

# What's in a Hashtag?

## Content based Prediction of the Spread of Ideas in Microblogging Communities

Oren Tsur  
Department of Computer Science  
The Hebrew University  
Jerusalem 91904, Israel  
oren@cs.huji.ac.il

Ari Rappoport  
Department of Computer Science  
The Hebrew University  
Jerusalem 91904, Israel  
arir@cs.huji.ac.il

### ABSTRACT

Current social media research mainly focuses on temporal trends of the information flow and on the topology of the social graph that facilitates the propagation of information. In this paper we study the effect of the content of the idea on the information propagation. We present an efficient hybrid approach based on a linear regression for predicting the spread of an idea in a given time frame. We show that a combination of content features with temporal and topological features minimizes prediction error.

Our algorithm is evaluated on Twitter hashtags extracted from a dataset of more than 400 million tweets. We analyze the contribution and the limitations of the various feature types to the spread of information, demonstrating that content aspects can be used as strong predictors thus should not be disregarded. We also study the dependencies between global features such as graph topology and content features.

### Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

### General Terms

Theory, Measurements

### Keywords

Social media, information diffusion, microblogging, Twitter, hashtags

## 1. INTRODUCTION

Social media is a natural platform for the spread of thoughts and ideas, sometimes called *memes*. The spread and propagation of a meme through a social network attract a lot of research in recent years. Successful prediction of the spread

of memes can improve “marketing” efforts, whether the target is a commercial product or an idea being promoted. It could also help in real time identification of new trends, both commercial and ideological.

From a graph topology perspective the social network is viewed as a graph of nodes (users/posts) connected by edges (readers in blogs, fans in Facebook, followers on Twitter etc.). Recent studies focused on the *topology* of the social graph, investigating what topologies and what activation patterns facilitate efficient propagation of memes. While graph topology plays an important role in the spread patterns of ideas, the *content* of the meme is also of great importance to the acceptance and promotion of a meme within the community. To the best of our knowledge, no prior work studies the way the inherent content features affect propagation.

In this work we propose a complementary framework: given an idea/meme  $m$ , and a time frame  $t$ , can we predict the acceptance of  $m$  in the community (a social network) within this horizon? We are interested in the following questions: can we accurately predict the acceptance of a meme based solely on the meme's content? Does the meme's context improve the prediction? What are the relations between the graph topology and the content and how do they integrate to facilitate efficient propagation?

Popularity is captured by the normalized count of the meme's occurrences in the given time frame. It is important to clarify that we are not interested in the temporal spreading patterns, although the temporal pattern serves as a feature type in our hybrid model.

The rationale behind focusing on content is threefold. First and foremost, there is a genuine interest in the way the content and structure of an idea drives its acceptance. Moreover, relying only on graph topology disregards content altogether, implicitly assuming all memes are born equal (if seeded equally). Second, the topology of the social graph is not always given, while the content is usually available online. Third, graph based algorithms are NP-hard and even approximation algorithms are ineffective for large graphs in real time.

A common definition of a meme is a short unit of text that passes relatively unchanged through many online sources [24]. We look at data from Twitter, considering hashtags as potential memes. A Twitter *hashtag* is a string of characters preceded by the hash (#) character. In many cases hashtags can be viewed as topical markers, an indication to the context of the tweet or as the core idea expressed

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WSDM'12, February 8–12, 2012, Seattle, Washington, USA.  
Copyright 2012 ACM 978-1-4503-0747-5/12/02 ...\$10.00.

in the tweet, therefore hashtags are adopted by other users that contribute similar content or express a related idea. A few examples of the use of hashtags are: “ask GAGA anything using the tag #GoogleGoesGaga for her interview! RT so every monster learns about it!!” referring to an exclusive interview for Google by Lady Gaga (singer), “Whoever said ‘youth is wasted on the young’ must be eating his words right now. #March15 #Jan25 #Feb14”, referring to the protest movements in the Arab world<sup>1</sup>, or “Speaker refers to #Lansseys ‘abysmal ignorance’ as demonstrated on alcohol strategy; this SoS #notfitforpurpose #nhs #savethenhs #healthbill”, referring to the “national health services” and the “health bill”.

The acceptance of a hashtag is captured by the (normalized) count of its appearance in a time interval. Given a hashtag, we aim at predicting its frequency after some time. We cast the problem of predicting the hashtag frequency as a regression task. We test our model on a large dataset which consists of more than 4 million tweets and thousands of hashtags. Our experiments show that there are three main factors to the acceptance of a meme: the meme’s content, the meme’s context and the social graph. A hybrid model combining all three factors performs best. In this paper we report several combinations of feature types, including models that incorporate early temporal patterns.

This work has two main contributions: (a) to the best of our knowledge, it is the first study that shifts the emphasis towards content features, and (b) we present a simple yet robust framework that efficiently models the exposure and acceptance of an idea/meme, using only global features, avoiding costly graph based algorithms.

This paper is arranged as follows. The next section surveys related work. Section 3 portrays the corpus we used and some preprocessing steps. Section 4 formally presents the prediction model and the feature types we employed, while Section 5 presents the experimental setup. Results are given in Section 6 and an elaborated discussion is offered in Section 7. In Section 8 we conclude and offer directions for future research.

## 2. RELATED WORK

Diffusion of information has become an active research area. Most works focus on the topology of the social graph, trying to model the propagation process, maximize the spread of information in a minimal effort by finding the most influential nodes, and maximize purchases induced by viral marketing and social recommendation networks [18, 19, 23, 4, 5, 13], or model temporal dynamics of information spread [12, 15, 34]. Collaborative filtering is used to predict the probability of a tweet to get retweeted [36].

Tweet’s content is used in tasks such as profiling users according to substance, style, status, and social tendency [27] and for modeling the inter-influence of the linguistic style between participants in the discourse [7]. The effect of hashtag length (in characters) on its frequency is studied in [6]<sup>2</sup>.

Some works do refer to the topic of the information passed [14, 24, 34], though topic is addressed in coarse granular-

<sup>1</sup>The protest movements are named after the date of the first demonstration, i.e. Jan25 is the Egyptian movement.

