Sonora: A Prescriptive Model for Message Authoring on Twitter

Pablo N. Mendes, Daniel Gruhl, Clemens Drews, Chris Kau, Neal Lewis, Meena Nagarajan, Alfredo Alba, and Steve Welch

IBM Research, USA

Abstract. Within social networks, certain messages propagate with more ease or attract more attention than others. This effect can be a consequence of several factors, such as topic of the message, number of followers, real-time relevance, person who is sending the message etc. Only one of these factors is within a user's reach at authoring time: how to phrase the message. In this paper we examine how word choice contributes to the propagation of a message.

We present a prescriptive model that analyzes words based on their historic performance in retweets in order to propose enhancements in future tweet performance. Our model calculates a novel score (SONORA SCORE) that is built on three aspects of diffusion - volume, prevalence and sustain.

We show that SONORA SCORE has powerful predictive ability, and that it complements social and tweet-level features to achieve an F1 score of 0.82 in retweet prediction. Moreover, it has the ability to prescribe changes to the tweet wording such that when the SONORA SCORE for a tweet is higher, it is twice as likely to have more retweets.

Lastly, we show how our prescriptive model can be used to assist users in content creation for optimized success on social media. Because the model works at the word level, it lends itself extremely well to the creation of user interfaces which help authors incrementally – word by word – refine their message until its potential is maximized and it is ready for publication. We present an easy to use iOS application that illustrates the potential of incremental refinement using SONORA SCORE coupled with the familiarity of a traditional spell checker.

Introduction

It is estimated that 72% of online adults use social media sites [11]. This percentage is even higher within the subgroup of young adults. Perhaps more surprisingly, the presence of senior citizens has roughly tripled in recent years. As the usage of social networking websites become routine for adults of all ages, these platforms represent an ever increasing opportunity for content sharing for virtually any content-producing professional or institution.

Authoring popular content for social media is challenging, especially considering the many variables that contribute to the "uptick" of a message [8,19,21]. Nagarajan et al. [17] show that presence or absence of attribution can largely dictate how retweetability is observed in a diffusion network. Hansen et al. [6] show that negative sentiment enhances virality in the news segment, but not in the non-news segment. Of these many

B. Benatallah et al. (Eds.): WISE 2014, Part II, LNCS 8787, pp. 17–31, 2014.

attributes that enhance a message's tendency to propagate, word choice is the only one controllable at the time of writing.

Given parlance variation among different demographics and communities, it is intuitive that word choice impacts an audience's reception of content. This variation is quite pronounced in the "New Media" age [12]. For instance, the word choice of middle age professionals discussing their product goals most certainly differs from teenage students discussing their music interests. This highlights a fundamental reason why word choice is important: by speaking the wrong vernacular one can not only distort the core of a message but also its reach.

In this paper we primarily concentrate upon the impact of word or phrase selection on the propagation of a message throughout its intended audience. To this end, we have developed a measure we call SONORA SCORE to prescribe word changes for an uptick in retweetability. SONORA SCORE estimates how well the language used in a tweet has performed, based on the observation of past data. The intuition behind SONORA SCORE is that certain words may 'resonate' better within a community. Based on this sound-related metaphor, we introduce measures of how 'loud' a word sounds, how prevalent it sounds within a time period, and for how long it sounds.

Authors, if well instructed, can modify their word choices to increase their retweetability. Although it would be beneficial to have a 'spell checker'-equivalent system to help improve a message for highest impact, no such solution has yet been disclosed in the literature. SONORA SCORE can be used in realtime to assist authors by prescribing word changes. In this paper, we highlight the impact of SONORA SCORE on 'words' in a tweet. However the proposed SONORA SCORE is easily extended to n-grams of any length and to other kinds of data such as email, blogs, news articles etc. A mobile application prototype using the SONORA SCORE is shown in Figure 1. The figure illustrates the process of scoring words and helping the user identify those words that have potential for improvement. If the user wishes to change a word the prototype proposes alternative words which our model sees as more effective.

In the remainder of this paper, we start by discussing related research in this field, and providing a high level look at how SONORA SCORE works. Our datasets and experiments are described next, exploring the predictive features of retweetability and the prescriptive function of SONORA SCORE. We also include an application section to illustrate the usage of our score in practice. Lastly, we present our conclusions and propose future work.

Related Work

Most of the analysis of message success in the Twitter-sphere has focused on asking questions of "global" features that might help us understand this phenomenon [21] – do tweets with URLs tend to get shared more? Do past retweets of an author boost future retweetability? What role does the topic of the tweet play? etc. Our focus is somewhat different. The system presented in this work aims to identify **wording-based features** that can be used to prescribe more successful features at message authoring time.

