram (word:POS pair) and bigram features that appear more than 10 times in the cross-validation data. This yields 3,568 unigram features and 4,095 bigram features, for a total of 7,663 so-called *1,2-gram features*. Values for each feature are normalized by linear transformation across all tweets in the training data to lie in the range $[0, 1]$.[18]

Table 10: Generality.

	effective?	author-preferred?
indefinite articles (a,an)	↑↑↑ *	——
definite articles (the)	——	YES (52%)

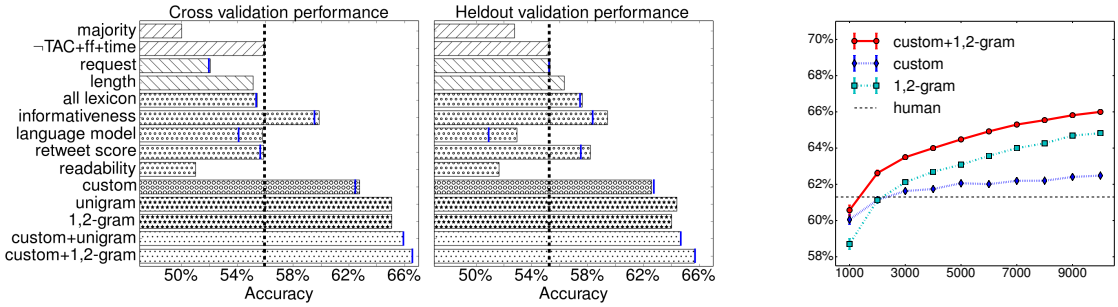
Table 11: Readability.

	effective?	author-preferred?
reading ease	↑↑	YES (52%)
negative grade level	↑	YES (52%)

For a given TAC pair, we construct its feature vector as follows. For each feature being considered, we compute its normalized value for each tweet in the pair and take the difference as the feature value for this pair. We use L2-regularized logistic regression as our classifier, with parameters chosen by cross validation on the training data. (We also experimented with SVMs. The performance was very close, but mostly slightly lower.)

A strong non-TAC alternative, with social information and timing thrown in. One baseline result we would like to establish is whether the topic and author controls we have argued for, while intuitively compelling for the purposes of trying to determine the best way for a given author to present some fixed content, are really necessary in practice. To test this, we consider an alternative binary L2-regularized logistic-regression classifier that is trained on unpaired data, specifically, on the collection of 10,000 most retweeted tweets (gold-standard label: positive) plus the 10,000 least retweeted tweets (gold-standard label: negative) that are neither retweets nor replies. Note that this alternative thus is granted, by design, roughly *twice* the training instances that our classifiers have, as a result of having roughly the same number of tweets, since our instances are pairs. Moreover, we additionally include the tweet author's follower count, and the day and hour of posting, as features. We refer to this alternative classifier as $\neg$*TAC+ff+time*. (Mnemonic: "ff" is used in bibliographic contexts as an abbreviation

[18] We also tried normalization by *whitening*, but it did not lead to further improvements.



(a) Cross-validation and heldout accuracy for various feature sets. Blue lines inside bars: performance when custom features are restricted to those that pass our Bonferroni correction (no line for readability because no readability features passed). Dashed vertical line: ¬TAC+ff+time performance.

(b) Cross-validation accuracy vs data size. Human performance was estimated from a disjoint set of 100 pairs (see §4).

Figure 2: Accuracy results. Pertinent significance results are as follows. In cross-validation, custom+1,2-gram is significantly better than ¬TAC+ff+time (p=0) and 1,2-gram (p=3.8e-7). In heldout validation, custom+1,2-gram is significantly better than ¬TAC+ff+time (p=3.4e-12) and 1,2-gram (p=0.01) but not unigram (p=0.08), perhaps due to the small size of the heldout set.

for "and the following".) We apply it to a tweet pair by computing whether it gives a higher score to t_2 or not.

Baselines. To sanity-check whether our classifier provides any improvement over the simplest methods one could try, we also report the performance of the majority baseline, our request-for-sharing features, and our character-length feature.

Performance comparison. We compare the accuracy (percentage of pairs whose labels were correctly predicted) of our approach against the competing methods. We report 5-fold cross validation results on our balanced set of 11,404 TAC pairs and on our completely disjoint heldout data[19] of 1,770 TAC pairs; this set was never examined during development, and there are no authors in common between the two testing sets.

Figure 2(a) summarizes the main results. While ¬TAC+ff+time outperforms the majority baseline, using all the features we proposed beats ¬TAC+ff+time by more than 10% in both cross-validation (66.5% vs 55.9%) and heldout validation (65.6% vs 55.3%). We outperform the average human accuracy of 61% reported in our Amazon Mechanical Turk experiments (for a different data sample); ¬TAC+ff+time fails to do so.