<sup>2</sup>This study refer to hashtags relevant to three events: Micheal Jackson death, the swine flu outbreak of 2009 and to the Twitter idiom ‘music-monday’.

ity (e.g. ‘sports’, ‘news’, ‘professional blogs’, ‘Apple’, ‘Microsoft’) and not very specific events or sentiment (e.g. ‘save the national health service’, ‘Gaga Video Music Awards’, ‘free Iran’). Leskovec et al. [24] look at the content in their meme-tracker, but textual content is only used in order to trace the evolution of a meme as it propagates.

Another line of works study the spread patterns of Twitter hashtags. Yang and Leskovec [34] propose a linear model that uses the implicit network rather than the explicit network. Cataldi et al. [3] detect emerging topics in Twitter although they refer to term frequency of tokens in Twitter. Although not using the explicit graph topology, both works use the level of influence of single nodes in order to model spread of information.

Romero, Meeder and Kleinberg [29] observe that different topical categories of hashtags have different propagation pattern. They introduce the distinction between ‘stickiness’ and ‘persistence’ arguing that some classes are more persistent than other. This work is the first to address attributes that are inherent to the meme and that goes beyond the most coarse topical propagation.

Our work shares these observations. However, it differs in a number of fundamental aspects. First, our task definition is different as we are interested in predicting the exposure of a meme in a given time frame while they are interested in the temporal spreading patterns. Second, we do not use annotators to classify hashtags, and third, while they model the spread patterns of only the most popular hashtags, we aim at predicting the spread of all hashtags but the least frequent ones.

Unlike other works, we are interested in the acceptance of a meme in a long time frame, as it is better to differentiate between sticky ideas (and successful campaigns) and short term trends that flare, infect many users but quickly disappear.

As we are interested in content based analysis, the sentiment of a tweet is of great importance. Sentiment analysis of Twitter data is studied in many works, [8, 9] among others. In this work we use the LIWC categories [32] and assign sentiment labels to tweets and hashtags based on the LIWC lexicons.

To the best of our knowledge, beyond the hashtag’s coarse topic or its number of characters, there is no prior work addressing the features inherent to the information (meme, hashtag), such as location, orthography, number of words, lexicality, ease of cognitive process and emotional effect on various cognitive dimensions. These attributes are described in detail in Section 4.3.

## 3. DATA

### Twitter & hashtags.

Twitter is a popular microblogging platform. A Twitter posting is called a *tweet*. A tweet is restricted to 140 characters in length. This length constraint makes characters “expensive”, hence tweets present an informal, sometimes ungrammatical, language, as well as introducing many abbreviations. Twitter allows the usage of two meta characters: @ marking a user name (e.g. @BarackObama), and # marking a hashtag: a sequence of non whitespace characters preceded by the hash character (e.g. #healthCareReform). An extensive survey of Twitter, its uses and its social as-

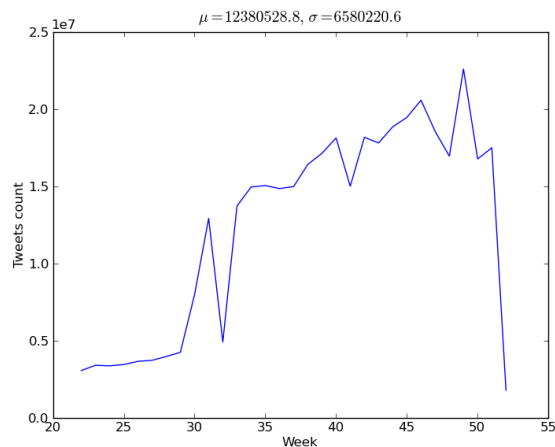


Figure 1: Plot of the number of tweets collected in weeks 22–52 (June–December) of 2009.

pects in the first years can be found in [20]<sup>3</sup>. The use of hashtags is a popular way to give the context of a tweet, an important function due to the length constraint. For example, the hashtag *#savethenhs*, reads as ‘save the national health service’, gives the context relevant to the tweet “Speaker refers to *#Lanseys* ‘abysmal ignorance’ as demonstrated on alcohol strategy; this SoS *#notfitforpurpose* *#nhs* *#savethenhs* *#healthbill*”. The hashtags *#iranelections* and *#freeiran* give the context and ideology behind the tweet “AP: Report: *#Iran*’s paramilitary launches cyber attack <http://is.gd/HiCYJU> *#iranelections* *#freeiran*”. Also note the use of *#Iran* both as a hashtag and as a crucial part of the sentence.

Analysis of the content of hashtags poses some interesting technical problems. The structure of the hashtag has no restrictions. A hashtag can be a lexical word (*#iran*), a compound of lexical words (*#freeiran*), or (compound with) abbreviations (*#savethenhs*). Moreover, while the Twitter engine is case insensitive, users do use variations such as *#freeiran* (2203), *#FreeIran* (729), *#freeIran* (319), *#Freeiran* (46), *#FREEIRAN* (44) or *#freeIRAN* (30). Although semantically identical, the usage frequencies of these hashtags differ greatly, as indicated in parenthesis.

### The Corpus.

Our data consists of more than four hundred million tweets tweeted between June–December 2009<sup>4</sup>. The data was collected using Twitter API (the Spritzer and Gardenhose services) and was submitted to Twitter’s streaming policy controlling the sample size collected every day (5%-15% sample of the non-private tweets, uniformly distributed). As Twitter gained popularity and as the sampling rate changed from time to time, the total number of tweets collected every week varies greatly. Figure 1 presents the weekly change in the sample size with standard deviation of 6.5 million.

We filtered out tweets that contain non-Latin characters,

<sup>3</sup>Their survey is based on about one million tweets in the span of three years. The current stream of Twitter messages is orders of magnitude larger.

<sup>4</sup>The data was collected and used for [26] and is generously shared at: <http://www.ark.cs.cmu.edu/tweets/>.