Recently, several studies have used measures of retweets, replies, and likes on Twitter as proxies to measure message virality. Suh et al. [21] found that the age of a user's account on a medium, number of friends they had, and the presence of URLs and hashtags

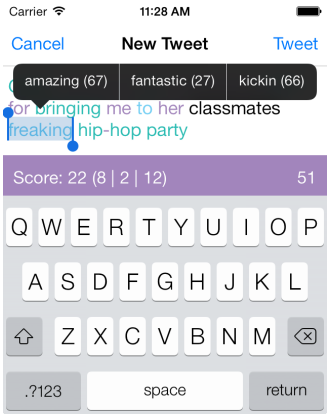


Fig. 1. Screenshot of prescriptive system for iterative content refinement. A potential tweet is entered, and the system dynamically computes the SONORA SCORE of each word. Words below a threshold are colored to suggest that better options are available. Selecting a word reveals alternatives with their respective SONORA SCORE. The user can choose from those suggestions or enter an alternative.

in the content had strong relationships with retweetability. This finding is supported by Petrović and Osborne [19], who concluded that the number of followers and the number of times a user was added to 'lists', along with content features were strong predictors of message retweetability. Yang et al. [23] found that the author's rate of being mentioned by other people predicts how fast a tweet spreads; including links in tweets often generates more number of tweets; and greater number of posts and mentions for a user are better predictors of longer diffusion hops.

Bigonha et al. [3] found that the readability quality of a message was useful (along with network features) in identifying the most influential members of a network. This stream of work has also been applied outside Twitter. Utilizing linguistic categories, [5] show that the presence of certain stylistic features (such as using assertive words and fewer tentative words) and better readability of abstracts correlated positively with the viral capability of a scientific article. Virality was defined as number of article downloads, citations, and bookmarks received by an article. These studies essentially support the motivation behind our work – content features have undeniable effects on message diffusion and author perception.

Our work is motivated in part by psycholinguistic analysis and readability tests that explore the effect of word and language choice [2], [4]. It addresses limitations of the above conclusions in their lack of prescribing what an author can do to write a more effective message. While we acknowledge that the multitude of features on a medium are central to understanding message "uptake", the goal of our work in contrast to that of our predecessors is not to achieve a global optimization of the problem (how to achieve the most retweets for any tweet) but rather a local optimization problem (what word choices will increase the likelihood of tweet retweetability).

Lakkaraju, McAuley and Leskovec [14] study resubmissions of images on the social network Reddit, and find that wording features can be predictive of whether a title will be a successful resubmission. It is unclear from their evaluation if their language model retains its success without features resubmission-specific features – e.g. words that have been previously successful with a given image. A contemporary paper to ours by Tan, Lee and Pang [22] (their results were not published when we conducted our research) has also confirmed that words have strong predictive ability. They report, for example, that words such as 'rt', 'retweet', 'win', and 'official' are strong predictors of retweetability. However, their discussion concentrates on globally successful words – with the danger of including words that are prevalent in spam messages (e.g. 'win'). They do not discuss prescribing word substitutions, which is the main interest in our work.

Approach

The algorithm consists of two disjoint steps: firstly selecting a set of historical tweets and extracting a few pieces of metadata. Secondly, combining metadata into a single score. In pseudo-code, the first part of the algorithm works as follows: (1) Select the word you wish to evaluate. Say, "popcorn".[1](2) From the corpus of tweets defining your target audience, select all tweets that mention the word "popcorn" (note: while we are looking at a single token here, we support n-grams such as "buttered popcorn") (3) Examine for each of these tweets the following (a) What is its root tweet[2]; (b) How often was that root tweet retweeted; (c) How long did the retweeting go on for; (d) When did the root tweet occur. Computing the SONORA SCORE can be defined as an operation on this set of posts and metadata. In the remainder of this section we will formally define these calculations.

Definitions

While it is possible to focus on different outcomes the most generic form of SONORA SCORE is calculated as a combination of three sub scores that can be understood in an analogy to sounds – i.e. given a word, how well will it sound in the social network – in terms of: VOLUME, PREVALENCE, SUSTAIN.

Posts. Let D be a subset of posts of size $n = |D|$, selected via the mechanism above from a universe (e.g. tweets) as a representative sample of a community c. Hereafter, we will use $D(w)$ as the shorthand notation for the subset of all posts in D that contain the word w.

Retweets. Let $RT(t)$ be the observed *retweetability* of a tweet t where $RT(t)$ is the number of times a tweet has been forwarded by a user through the retweet[3] function on

[1] While we use the example of a word here, any attribute of a tweet can define a set. e.g., author, class of influencers, time of tweet, geographic location of tweet, etc.

[2] We use 'root tweet' to refer to the tweet that was the initial source of all retweets.

[3] `https://support.twitter.com/articles/77606-faqs-about-retweets-rt`

Twitter™. Let NZRT (non-zero retweets) be the set of retweeted posts, and ZRT (zero retweets) be the set of non-retweeted posts. Similarly, NZRT(w) and ZRT(w) are the corresponding subsets of posts containing the word w.

Word Volume. VOLUME of a word w captures the intuition of how 'loud' w 'resonates' in the subset of posts. It is represented by the sum of the retweet counts of all posts in $D(w)$ that have a non-zero retweet count: $V(w) = \sum_{\forall t_i \in \text{NZRT}(w)} RT(t_i)$.