The importance of topic and author control can be seen by further investigation of ¬TAC+ff+time's performance. First, note that it yields an accuracy of around 55% on our alternate-version-selection task,[20] even though its cross-validation accuracy on the larger most- and least-retweeted unpaired tweets averages out to a high 98.8%. Furthermore, note the superior performance of unigrams trained on TAC data vs ¬TAC+ff+time — which is similar to our unigrams but trained on a larger but non-TAC dataset that included metadata. Thus, TAC pairs are a useful data source even for non-custom features. (We also include individual feature comparisons later.)

Informativeness is the best-performing custom feature group when run in isolation, and outperforms all baselines, as well as ¬TAC+ff+time; and we can see from Figure 2(a) that this is not due just to length. The combination of all our 39 custom features yields approximately 63% accuracy in both testing settings, significantly outperforming informativeness alone ($p<0.001$ in both cases). Again, this is higher than our estimate of average human performance.

Not surprisingly, the TAC-trained BOW features (unigram and 1,2-gram) show impressive predictive power in this task: many of our custom features can be captured by bag-of-word features, in a way. Still, the best performance is achieved

[19] To construct this data, we used the same criteria as in §3: written by authors with more than 5000 followers, posted within 12 hours, $n_2 - n_1 \geqslant 10$ or $\leqslant -15$, and cosine similarity threshold value the same as in §3, cap of 50 on number of pairs from any individual author.

[20] One might suspect that the problem is that ¬TAC+ff+time learns from its training data to overrely on follower-count, since that is presumably a good feature for non-TAC tweets, and for this reason suffers when run on TAC data where follower-counts are by construction non-informative. But in fact, we found that removing the follower-count feature from ¬TAC+ff+time and re-training did not lead to improved performance. Hence, it seems that it is the non-controlled nature of the alternate training data that explains the drop in performance.

by combining our custom and 1,2-gram features together, to a degree statistically significantly better than using 1,2-gram features alone.

Finally, we remark on our Bonferroni correction. Recall that the intent of applying it is to avoid false positives. However, in our case, Figure 2(a) shows that our potentially "false" positives — features whose effectiveness did not pass the Bonferroni correction test — actually do raise performance in our prediction tests.

Size of training data. Another interesting observation is how performance varies with data size. For $n = 1000, 2000, \ldots, 10000$, we randomly sampled n pairs from our 11,404 pairs, and computed the average cross-validation accuracy on the sampled data. Figure 2(b) shows the averages over 50 runs of the aforementioned procedure. Our custom features can achieve good performance with little data, in the sense that for sample size 1000, they outperform BOW features; on the other hand, BOW features quickly surpass them. Across the board, the custom+1,2-gram features are consistently better than the 1,2-gram features alone.

Top features. Finally, we examine some of the top-weighted individual features from our approach and from the competing ¬TAC+ff+time classifier. The top three rows of Table 12 show the best custom and best and worst unigram features for our method; the bottom two rows show the best and worst unigrams for ¬TAC+ff+time. Among custom features, we see that community and personal language models, informativeness, retweet scores, sentiment, and generality are represented. As for unigram features, not surprisingly, "rt" and "retweet" are top features for both our approach and ¬TAC+ff+time. However, the other unigrams for the two methods seem to be a bit different in spirit. Some of the unigrams determined to be most poor only by our method appear to be both surprising and yet plausible in retrospect: "icymi" (abbreviation for "in case you missed it") tends to indicate a direct repetition of older information, so people might prefer to retweet the earlier version; "thanks" and "sorry" could correspond to personal thank-yous and apologies not meant to be shared with a broader audience, and similarly @-mentioning another user may indicate a tweet intended only for that person. The appearance of [hashtag] in the best ¬TAC+ff+time unigrams is consistent with prior research in non-TAC settings (Suh et al., 2010; Petrović et al., 2011).

Table 12: Features with largest coefficients, delimited by commas. POS tags omitted for clarity.

Our approach	
best 15 custom	twitter bigram, length (chars), rt (the word), retweet (the word), verb, verb retweet score, personal unigram, proper noun, number, noun, positive words, please (the word), proper noun retweet score, indefinite articles (a,an), adjective
best 20 unigrams	rt, retweet, [num], breaking, is, win, never, ., people, need, official, officially, are, please, november, world, girl, !!!, god, new
worst 20 unigrams	:, [at], icymi, also, comments, half, ?, earlier, thanks, sorry, highlights, bit, point, update, last, helping, peek, what, haven't, debate
¬TAC+ff+time	
best 20 unigrams	[hashtag], teen, fans, retweet, sale, usa, women, butt, caught, visit, background, upcoming, rt, this, bieber, these, each, chat, houston, book
worst 20 unigrams	:, ..., boss, foundation, ?, ∼, others, john, roll, ride, appreciate, page, drive, correct, full, ', looks, @ (not as [at]), sales, hurts

6 Conclusion

In this work, we conducted the first large-scale topic- and author-controlled experiment to study the effects of wording on information propagation.