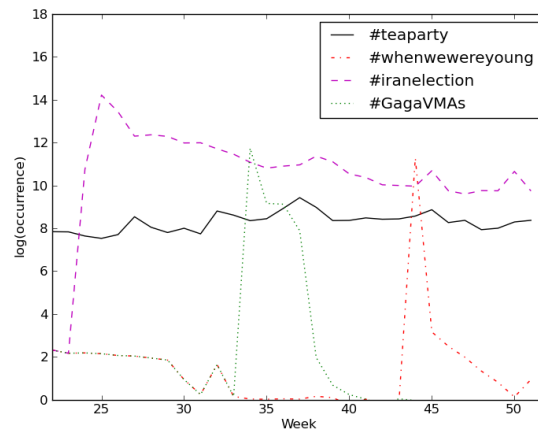


Figure 2: Four typical temporal trends (unnormalized counts).

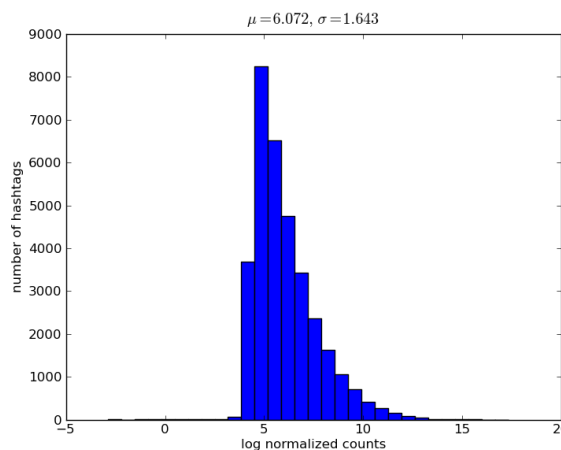


Figure 3: Histogram of the log count of all the hashtags in our dataset (7 months).

trying to maintain a corpus of only English tweets. Although we managed to remove all East Asian tweets, our corpus still contained some non-English tweets mainly in Spanish and Dutch. Non-English tweets can bias the results in case of hashtags that are composed of non-English words or English hashtags that appear in non-English content. A less noisy corpus can lead to improved results.

Another possible bias may be introduced due to spam tweets and spam hashtags. Identifying spam in Twitter is beyond the scope of this work, thus we cannot estimate the biased introduced by spammers.

### Normalization of counts.

The great variance in the weekly sample size introduces some bias to the hashtag counts, unless one assumes that all hashtags have the same parametric distribution (over time). This assumption does not hold as different hashtags present different temporal patterns. The same hashtag could have gained different number of counts in our dataset in case it

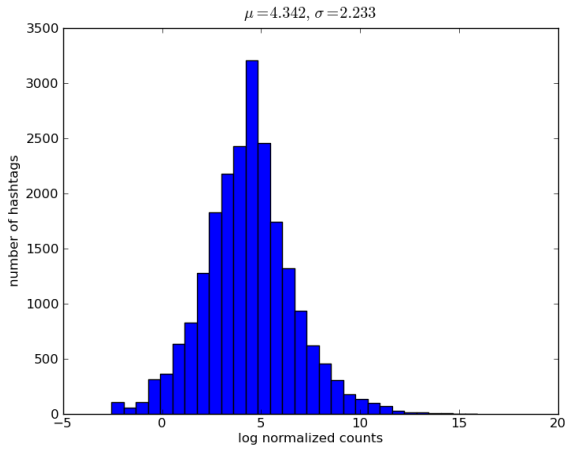


Figure 4: Histogram of the log count of ‘fresh’ hashtags in 10 weeks time frame.

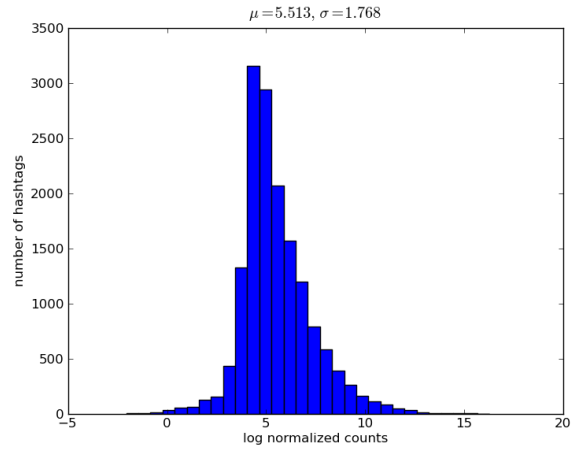


Figure 6: Histogram of the log count of ‘fresh’ hashtags in 20 weeks time frame.

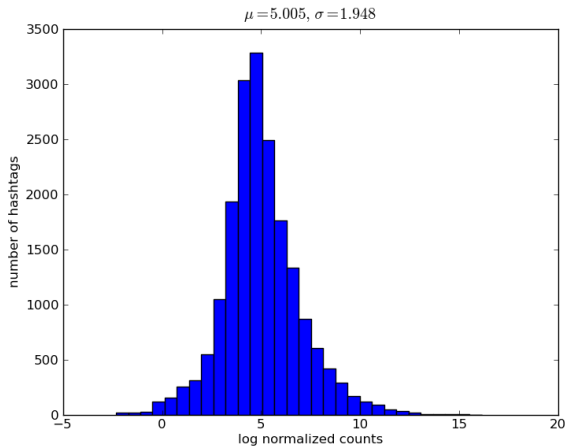


Figure 5: Histogram of the log count of ‘fresh’ hashtags in 15 weeks time frame.

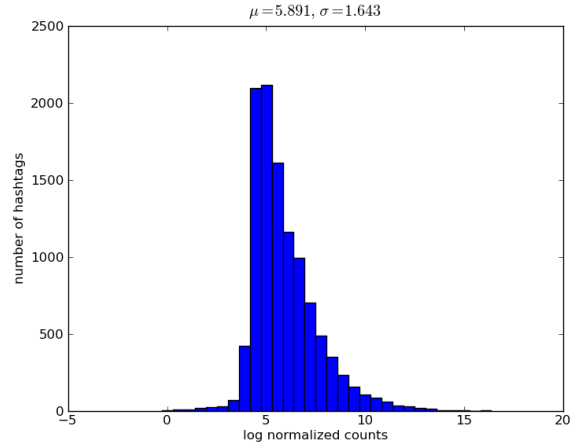


Figure 7: Histogram of the log count of ‘fresh’ hashtags in 25 weeks time frame.

was introduced in different weeks simply due to the variance in the sample size. See the weekly unnormalized counts of a few hashtags in Figure 2 (and compare to the weekly sample size in Figure 1). In order to address this bias we normalize the hashtag counts in the following way:

$$N(ht^i) = \sum_{j \in \text{weeks}} \text{count}(ht_j^i) \cdot \frac{w'}{w_j}$$

Where  $ht^i$  is the  $i$ -th hashtag,  $\text{weeks} = \{22, \dots, 52\}$ , the sampled weeks,  $\text{count}(ht_j^i)$  is the number of times hashtag  $i$  occurs in the sample of week  $j$ ,  $w_j$  is the sample size in week  $j$  and  $w'$  is a constant, we use  $w' = w_1$  so the sample size of the first week is the base for the normalization.