Word Amplitude. AMPLITUDE of a word w is a variant of Volume that models the difference in volume for a word w in a subset of posts that were retweeted versus a subset of posts that were not retweeted: $A(w) = \phi(|\text{NZRT}(w)|) - \phi(|\text{ZRT}(w))|$. Here we use $\phi(x \in X) = x - mean(X)$ as a mean-centering transformation function to center the counts around 0.

Word Prevalence. PREVALENCE captures the notion that some words are more common than others over a timespan of interest. It is computed based on the number of elements in $D(w)$ that occur for each day in the timespan $\{\tau-, \tau+\}$ of interest. $P(w) = \sum_{d \in \{\tau-,\tau+\}} DC(w, d)$, and the daily count $DC(w, d)$ is the cardinality of the set of posts t_i in day d that contain the word w: $DC(w, d) = |\{\forall t_i \in D(w) | date(t_i) \in d\}|$.

Word Sustain. SUSTAIN of a word captures the notion of how long a word 'resonates' in a subset. Let us define the r_s score to be the sum of the number of hours from first to last tweet of all t_i where this number is non zero: $S(w) = \sum_{\{\forall t_i \in D(w) | \text{HR}(t_i) > 0\}} \text{HR}(t_i)$. The number of hours (HR) is calculated from the first time a tweet id appeared in the dataset, until the last time it appeared.

Tweet Scores from Word-Level Features. So far, we have defined the word-level scores $V(w)$, $P(w)$ and $S(w)$ as functions over words, i.e. functions of the type $g : w \to \mathbb{R}$. It is also possible to define a corresponding score $V_f(t)$ as an aggregation of word scores $V(w_i)$ for a tweet $t = \{w_1, ..., w_n\}$ according to an aggregation function $f : \{x_1, ..., x_n\} \in \mathbb{R} \to \mathbb{R}$. Corresponding definitions of aggregation functions apply to $P_f(t)$ and $S_f(t)$. Possible choices for f are mean, min, max and stdev.

We can now define $Sonora(t, c)$ for a tweet t and community c as an estimate of how well the tweet's words $t = \{w_1, ..., w_n\}$ resonate with the community c. The SONORA SCORE of a tweet is computed through an aggregation of the $V_f(t)$, $P(t)$ and $S(t)$ scores based on its words: $Sonora(t) = \alpha \cdot V_f(t) + \beta \cdot P_f(t) + \gamma \cdot S_f(t)$. Here α, β, γ are mixture weights to control the influence of each component on the final SONORA SCORE, according to the desired outcome ($\alpha + \beta + \gamma = 1$).

Evaluation

For experiments in this work, we collected roughly 225M tweets from an unfiltered 1% feed from Twitter over a three month period, with a few days missing due to network connectivity challenges. To test the predictive nature of SONORA SCORE and other features, the the sample was split in training/testing sets. We chose an arbitrary point τ in time (Apr 01 06:59:59), and divided the data into two sets: $D\tau+$ and $D\tau-$with all tweets before and after τ, respectively. A total of 3.8% of the tweets in $D\tau+$ has either been tweeted or retweeted in $D\tau-$. The remaining 96.2% are unseen tweets in $D\tau-$.

In order to control for the effect that spam may have on our analysis, we have a collected a set of words that commonly appear in spam messages [1]. We then removed from our dataset every tweet that contained a word from this list. After the spam removal, from an evaluation on 300 tweets, we observed that none of the highly retweeted messages ($\geq$1000) were spam, while 3.85% of the mid-retweet and 3.23% of the low-retweet messages were considered to be spam ($\pm$ 5.66, confidence level 95%). The ranges for defining low-, mid- and high-retweets used were $[0, 10)$, $[10, 1000)$, $[1000, +\infty)$. We consider this an acceptable level of noise for the experiments in this paper. In future work we plan to evaluate the SONORA SCORE's robustness to noise.

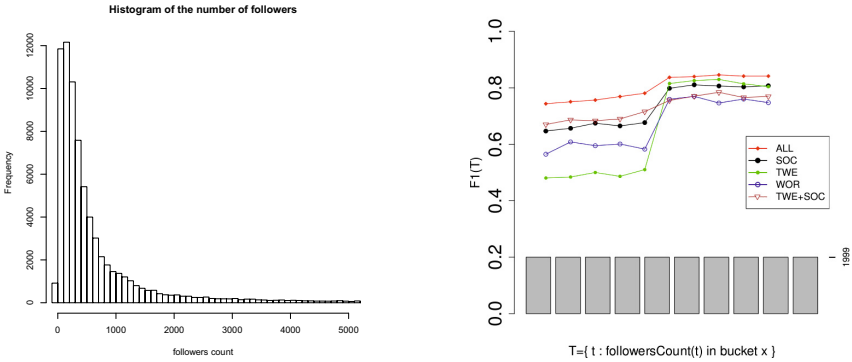
We have sampled $100,000$ tweets from $D\tau+$ that were written in English. We chose English as initial focus, as it is a common language between the authors, but the methods we present are not limited to any particular language. We used *langid.py* [15], an open source automated language detection tool with reported accuracy of 94% on Twitter text [15]. We performed an informal evaluation in our dataset with a random 100 tweets and verified that 87% were correctly detected as English. Tokenization was performed with a state-of-the-art tweet tokenizer from [20] that is aware of Twitter entities such as users, URLs and hashtags. Therefore, barring tokenization errors, occurrences of users, hashtags and URLs can also be treated as "words" for the features described in our approach section.