The features we developed to choose the better of two alternative wordings posted better performance than that of all our comparison algorithms, including one given access to author and timing features but trained on non-TAC data, and also bested our estimate of average human performance. According to our hypothesis tests, helpful wording heuristics include adding more information, making one's language align with both community norms and with one's prior messages, and mimicking news headlines. Readers may try out their own alternate phrasings at `http://chenhaot.com/retweetedmore/` to see what a simplified version of our classifier predicts.

In future work, it will be interesting to examine how these features generalize to longer and more extensive arguments. Moreover, understanding the underlying psychological and cultural mechanisms that establish the effectiveness of these features is a fundamental problem of interest.

Acknowledgments. We thank C. Callison-Burch, C. Danescu-Niculescu-Mizil, J. Kleinberg, P. Mahdabi, S. Mullainathan, F. Pereira, K. Raman, A. Swaminathan, the Cornell NLP seminar participants and the reviewers for their comments; J. Leskovec for providing some initial data; and the anonymous annotators for all their labeling help. This work was supported in part by NSF grant IIS-0910664 and a Google Research Grant.

References

Yoav Artzi, Patrick Pantel, and Michael Gamon. 2012. Predicting responses to microblog posts. In *Proceedings of NAACL (short paper)*.

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of EMNLP*.

Eitan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of WSDM*.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

Youmna Borghol, Sebastien Ardon, Niklas Carlsson, Derek Eager, and Anirban Mahanti. 2012. The untold story of the clones: Content-agnostic factors that impact YouTube video popularity. In *Proceedings of KDD*.

Dennis Chong and James N. Druckman. 2007. Framing theory. *Annual Review of Political Science*, 10:103–126.

Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. 2012. You had me at hello: How phrasing affects memorability. In *Proceedings of ACL*.

John DiNardo. 2008. Natural experiments and quasi-natural experiments. In *The New Palgrave Dictionary of Economics*. Palgrave Macmillan.

Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. 2013. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of ACL*.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of NAACL (short paper)*.

David Godes, Dina Mayzlin, Yubo Chen, Sanjiv Das, Chrysanthos Dellarocas, Bruce Pfeiffer, Barak Libai, Subrata Sen, Mengze Shi, and Peeter Verlegh. 2005. The firm's management of social interactions. *Marketing Letters*, 16(3-4):415–428.

Marco Guerini, Carlo Strapparava, and Gözde Özbal. 2011. Exploring text virality in social networks. In *Proceedings of ICWSM (poster)*.

Marco Guerini, Alberto Pepe, and Bruno Lepri. 2012. Do linguistic style and readability of scientific abstracts affect their virality? In *Proceedings of ICWSM (poster)*.

Lars Kai Hansen, Adam Arvidsson, Finn Årup Nielsen, Elanor Colleoni, and Michael Etter. 2011. Good friends, bad news-affect and virality in Twitter. *Communications in Computer and Information Science*, 185:34–43.

Chip Heath, Chris Bell, and Emily Sternberg. 2001. Emotional selection in memes: The case of urban legends. *Journal of personality and social psychology*, 81(6):1028.

George C. Homans. 1958. Social Behavior as Exchange. *American Journal of Sociology*, 63(6):597–606.

Liangjie Hong, Ovidiu Dan, and Brian D. Davison. 2011. Predicting popular messages in Twitter. In *Proceedings of WWW*.

Carl I. Hovland, Irving L. Janis, and Harold H. Kelley. 1953. *Communication and Persuasion: Psychological Studies of Opinion Change*, volume 19. Yale University Press.

J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media? In *Proceedings of WWW*.

Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. 2013. What's in a name? Understanding the interplay between titles, content, and communities in social media. In *Proceedings of ICWSM*.

Annie Louis and Ani Nenkova. 2013. What makes writing great? First experiments on article quality prediction in the science journalism domain. *Transactions of ACL*.

Zongyang Ma, Aixin Sun, and Gao Cong. 2012. Will this #hashtag be popular tomorrow? In *Proceedings of SIGIR*.

Neil McIntyre and Mirella Lapata. 2009. Learning to tell tales: A data-driven approach to story generation. In *Proceedings of ACL-IJCNLP*.

Katherine L Milkman and Jonah Berger. 2012. What makes online content viral? *Journal of Marketing Research*, 49(2):192–205.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2011. RT to win! Predicting message propagation in Twitter. In *Proceedings of ICWSM*.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of EMNLP*.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of EMNLP*.

Daniel M. Romero, Chenhao Tan, and Johan Ugander. 2013. On the interplay between social and topical structure. In *Proceedings of ICWSM*.

Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856.

Matthew P. Simmons, Lada A Adamic, and Eytan Adar. 2011. Memes online: Extracted, subtracted, injected, and recollected. In *Proceedings of ICWSM*.

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 SocialCom*.

Tao Sun, Ming Zhang, and Qiaozhu Mei. 2013. Unexpected relevance: An empirical study of serendipity in retweets. In *Proceedings of ICWSM*.

Oren Tsur and Ari Rappoport. 2012. What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities. In *Proceedings of WSDM*.

Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In *Proceedings of WSDM*.