In our experiments we looked only at hashtags which occur over 100 times; assuming that less frequent hashtags are either typos (typically a missing trailing white space hence creating a concatenation of tokens that are not part of the tag), used only in a weakly connected component of the

social graph or were introduced at the end of our date range therefore did not have the chance to be widely adopted (or ignored).

We look only on ‘fresh’ hashtags – hashtags that did not get popular before our corpus was collected (e.g. unlike *#teaParty* in Figure 2). We define a hashtag as fresh if it did not appear in the first week (week 22) or if its normalized count in the first week is less than 10% of its normalized count in its peak. Figures 3, 4, 5, 6 and 7 show the differences between the long tailed distributions of all the hashtags vs. the normal distribution of the fresh ones (we elaborate on the differences between different horizons in the General Discussion section).

As mentioned above, all hashtags that appear less than 100 times ( $< 250e^{-9}\%$ ) were removed from the dataset.

## Preprocessing.

Identifying the distinct words composing a hashtag is a mandatory step in order to process it and convert it to its vectorial representation. Matching hashtags against a lexicon of English words and using dynamic programming to perform segmentation of compound hashtags seems straightforward. However, many compound hashtags consists of non standard words and abbreviations that do not appear in standard dictionaries, e.g. ‘luv’ instead ‘of love’ in *#welwvjb* (reads as ‘we love Justin Beiber’), and *#savethenhs* (reads as ‘save the national health service’). In order to perform the segmentation of compound hashtags, besides matching hashtags against lexicons, we employed a simple heuristic algorithm that exploits redundancy of hashtags that differ only orthographically. The algorithm matches tuples like *#freeiran*, *#FreeIran* and *#FreeIran* and performs segmentation according to the capital letters, assuming they are used for visual segmentation (just like naming conventions of variables in programming). This heuristic algorithm achieves high precision and decent recall. However, we spend a few hours validating the results and manually fixing some untokenized hashtags.

## 4. PREDICTION MODEL

In order to predict the acceptance of a hashtag we learn a regression model. In this section we describe our regression model and the features we utilize in our model.

### 4.1 The Target Function

In a regression task, we want to learn a target function  $f(ht) = n$ , where  $ht$  is a vector space representation of a given hashtag, and  $n \in \mathbb{N}_0$  is the normalized count of its occurrences in a time frame. Function  $f$  is typically learned from a training set of example hashtags and their counts,  $\{ht_i, n_i\}$ . In this work we learn the transformed target function  $f'(ht) = \log(n)$ .

Using  $f'$  instead of  $f$  serves two purposes: (a) We are interested in predicting the magnitude of the acceptance in a time frame, thus while a hashtag occurring 500 times is very different from a hashtag occurring 1000 times, occurrence of 30000 is similar to 30500.  $f'$  captures this observation. (b)  $f'$  allows us better smoothing of temporal variations induced due to different times of first appearance in the dataset.

### 4.2 Regression Model

We denote the training set by  $(X, Y) = \{x_i, y_i\}$ , where  $x_i$  is a feature vector representation of hashtag  $ht_i$  and  $y_i = \log(n_i)$ , is the log of the normalized number of its occurrences. A linear function  $b + w^T X = b + \sum_j w_j^T X^j$  is a simple yet robust way to model the dependency of  $Y$  on  $X$  [11].

In its standard setting, the linear function is found analytically, however, this requires inversion of large matrices, thus impractical. Instead we use stochastic gradient descent. We use an  $L1$  regularization to learn the optimal model parameters  $b$  and  $w$ :

$$L_r(b, w) = \frac{1}{2} \sum_i \left( y_i - (b + \sum_j w_j^T x_i^j) \right)^2 + \frac{1}{2} \lambda \|w\|$$

Where  $i$  denotes the  $i$ -th training example and  $j$  denotes the  $j$ -th attribute in  $x_i$ . Regularization is introduced in order to overcome sparsity and to lower the risk of overfitting.

To learn the parameter values we use Stochastic Gradient Descent (SGD) [2], for which the parameter update is as follows:

$$\begin{aligned} \Delta b &= \eta_t (y_i - (b + w^T x_i)) \\ \Delta w_i &= \eta_t (y_i - (b + w^T x_i)) x_i - \lambda w_i \end{aligned}$$

We cycle through random permutations of the observations to achieve convergence. For each learning rate we use a schedule of the form  $\eta_t = \frac{\eta_0}{t+\tau}$  where  $\tau > 0$ , and  $t$  is the number of epochs. The schedule satisfies the Robbins-Monro conditions [28],  $\sum \eta_t = \infty$  and  $\sum \eta_t^2 < \infty$ , hence convergence is guaranteed.

The final model strongly depends on a proper choice of hyper-parameters,  $\eta$  and  $\lambda$ . We use the Nelder-Mead [25] method in order to find the optimal values of the hyper parameters. Though not guaranteed to converge to the function’s minimum [21], it is a widely used algorithm with excellent results on real world scenarios [33].

### 4.3 Model Features

In order to learn our regression models we represent each hashtag as a binary vector. We define four types of features:

1. Hashtag content – features that can be extracted from the hashtag itself.
2. Global tweet features – features related to the content of the tweets containing the hashtag.
3. Graph topology features – features related to graph topology and retweet statistics.
4. Global temporal features – features related to temporal pattern of the use of the hashtag.