Features

Our experiments evaluate three categories of features: social features, tweet-level features and word-level features. The **social** features include characteristics related to the Twitter user network and are intended to help model a tweet's prior probability of getting retweeted based only on who is tweeting it to whom. Here, we use the number of followers and number of friends of a user[4]. The **tweet-level** features are intended to describe the tweet's *a priori* likelihood of getting retweeted without looking at the individual words they include or the message they convey. In this work we include the number of hashtags, number of URLs, number of user mentions and the number of stopwords present in a tweet. The **word-level** features, which are the novel contributions of this work and are formally defined in our approach section, focus on the mentions of words in tweets within a period or community of interest. The volume captures how frequently and to what extent tweets containing a word have been retweeted, the prevalence seeks to capture how steadily the word has appeared, and sustain captures for how long a "discussion" continues in which the word appears. For a comprehensive list of features considered by prior art and their predictive abilities please refer to the related work section.

Note that social and tweet-level features are observations, while word-level features are estimates. For instance, the number of hashtags is counted directly from each tweet being evaluated, and so is the number of followers. Meanwhile, the word-level features compute scores estimated from past tweets in $D\tau-$, i.e. not in the $D\tau+$ set from which the testing examples were sampled.

[4] Twitter API's 'Get Friend' request returns a collection of users the specified user is following. The 'Get Followers' request returns a collection of users following the specified user.



(a) The majority of users have less then 500 followers. Histogram plot clipped at 5000 for presentation purposes.

(b) Prediction F1 performance (y) varies with the number of followers (x). Lines represent different feature sets. Bars represent the sizes of each bin

Fig. 2. F1 improvement obtained from including SONORA SCORE features is higher for buckets with lower number of followers

Prediction Evaluation

In order to validate the effectiveness of the SONORA SCORE, we look to prior work that treats retweetability as an information retrieval problem – i.e. retrieve, from a collection of tweets, those that were retweeted. First, our datasets were preprocessed, assigning a label of 1 to the tweets where $RT(t) > 1$, and 0 otherwise. Under this experimental setting, different approaches can be compared by their ability to correctly retrieve as many of the retweeted posts as possible (high recall) while making as few classification mistakes as possible (high precision).

We selected 80% of the data for training, and performed testing with the remaining 50,000 tweets. Social features and tweet features were directly extracted from each tweet being tested. For the word features, the SONORA SCORES were computed from word occurrences in $D\tau-$, in order to avoid any bias. We experimented with three different popular models for predicting usefulness of features. In all cases, the models have been tested on the same sample from $D\tau+$.

First, we were interested in learning optimal weights to linearly combine our features into one SONORA SCORE that optimizes the likelihood a message will get retweeted. For that purpose, we have trained several Generalized Linear Models (GLM) [18] using subsets of our features. An advantage of GLMs is that they are transparent and friendly for user-interaction – the weights can be displayed as knobs or sliders on an interface, giving the users the freedom to disagree with our model's suggestions.

To explore non-linear combinations of our features, we also trained Conditional Inference Tree (CIT) models [9]. CITs learn relationships between features and the retweetability labels by recursively performing binary partitions in the feature space until no significant association between features and the labels can be stated.

Finally, we also included in our experiments Random Forests (RF) models [7], as they have been repeatedly shown in literature to perform well in a multitude of classification tasks. RFs learn a large number of decision trees that can be used as an ensemble to collectively predict the label at classification time. One disadvantage of RFs is that they operate as a black box, making it harder to take user input into consideration when tweaking the model.

Findings for Social and Tweet-Level Features. Table 1 shows the performance of each method (rows) in terms of F1 for each of the feature sets (columns). The social features have the best individual group performance, reflecting the intuition that the more popular you are, the more retweets you tend to receive. Out of the scores we tested, FOLLOWERS COUNT, USER MENTIONS, NUMBER OF HASHTAGS and NUMBER OF URLS were the most successful in predicting which tweets were retweeted.