All features are binary therefore each attribute in each feature type is represented by a short feature vector. The concatenation of these vectors spans our vectorial space. In the remaining of this section we give a detailed description of the extracted features along with the intuition behind using these features.

#### 4.3.1 Hashtag content

For each hashtag we extract attributes belonging to a number of main categories: length (characters and words), hashtag orthography, emotional and cognitive dimension, hashtag location, and match with lexical lists such as general dictionary, proper names, holidays and celebrity names. These attributes are extracted based on the hashtag alone, e.g. *#freeIran* is 9 characters long, 2 words long, both words are listed lexical items, Iran is a state, the word free is a positive sentiment word etc.

#### Character length.

As a tweet’s length is limited to 140 characters, each character becomes expensive. Using a hashtag as part of the tweet consumes space that could be used for the free text. On the other hand, if a hashtag is too short it might not be understood and will not serve its purpose. We used 7 bins for this attribute, bins capture hashtag of 2,3,4,5,6–9,10–14 and >14 characters long.

### *Number of words.*

55% of the hashtags in our data are compounds of more than one word (e.g. *#freeIran*, *#GoogleGoesGaga*). A word compound can make a hashtag/meme clearer or too complex, thus more/less appealing to be adopted by users. Four bins were used for this feature: 1 word, 2,3 and 4 words or more. Acronyms (e.g. *#NYC*, *#nhs* and *#ff*) were considered one word.

### *Orthography.*

Hashtags can be written in capital letters, contain some capital letters and/or digits, e.g. *#myheart4JB*, reads as ‘my heart for Justin Beiber’. Using the ‘right’ writing style can make the hashtag readable (*savethenhs* vs. *saveTheNhs*), while on the other hand it requires more typing effort (requires using the shift key). We use four attributes for this group: no caps, some caps, all caps and contains digits.

### *Lexical items.*

Working on a development set, we noticed that internet memes sometimes evolve around celebrities, holidays or locations. We match the hashtag or its words against five predefined lists: (1) a general lexicon containing all words from a large portion (612MB) of English Wikipedia (approximating that the hashtag is a proper English word), (2) a list of proper names taken from the name list compiled at the US census of 1995 and published online, (3) a list of celebrity names compiled from Forbes’ ‘The Celebrity 100’ lists of 2008–2010<sup>5</sup>. (4) A short list of holidays and days of the week and (5) a list of all the world’s countries. Each of these five lexicons is an attribute in our vector.

### *Location.*

Hashtags can appear anywhere in a tweet. The location of a hashtag can give an indication to the way it is used<sup>6</sup>. For example, if located in the middle of the tweet, the hashtag also serves as part of the sentence and not only as a meta tag. We acknowledge three locations: prefix, infix and suffix. We treat sequences of hashtags as a single location, thus, for example, the two last hashtags in “*AP: Report: #Iran’s paramilitary launches cyber attack http://is.gd/HiCYJU #iranelections #freeiran*” are both considered to be suffixes, while *#Iran* is infix. Generally, if a hashtag has more than 25% of its occurrences in one position it is considered to fit this position role. We note that a hashtag can, therefore, have up to three locations.

### *Collocation.*

Some hashtags tend to collocate with other hashtags. This was captured with one binary attribute, where the value 1 is assigned if more than 40% of the hashtag occurrences are collocated with other hashtags.

### *Cognitive dimension.*

Words can be classified according to a number of psy-

<sup>5</sup>Based on a development set we added 3 names of recently emerged celebrities to the celebrity lexicon.

<sup>6</sup>Ideally, we would like to know what POS tag is most suitable for a given hashtag, however, the informal nature of tweets along with the fact that hashtags can be composed of a number of words, sometime a complete clause, along with the lack of POS annotated Twitter data, makes current POS tagging techniques inapplicable.

chological and cognitive dimensions. Some words trigger specific emotions and encourage specific behavior, such as increased empathy, involvement and cooperation [22]. The psychological dimensions in which a hashtag is located, are assumed to influence its spread. The LIWC project [32] assigns words to a number of emotional and cognitive dimensions in various granularities. Among the dimensions are positive sentiment, negative sentiment, physical, social, optimistic, self, anger etc., 69 attributes in total.

### *4.3.2 Global Tweet Features*

Although an ideal meme is self contained, in the Twitter domain, the context in which the hashtag appears might contribute to its acceptance as discourse communities on Twitter (and offline) tend to converge linguistically and stylistically [7]. In order to capture the context in which hashtags appear, we looked on the cognitive dimension of the 1000 most frequent words each tag cooccurred with.

### *Cognitive dimension.*

Similarly to the cognitive dimension for the hashtags, the most frequent words appearing with the hashtag are mapped to the 69 LIWC categories. Thus if a hashtag *ht* tends to appear with the word ‘great’ – the contextual positive sentiment attribute will be 1.

### *4.3.3 Graph Topology Features*

It is clear that graph topology plays an important role in the spread patterns [18, 19, 23, 4, 5, 13, 29]. Since we are mainly interested in how the content of an idea affects its acceptance, we only use basic topological features with the strongest predictive power.

### *Average number of followers.*

We divide the average number of followers of users who used the hashtag to 19 bins on a sub logarithmic scale.

### *Max number of followers.*

Although a user with many followers does not necessarily have influence proportional to the number of his/her followers [4], the number of followers could be a crude estimator to the influential power of the user. We divide the max number of followers observed for users who used the hashtag to 19 bins on a logarithmic scale.

### *Retweets ratio.*

This attribute captures the tendency of a hashtag to appear in retweeted messages. We are interested in this feature in order to create dependency with the orthographic feature set as retweeting is a simple operation that does not require extensive use of the shift key in case the hashtag has many capital letters. We used 25 bins for different retweet rates from less than 1% of retweets (first bin) to 60%-100% of retweets (last bin).

### *4.3.4 Global Temporal Features*

Different hashtags are characterized by different temporal trends observed in close proximity to peak hours of usage of a hashtag [35]. We are modeling the acceptance of a hashtag in a longer time frame of weeks and month, therefore we model the temporal trend in the first few weeks to the appearance of a hashtag, using it to project about the future trend.

### Temporal features.

We sampled the normalized weekly counts of each hashtag in four time stamps:  $w_i$ ,  $i \in \{t, t + 1, t + 2, t + 6\}$ , where  $t$  is the first week the hashtag occurred in our data, and  $t + j$  is the  $j$ -th week after the first occurrence. The four samples give three lag values, denoted as  $d_{k \in \{1, 2, 3\}}$ .  $d_k$  is the ratio of change from the previous time stamp. These 3-lag values capture the change in the meme usage in the first two weeks and then again after another four weeks, approximating ‘stickiness’ and ‘persistence’.

For each lag value we use 17 bins on a logarithmic scale from a change larger than -200% (decreased usage) to more than 200% change (increased usage). If the  $d$  ratio falls between -5% and 5% we consider the hashtag to be stable in that week.

We call this feature type ‘global’ since it is based on the global count of hashtags occurrences in a given week and not on occurrences induced locally by cascading from specific nodes.

## 5. EXPERIMENTAL SETUP

In our experiments we were trying to learn three aspects in the prediction: (i) what is the attribute combination that yields the best prediction? (ii) what are the strongest attributes and how do they complement each other? (iii) how does the prediction accuracy change given different time frames.

All our experiments were executed in a 10-fold cross validation manner. Performance is measured by the mean square error (MSE), a standard measure for regression tasks [1].

We note that when computationally possible, besides the stochastic gradient descent (SGD) described in Section 4.2, we also applied analytic regression algorithm and Support Vector Machine for regression (SVR) [31] with various kernels and optimization of hyper parameters (both implemented as part of the Weka toolkit [16]). Both SVR and analytic regression outperformed the SGD insignificantly. We backed to the SGD as it is orders of magnitude more efficient on high dimensional vectors and since the regression model produced by the SGD is more interpretable than some kernel models.

### Baseline model.

Our basic model is a regression that optimizes only the intercept  $\beta$ , which always converges to the average, predicting the log-average normalized number of occurrences as a fixed function, keeping the error close to  $\sigma^2$ , the variance of the observed counts.

### Other basic models.

Each of the four feature types in Section 4.3 was used in a different regression model.

### Hybrid models.

Finally, we experimented with different combinations of the four feature types, studying the mutual effect of various feature types.

## 6. RESULTS

Table 1 shows our results for the baseline, the four basic models and the hybrid model. Results are presented for four different horizons. Each of the basic models out-

Model	MSE <sub>10</sub>	MSE <sub>15</sub>	MSE <sub>20</sub>	MSE <sub>25</sub>
baseline	4.988	3.796	3.125	2.698
HT <sub>all</sub>	4.380	3.410	2.902	2.565
TW <sub>content</sub>	4.776	3.509	2.743	2.221
Graph	4.295	3.144	2.404	1.923
Temporal	3.294	2.893	2.507	2.112
Hybrid <sub>all</sub>	2.584	2.098	1.685	1.315

Table 1: MSE of basic models and the hybrid model in horizons.  $MSE_n$  indicates results for acceptance prediction in an  $n$  weeks time frame.

Time frame	weeks <sub>10</sub>	weeks <sub>15</sub>	weeks <sub>20</sub>	weeks <sub>25</sub>
Sample size	21683	19020	15635	10864
Mean avg.	4.342	5.005	5.513	5.891
STD	4.986	3.794	3.125	2.699

Table 2: Sample size, average log of normalized counts and standard deviation of ‘fresh’ hashtags in different time frames.

performs the baseline, while the global models (graph and temporal) outperforms the content based models. The hybrid model, combining all feature types is significantly better than all partial models. All models achieve better prediction as prediction horizon grows. This can be attributed to the decreased standard deviation of the counts as shown in Table 2 and in Figures 4,5,6 and 7.

Table 3 presents the contribution of different combinations of feature types, demonstrating that all feature types contribute to the prediction and no feature type is masked under a strong signal produced by other feature types. Table 3 also presents the correlation coefficient between the predicted values and the real values. The baseline, always optimized so the MSE is close to  $\sigma^2$ , presents no correlation with the target values as it is a fixed function. A correlation of 0.669, achieved in the all-inclusive hybrid model, shows that our linear model serves as a good approximation to the actual model. We also note that even the 0.268 correlation achieved by the  $HT_{all}$  is reasonable in real-world noisy data.

It is interesting to note that even though the content attributes of the hashtag and the tweet context show only a small improvement over the baseline with MSE of 3.41 and 3.509, combining both content types presents a much smaller error with MSE of 2.967.

## 7. GENERAL DISCUSSION

Model	MSE	Corr-coeff
baseline	3.796	-0.021
Ht <sub>all</sub>	3.410	0.319
TW <sub>cont</sub>	3.509	0.275
Graph	3.144	0.414
Temporal	2.893	0.487
HT <sub>cont</sub> + TW <sub>cont</sub>	2.967	0.467
HT <sub>cont</sub> + TW <sub>cont</sub> + Graph	2.546	0.573
HT <sub>cont</sub> + TW <sub>cont</sub> + Temp	2.321	0.6234
Graph+Temporal	2.450	0.594
Hybrid <sub>all</sub>	2.098	0.669

Table 3: MSE and correlation coefficient for various combinations of feature types for a 15 weeks time frame.

Features	MSE	corr-coeff
baseline	3.796	-0.021
number of words	3.695	0.162
length chars	3.693	0.163
orthography	3.781	0.061
lexicons	3.667	0.183
lc+nw	3.662	0.187
collocation	3.651	0.195
cognitive	3.637	0.204
location	3.613	0.218
lc+nw+ortho	3.654	0.192
lc+nw+cog	3.637	0.204
lc+nw+lex	3.631	0.207
Ht <sub>all</sub>	3.410	0.319
TW <sub>cont</sub>	3.509	0.275
HT <sub>all</sub> + TW <sub>cont</sub>	2.967	0.467

Table 4: MSE and correlation coefficient of each single content feature set and combinations of content features for a 15 weeks time frame.