Table 1. F1 results on our 50K sample from $D\tau+$using social features (SOC), tweet-level features (TWE), both social and tweet-level features (SOC+TWE), our word-level features (WOR) and a combination of all features (ALL)

	SOC	TWE	SOC+TWE	WOR	ALL
CIT	0.71	0.70	0.73	0.59	0.74
RF	0.76	0.70	0.75	0.70	**0.82**
GLM	0.65	0.69	0.66	0.56	0.68

It is a fact that the prior probability of a message getting retweeted at all, independently of what one is writing about, depends on how many people will see a particular tweet. A person cannot retweet a post they have not seen. The number of people that see a tweet, in turn, depends on a number of factors. Intuitively, the more followers a user has, the higher the likelihood that their tweets will be seen. One the other hand, hashtags are topic markers that are commonly used for searching, therefore could serve as a way to send the message beyond the stream of followers. Similarly, tweet analysis software will notify users when they are mentioned, which is another way to reach out to users that are not followers. Other features that may impact message visibility include time of the day, trendiness of topic, and other variables that are out of the scope of this work.

We argue that the high performance of the social and tweet-level features in our results confirms the intuition that targeting a larger audience increases the chances that someone will retweet a message.

Findings for Word-Level Features. Note that both social and tweet-level features were extracted from the actual tweet being evaluated, while the word-level features were estimated from data from the previous month i.e. $D\tau-$. We find it remarkable that they generalize well over time and offer powerful predictive ability. The best word-level model results (WOR) – i.e. based on RFs – are only .06 F1 points away from the best social (SOC) and tweet-level models (TWE). Moreover, by aggregating our word-level features with social and tweet-level (ALL), we are able to obtain an increase in .06 F1 points over the best social model (SOC), .12 over the best tweet-level model (TWE) and .07 over the combination of social and tweet-level features (SOC+TWE). This is

encouraging evidence that word-level features can support the detection of aspects of a tweet's message that lead to better or worse retweetability.

This is encouraging evidence that word-level features can support us detecting aspects of a tweet's message that leads to better or worse retweetability. This characteristic also sets us apart from previous work, as it does not suffice to focus on identifying "good" features, but also identifying "bad" features that when changed might provide better "uptake".

Prescription Evaluation

A prescriptive setting is substantially different from the tweet classification/retrieval setting discussed in previous work and evaluated in our previous section. When guiding users in formulating tweets to garner higher "uptake", our choice of features is limited to what can be changed at tweet authoring time.

For instance, in this setting the social features are fixed. If users want to achieve higher "uptake", they cannot easily enhance their social features instantaneously. Similarly, the tweet-level features that performed well in the predictive setting are rather vague in a prescriptive setting. Although we show that tweets with hashtags often get retweeted more, it is unclear from the tweet-level features which hashtags should be used. On the other hand, word-level features have the potential to help us to prescribe words (or hashtags) that have shown good historical performance in terms of a particular aspect that the user may want to explore.

In summary, we isolate and evaluate the performance of word-level features. We go about it in two ways. First, simulate a prescriptive setting by investigating the performance of each approach when the social features are fixed and tweet-level features are known. Second, we evaluate the likelihood that a suggestion given based on the SONORA SCORE will yield a higher retweet rate.

The effect of fixed social features. The histogram for the number of followers in Figure 2a shows that a large fraction of users in our sample have between 0-500 followers. Examining a particularly dense area (100-200 followers) shows (see Table 2) the F1 performance of their tweets which contain user mentions. This simulates a particular prescriptive setting, where a given user may be required to mention someone, while having at that point in time between 100-200 users. It can be noted that social features loose predictive power in this setting, as their variability decreases. The word-level features, however, not only retain their predictive power but are also prescriptive, allowing us to point out which words in a tweet have a low score.

While Table 2 shows a fixed number of followers and user mentions, Figure 2b shows the performance of each approach across groups of users with distinct follower counts. We start by selecting subsets of data with the same or similar number of followers. For that purpose, we sort the tweets based on their number of followers, sweep the space of FOLLOWERS COUNT, and partition the data into 10 bins. In Figure 2b, each bar on the horizontal axis represents one bin, each containing 2000 tweets. The small vertical axis on the right-hand size of the figure displays the scale of bin sizes.

Note, for users with a large number of followers, it is possible to predict retweetability very well based on social features alone. However, for tweets that were sent to

Table 2. F1 results on a subset with fixed settings: containing user mentions and with number of followers between 100 and 200. Columns refer to social features (SOC), tweet-level features (TWE), word-level features (WOR) and a combination of all features (ALL).

	SOC	TWE	WOR	ALL
CIT	0.63	0.53	0.63	0.45
RF	0.71	0.45	**0.79**	**0.83**
GLM	0.53	0.45	0.56	0.56

less than 800 followers, closer inspection of content is warranted. In those cases, adding word-level and tweet-level features to the mix significantly increases performance.

A similar effect to the followers count is seen with tweet-level features, which underperforms in tweets with lower number of followers and contributes more for popular users. In all cases, adding word-level features has helped increase F1 for all buckets for a given fixed social setting.