Model	MSE	Corr-coeff
d <sub>1</sub>	3.236	0.383
d <sub>2</sub>	3.39	0.326
d <sub>3</sub>	3.44	0.303
d <sub>1</sub> + d <sub>2</sub>	3.088	0.431
d <sub>1</sub> + d <sub>3</sub>	2.97	0.464
d <sub>2</sub> + d <sub>3</sub>	3.19	0.398
d <sub>1</sub> + d <sub>2</sub> + d <sub>3</sub>	2.893	0.487

Table 5: MSE and correlation coefficient for different number of lags and different distances between sampling points in 15 weeks horizon.  $d_i$  indicates the the  $i$ -th lag described in Section 4.3.4.

In this section we discuss some of the results in greater detail.

### Temporal based prediction.

Our temporal feature set (Section 4.3.4) is based on lags computed from samples in four time stamps, assuming that the lags portrait the tendency of a hashtag to stick around and propagate. Table 3 shows that hybrid models that include temporal features always outperform other models for a 15 weeks horizon. Table 1 shows that the temporal models performs better than other feature types in 10 and 15 weeks horizon while the graph-based model performs better in 20 and 25 weeks horizon.

Looking at our data, 4766 hashtags have a single spike. While our temporal features correctly model single-spiked hashtags, they fail to model hashtags with multiple spikes (see Figure 8), thus do not perform as well on longer horizons and hashtags with reoccurring spikes. An algorithm for clustering and predicting patterns of short temporal variations was proposed by [34, 35]. Table 5 illustrates the MSE for different numbers and distances of lags.

A close examination of spiked hashtags show that these are usually related to Twitter games and idioms that flare for an instant then quickly vanish, sometimes reoccurring for another playful round. Twitter games can be sometimes captured by our content features. An analysis of the information diffusion of (the most popular) Twitter idioms was done in [29].

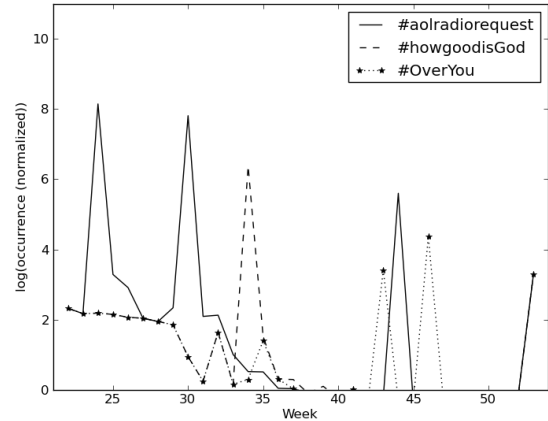


Figure 8: Hashtags with one, two and three spikes.

### The effect of cognitive attributes.

Table 4 presents various combinations of attributes from the content type. While the baseline’s MSE is 3.796, using all hashtag attributes (listed in Section 4.3) achieves MSE of 3.41, each attribute having a small contribution to the model. One of our hypotheses was that the hashtag’s cognitive attributes have a great effect on the hashtag’s popularity. While this hypothesis does not hold, the LIWC cognitive categories had a positive impact when applied to the context of the tweet ( $TW_{cont}$ ), still supporting the intuition that the use of certain words drives users to adopt a hashtag or get involved in a certain discourse (defined by specific hashtags). We attribute the marginal contribution of the cognitive attributes of the hashtag to its ‘sparseness’, as these attributes were useful in the richer context.

These findings also validate sociolinguistic theories about the correlation between the use of certain words and personality traits such as the image of self and on team work [10, 30, 22, 17, 32].

### Dependencies and collinearity.

In order to better understand the dependencies between attributes and feature types, we trained models with dummy attributes. Dummy attributes are second order attributes created by a Cartesian product over the feature space. Since all features are binary, a new feature type  $x_{ij}$  was added:  $x_{ij} = 1$  iff  $x_i = 1 \wedge x_j = 1$ , where  $x_i$  and  $x_j$  are attributes in the standard model. For example, while in the standard (hybrid) model we have two different attributes *containsCaps* (some, but not all, letters in the hashtag are capitalized) and *rtRate<sub>0.4-0.6</sub>* (the hashtag has retweet rate of 40%–60% of its appearances); a dummy attribute in our new model is *containsCaps + rtRate<sub>0.4-0.6</sub>* which equals 1 iff both attributes had 1 in the hashtag’s vector (e.g. the hashtag *#BoyFriends*). The dimension of our Cartesian models is very large (thousands) compared to the size of the sample (see Table 2), thus overfitted. While overfitted, analysis of model coefficients captures the relations between different variables.



Attribute	Coefficient
$chars_{long} + rtRate_{0.4-0.6}$	0.7677
$chars_5 + rtRate_{0.4-0.6}$	0.5718
$chars_3 + rtRate_{0.4-0.6}$	0.3543
$chars_2 + rtRate_{0.4-0.6}$	0.3469
$chars_{6-9} + rtRate_{0.4-0.6}$	0.3417
$chars_{10-14} + rtRate_{0.4-0.6}$	0.3060
$chars_4 + rtRate_{0.4-0.6}$	0.1266
$words_1 + rtRate_{0.4-0.6}$	0.6086
$words_2 + rtRate_{0.4-0.6}$	0.4224
$words_3 + rtRate_{0.4-0.6}$	0.2581
$words_{long} + rtRate_{0.4-0.6}$	0.0260
$containscaps + rtRate_{0.4-0.6}$	0.3131
$nocaps + rtRate_{0.4-0.6}$	0.002
$allcaps + rtRate_{0.4-0.6}$	-0.3437

**Table 6: model coefficients dummy attributes of hashtag length and orthography combined with high retweet rates (40%-60% of the message containing these hashtags are retweeted),  $chars_{long}$  indicated hashtags longer than 14 characters,  $words_{long}$  indicates hashtags that contain more than 3 distinct words.**

### Cognitive load and physical constraints.

Choosing attributes for our model we had several hypotheses about the the acceptance of hashtags. One such hypothesis is inspired by the optimality theory framework. A successful hashtag should be clear, informative and yet not too complex. The complexity of a hashtag can be measured by its length (in chars), the number of words it consists of and its orthographic features. Longer hashtags are harder to type, they are not economical unless very informative (due to the Twitter 140 characters policy) and they can be harder to interpret (a sequence of words with no white spaces). All these aspects can be broken down if one takes retweeting into account, as a retweeted message doesn't require retyping a complex hashtag. Table 6 illustrates the differences in the coefficients when the typing constraint is accounted for by looking only on hashtags with high retweet rate.