Success probability. Finally, we test how often a prescription from our system yields enhancements on a tweet's chance to be retweeted based on a Monte Carlo-style evaluation. We randomly draw (1M times, with replacement) two distinct tweets A and B from the set, and check: if $Sonora(A) > Sonora(B)$ is it also the case that $RT(A) > RT(B)$? In essence, this tests the assertion "If the system tells you one tweet is better than the other, what are the odds it is correct".

Table 3 can be interpreted as follows: Independent of the number of followers the best individual word-level predictor of retweet seems to be the Volume (prior effectiveness of those words). If combined with hashtags (as has been discussed prior) achieves $\approx .65$. This value can be thought of as "if the SONORA SCORE tells you that tweet A has a better phrasing than tweet B, it will be right twice as often as it is wrong.

$$\begin{aligned} 0.574 \cdot A_{min} + 1.51 \cdot A_{max} - .00003 \cdot V_{min} + .0002 \cdot V_{mean} \\ -0.0002 \cdot V_{max} + 0.0002 \cdot V_{sum} - 0.000005 \cdot P_{sum} \\ -0.002 \cdot S_{mean} + .0003 \cdot S_{max} - 0.00008 \cdot S_{sum} + \\ 1.35 \cdot \text{nURLs} - 0.637 \cdot \text{nUserMentions} + 1.29 \end{aligned}$$

Fig. 3. Formula of the best performing GLM (weights truncated for presentation)

We have searched through the space of feature combinations for those that could be used in a prescriptive setting with highest probability of success. Table 3 reports the results. The best performing model relied on A_{min}, A_{max}, V_{min}, V_{mean}, V_{max}, V_{sum}, P_{sum}, S_{mean}, S_{max}, S_{sum} as defined in our approach section, as well as nURLs, and nUserMentions. Figure 3 shows the formula for the best performing GLM, which obtained .69. These numbers are skewed by the "social" features not in this prescriptive model – with them included, the predictor goes to $\approx .8$, or odds are that it will be right four times as often as it is wrong.

In short, there are many factors that impact a tweets "success", but not unlike a spelling or grammar checker, our prescriptive model gives a fair indication of what needs a second look.

Table 3. Relative frequency of a SONORA SCORE recommendation yielding an RT improvement

Model	Feature Sets	P(success)
CIT	Amplitude, Volume, Prevalence, Sustain, nURLs, nUserMentions	72.12%
GLM	Amplitude, Volume, Prevalence, Sustain, nURLs, nUserMentions	69.04%
CIT	Amplitude, Volume, Prevalance and Sustain	68.71%
GLM	Amplitude, Volume, Prevalence and Sustain	66.67%
CIT	Volume, nHashtags	65.44%

Prescribing word substitutes for an audience. In order to prescribe message enhancements at authoring time, our system needs to find word substitutes that convey an equivalent message and that have better performance within an audience. In this work we have used WordNet [16] synonyms as our source of equivalent words. Other lexical databases, thesauri, or techniques for automatic extraction of related words [13] could also be used. To estimate better performance within an audience, we use the aforementioned word-level features.

Table 4. Ranked Wordnet synonyms of 'great' for two datasets ($\rho = 0.74$). Omitting the top 3 suggestions (great, greatest and greater) as they did not vary between groups.

interest	**synonym of 'great'**
poetry	..., cracking, bully, **neat**, smashing, keen, **swell**, **nifty**, groovy, dandy, bang-up
science	..., **neat**, keen, smashing, **nifty**, cracking, bully, groovy, dandy, bang-up, **swell**

In this setting, two questions come to mind. First, do different words really resonate differently with different audiences? Second, do audiences vary only in their interests for different topics, or do they also differ in their preference for more topic-independent language constructs such as adjectives?

Given a word w_i, we look at how well each of the synonyms of w_i have performed for a target audience. For example, consider two audiences: one interested in poetry and one interested in science. If both groups seem to prefer the same synonyms, then there is little a system can do to prescribe word substitutions to enhance a message's score using this method. However, if there is a difference in preference for different synonyms between the groups, then there is a real opportunity to prescribe message changes to target different audiences.

We collected the set of all synonyms in WordNet for all words mentioned in our dataset from March and April, keeping the synsets with size greater than two. We also selected two focused tweet datasets, filtering for all users mentioning 'poetry' or 'science' in their user profiles. Then, for each set, we ranked the synonyms according to the number of times they appeared in a retweeted message. We then contrasted the rankings

with one another. As an example, Table 4 shows synonyms of 'great', and the rank for each of its synonyms in the 'poetry' and 'science' sets. Although the most common forms 'great', 'greater' and 'greatest' seem to be the most common within both groups, there is significant difference in the usage of 'nifty', 'neat' and 'swell'.

We used Spearman's rank correlation coefficient (ρ) to quantify the similarity in ranking between two synonym sets. A $\rho = 1$ indicates that the two groups have identical preference within that synonym set, while a $\rho = -1$ indicates opposite preference. The rankings shown in Table 4 have $\rho = 0.74$.