Examining the coefficients of the dummy attributes representing the hashtag length (chars and number of words) with the retweet rates, we observe an interesting phenomena. The coefficients show contradictory effects of the hashtag number of words and its character length. It seems that while users do not care so much about longer hashtags, they still prefer hashtags that are less complex. It might be because these are harder to remember or interpret (compare *#savethenhs* vs. *#technology*, both 10 characters long). While a larger number of words seems to be a burden, we observe that once capital letters are introduced to mark word boundaries, the coefficient of the dummy attribute that captures the dependency between number of words and retweets grows (compare *#savethenhs* vs. *#saveTheNHS*).

## 8. CONCLUSIONS

Predicting the spread of ideas in online communities is an interesting task from both commercial and psychological perspectives. Traditional approaches model the propagation of ideas in social media by analyzing the topology of the social graph. In this work we took a hybrid approach to predicting the spread of ideas according to their content as well as the topology of the social graph. We viewed Twitter hashtags as ideas and as potential memes and learned a regres-

sion model that predicts the spread of each hashtag/meme in a time frame. Our experiments demonstrated that the content of the idea plays an important role in its acceptance by the community. We demonstrated that an elegant hybrid approach, combining the meme's content, the meme's context, global temporal features and graph topology shows the best results, while maintaining computational efficiency. However, some of the feature types we used require more data that is not always available. In contrast, using the meme's content alone is free of the need of any global information but the meme itself. Our experiments also demonstrate some psychological and cognitive-linguistic principles on a large scale data.

There are many directions for future research. One of the main directions is to gain a better understanding of the mutual roles the different feature types play. Another direction is to better characterize psychological aspects and cognitive constraints and the ways they interact.

## 9. ACKNOWLEDGMENTS

We thank Gideon Dror and Idan Szpektor for valuable discussions and advice about this research.

## 10. REFERENCES

- [1] J. Bibby and H. Toutenburg. *Prediction and Improved Estimation in Linear Models*. John Wiley & Sons, Inc., New York, NY, USA, 1 edition, 1978.
- [2] L. Bottou. Stochastic learning. In O. Bousquet and U. von Luxburg, editors, *Advanced Lectures on Machine Learning*, Lecture Notes in Artificial Intelligence, LNAI 3176, pages 146–168. Springer Verlag, Berlin, 2004.
- [3] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on Twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, pages 1–10. ACM, 2010.
- [4] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Washington DC, USA, 2010.
- [5] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1029–1038. ACM, 2010.
- [6] E. Cunha, G. Magno, G. Comarela, V. Almeida, M. Gonçalves, and F. Benevenuto. Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. *ACL HLT 2011*, page 58, 2011.
- [7] C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754. ACM, 2011.
- [8] D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on*

- Computational Linguistics: Posters*, pages 241–249. Association for Computational Linguistics, 2010.
- [9] D. Davidov, O. Tsur, and A. Rappoport. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116. Association for Computational Linguistics, 2010.
- [10] D. Davis and T. Brock. Use of first person pronouns as a function of increased objective self-awareness and performance feedback. *Journal of Experimental Social Psychology*, 11(4):381–388, 1975.
- [11] N. R. Draper and H. Smith. *Applied Regression Analysis (Wiley Series in Probability and Statistics)*. Wiley-Interscience, third edition, April 1998.
- [12] M. Götz, J. Leskovec, M. McGlohon, and C. Faloutsos. Modeling blog dynamics. In *International Conference on Weblogs and Social Media*, 2009.
- [13] A. Goyal, F. Bonchi, and L. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM, 2010.
- [14] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 491–501, New York, NY, USA, 2004. ACM.
- [15] L. Guo, E. Tan, S. Chen, X. Zhang, and Y. Zhao. Analyzing patterns of user content generation in online social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 369–378. ACM, 2009.
- [16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.
- [17] J. Kahn, R. Tobin, A. Massey, and J. Anderson. Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American journal of psychology*, pages 263–286, 2007.
- [18] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.
- [19] D. Kempe, J. Kleinberg, and É. Tardos. Influential nodes in a diffusion model for social networks. *Automata, Languages and Programming*, pages 1127–1138, 2005.
- [20] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.
- [21] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the nelder-mead simplex algorithm in low dimensions. *SIAM Journal of Optimization*, 9:112–147, 1996.
- [22] G. Leshed, J. Hancock, D. Cosley, P. McLeod, and G. Gay. Feedback for guiding reflection on teamwork practices. In *Proceedings of the 2007 international ACM conference on Supporting group work*, pages 217–220. ACM, 2007.
- [23] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1, May 2007.
- [24] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. Citeseer, 2009.
- [25] J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, January 1965.
- [26] B. O Connor, R. Balasubramanyan, B. Routledge, and N. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 122–129, 2010.
- [27] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *International AAAI Conference on Weblogs and Social Media*. The AAAI Press, 2010.
- [28] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.
- [29] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 13th international conference on World Wide Web*, WWW '11, 2011.
- [30] S. Rude, E. Gortner, and J. Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18(8):1121–1133, 2004.
- [31] A. Smola and B. Scholkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.
- [32] Y. R. Tausczik and J. W. Pennebaker. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54, Mar. 2010.
- [33] M. Wright. Direct search methods: Once scorned, now respectable. In D. Griffiths and G. Watson, editors, *Numerical Analysis*, pages 191–208. Addison Wesley, Redwood City, 1995.
- [34] J. Yang and J. Leskovec. Modeling Information Diffusion in Implicit Networks. In *2010 IEEE International Conference on Data Mining*, pages 599–608. IEEE, 2010.
- [35] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM, 2011.
- [36] T. Zaman, R. Herbrich, J. Van Gael, and D. Stern. Predicting information spreading in twitter. In *Workshop on Computational Social Science and the Wisdom of Crowds, NIPS*, 2010.