In order to control for variation in language that may occur in a dataset by chance, we analyzed the distribution of ρ scores for our focused datasets in contrast with a base dataset of retweets in March and in April. Table 5 shows the average ρ values for our focused (poetry-science) and base datasets (mar-apr).

Table 5. Rank correlation ρ between two random samples (mar-apr) is higher than between two samples focused on different audiences (poetry-science)

data sets / ρ	adjectives	nouns	verbs
poetry-science	0.79	0.73	0.73
mar-apr	0.88	0.79	0.76

For all three classes of words, the rank correlation in our base dataset is higher than in our focused data sets. Therefore, different synonyms get retweeted at different rates for the poetry and science audiences(more so than two random samples). This highlighs the importance of word choice for message authoring, and shows that the differences in language go beyond interest in different topics.

Applications

The SONORA SCORE application prototype has been developed to help social media communication professionals message to varied audiences. However, for the sake of the argument, let us pick an intuitive example. Consider two authors: June, a teenage girl, and Augustus, a middle aged man. Both are trying to write a tweet thanking a friend for introducing them to people in a social situation. We select these two social groups as there is a strong difference in common diction between them; a difference that most people are at least somewhat familiar with.

Audience Specific Sonora Score. First, let's consider the two tweets being sent out. June is going to tweet. We highlighted (e.g. [green]) those non-stop words that particularly resonate (i.e., have a high SONORA SCORE) with her target audience: *"give a [shoutout] to my [awesome] BFF for bringing me to her classmate's [kickin] hip-hop party."* Contrast the language with Augustus' tweet: *"[thanks] to my social guru for helping me shamelessly [network] at the [xfactor] party - mucho appreciated."* Both of these authors clearly reflect (or understand) their target audiences and are writing

appropriately for their peer groups. Several of the words they are using have very good SONORA SCORE and none of them are particularly poor choices. But consider what would happen if June had sent a tweet phrased like Augustus' to her peer group. We highlighted (e.g. *red*) the words that may be particularly poor choices (i.e., have a low SONORA SCORE): *"thanks to my *social guru* for helping me *shamelessly* network at the *xfactor* party - *mucho appreciated*."* It can be noticed that many of the "good" words for the older target audience are just average with this one, and a number of the words drop to *red* warnings as likely to not "resonate" well for June. The story is not all that much better if Augustus were to use language from the "younger" diction in his tweet: *"give a *shoutout* to my *awesome BFF* for bringing me to her *classmate's kickin hip-hop* party."* Again we see that this language would likely not resonate very well with Augustus' peer group. This ability to analyze tweets for a particular audience provides the opportunity to do "synthetic market studies" of every tweet looking at a specific target audience. This can be a very useful tool for those who are directing their communications to, at times, strongly differing audiences.

Mobile App Prototype. Implemented as a simple Twitter posting client (see Figure 1), the iOS application prototype allows a user to type a proposed tweet (given a target audience), and get a real-time estimate of how well it will resonate. The design of the user interface was guided by 3 main themes established by the iOS Human Interface Guidelines [10] (i) **deference:** the user interface helps the user to understand and interact with the content, but never competes with it; (ii) **clarity:** text is legible at every size, icons are precise and lucid, adornments are subtle and appropriate, and a sharpened focus on functionality motivates the design; (iii) **depth:** visual layers and realistic motion impart vitality and heighten users' delight and understanding. More specifically we wanted the user interface to: (a) provide immediate feedback on the effectiveness of a tweet while composing it (b) display the SONORA SCORE for each word and the complete tweet; (c) provide detailed SONORA SCORES for individual words; (d) present the user with a list of synonyms with higher SONORA SCORES to replace a word with a low score; (e) follow established interaction patterns of the platform.

To achieve the first objective the user interface makes use of color to show the SONORA SCORE for each word as well as for the overall tweet to provide the user with an immediate understanding of the effectiveness of each word in the tweet. Additionally the overall SONORA SCORE for the tweet is shown in numerical representation, as is the number of characters remaining from the 140 character limit for a tweet. Tapping a word twice highlights the word and provides the detailed SONORA SCORE for the selected word. If synonyms are available they are presented in a context menu layered above or beyond the selected word. Tapping one of the synonyms replaces the selected word with the selected synonym. We chose to use a context menu due to it being a familiar interaction pattern for text manipulation on the iOS platform. The screenshot in Figure 1 demonstrates message composition using the above outlined techniques. In this case while the application notes "amazing" resonates slightly better, June feels the nearly as good "kickin" is a better word choice and goes with it. Again, SONORA SCORE may best be thought of as a "grammar checker" that is often right, but still needs a human to make the decisions on best phrasing.

Conclusion

Effective messaging is an integral part of building and engaging relevant communities. By using poorly chosen words and tone, one can distort a message rendering it ineffective or worse, alienating communities which may have a negative impact on brand image. A well "sounding" message not only gets shared more, it also increases the chances that the user will be cited in the future and be considered influential. In this work we presented SONORA SCORE, a new feature based on word distribution that is aimed at helping authors construct better messages on Twitter. We evaluated the score and investigated the role word choice plays in retweetability. SONORA SCORE models three aspects of retweetability - volume, prevalence and sustain - that allow us to predict message popularity and prescribe word choice for message optimization. While past work has focused on identifying features that make for a popular tweet, none of them are prescriptive in nature, i.e. offer suggestions on what to change for better "uptake".

We found that SONORA SCORE serves as a good predictor of retweetability, and complements known predictors such as social and tweet-level features. For users with fewer followers our score's predictive ability is even higher; suggesting that if you cannot change their popularity or their topic, word choice plays a very important role. Based on a Monte Carlo-style evaluation, we found that when SONORA SCORE gives a higher score to a tweet, it is twice as likely to also have higher retweet. Although the experiments presented here focus on tweets, the SONORA SCORE prescriptive model applies to practically all types of messages and networks – emails, blogs, news articles etc., since the basic building block for the SONORA SCORE is at the 'word' level.

This paper has focused on the retweet rate as a measure of success. In future work we plan to study also the length of the retweet cascade, among other aspects of successful tweets.

Acknowledgements. We thank the anonymous reviewers for their careful reading and their many insightful comments and suggestions.

References

1. Alexe, B., Hernandez, M.A., Hildrum, K.W., Krishnamurthy, R., Koutrika, G., Nagarajan, M., Roitman, H., Shmueli-Scheuer, M., Stanoi, I.R., Venkatramani, C., Wagle, R.: Surfacing time-critical insights from social media. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, pp. 657–660. ACM, New York (2012)
2. Areni, C.S.: The effects of structural and grammatical variables on persuasion: An elaboration likelihood model perspective. Psychology & Marketing 20(4), 349–375 (2003)
3. Bigonha, C., Cardoso, T.N., Moro, M.M., Gonçalves, M.A., Almeida, V.A.: Sentiment-based influence detection on twitter. Journal of the Brazilian Computer Society 18(3), 169–183 (2012)
4. Bormuth, J.R.: Readability: A new approach. Reading Research Quarterly, 79–132 (1966)
5. Guerini, M., Pepe, A., Lepri, B.: Do linguistic style and readability of scientific abstracts affect their virality? In: ICWSM (2012)

6. Hansen, L.K., Arvidsson, A., Nielsen, F.A., Colleoni, E., Etter, M.: Good friends, bad news-affect and virality in twitter. In: Park, J.J., Yang, L.T., Lee, C. (eds.) FutureTech 2011, Part II. CCIS, vol. 185, pp. 34–43. Springer, Heidelberg (2011)
7. Ho, T.K.: Random decision forests. In: Proceedings of the 3rd International Conference on Document Analysis and Recognition, August 14-16, pp. 278–282. IAPR (1995)
8. Hong, L., Dan, O., Davison, B.D.: Predicting popular messages in twitter. In: Proceedings of the 20th International Conference Companion on World Wide Web, pp. 57–58. ACM (2011)
9. Hothorn, T., Hornik, K., Zeileis, A.: Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics 15(3) (2006)
10. Apple Inc. ios human interface guidelines (2013) (Online: accessed October 3, 2013)
11. Smith, A., Brenner, J.: 72% of Online Adults are Social Networking Site Users (2013)
12. Ke, J., Gong, T., Wang, W.S.: Language change and social networks. Communications in Computational Physics 3(4), 935–949 (2008)
13. Kim, J.-K., de Marneffe, M.-C.: Deriving adjectival scales from continuous space word representations. In: EMNLP, pp. 1625–1630. ACL (2013)
14. Lakkaraju, H., McAuley, J.J., Leskovec, J.: What's in a name? understanding the interplay between titles, content, and communities in social media. In: International Conference on Weblogs and Social Media (2013)
15. Lui, M., Baldwin, T.: langid.py: An off-the-shelf language identification tool. In: ACL (System Demonstrations), pp. 25–30. The Association for Computer Linguistics (2012)
16. Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM 38(11), 39–41 (1995)
17. Nagarajan, M., Purohit, H., Sheth, A.P.: A qualitative examination of topical tweet and retweet practices. In: ICWSM (2010)
18. Nelder, J.A., Wedderburn, R.W.M.: Generalized linear models. Journal of the Royal Statistical Society. Series A (General) 135(3), 370–384 (1972)
19. Petrovic, S., Osborne, M., Lavrenko, V.: Rt to win! predicting message propagation in twitter. In: ICWSM (2011)
20. Ritter, A., Clark, S., Mausam, Etzioni, O.: Named entity recognition in tweets: An experimental study. In: EMNLP (2011)
21. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: 2010 IEEE Second International Conference on Social Computing (SocialCom), pp. 177–184. IEEE (2010)
22. Tan, C., Lee, L., Pang, B.: The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In: Proceedings of ACL (2014)
23. Yang, J., Counts, S.: Predicting the speed, scale, and range of information diffusion in twitter. In: ICWSM, vol. 10, pp. 355–358 (2010)