Competence Modeling in Twitter: Mapping Theory to Practice

John O’Donovan
Dept. of Computer Science,
University of California, Santa Barbara
Email: jod@cs.ucsb.edu

Byungkyu Kang
Dept. of Computer Science,
University of California, Santa Barbara
Email: bkang@cs.ucsb.edu

Tobias Höllerer
Dept. of Computer Science,
University of California, Santa Barbara
Email: holl@cs.ucsb.edu

ABSTRACT

Availability of “big data” from the Social Web provides a unique opportunity for synergy between the computational and social sciences. On one hand, psychologists and social scientists have developed and established models of human competence, credibility, trust and skill over many years. Currently, much research is being conducted by computer scientists to evaluate these human-behavioral aspects using real-world data from Twitter and other sources. However, many of these algorithms are formulated in an ad-hoc way, without much reference to established theory from the existing literature. This paper presents a framework for mapping existing models of human competence and skill onto a real world streaming data from a social network. An example mapping is described using the Dreyfus model of skill acquisition, and an analysis and discussion of resulting feature distributions is presented on four topic-specific data collections from Twitter, including one on the 2014 Winter Olympics in Sochi, Russia. The mapping is evaluated using human assessments of competence through a crowd sourced study of 150 participants.

I INTRODUCTION

At the beginning of 2014 the Twitter social network generated over 9,000 new messages every second, [1]. The volume and geographic diversity of these messages easily establish Twitter as a major information channel for news, media and conversation. It is a well known fact that where user-generated content exists, there is always a large amount of noisy or otherwise useless data. A key challenge to harnessing Twitter as an information source, is the ability to find relevant, reliable and trustworthy users to follow. Computer scientists in the fields of search and information retrieval (e.g.: recommender systems) have attempted to address this problem in other domains for several decades [2], while Behavioral scientists (Psychologists, Cognitive scientists, Social scientists) have studied the concepts of trust, reliability and competence for a far longer period of time, and have developed established theory for identifying and classifying these characteristics, both at the human level and the information level. [3,4] While many studies of Twitter in the computer science literature attempt to model and mine for these characteristics [5–7], their models and algorithms tend to be formulated in an ad-hoc manner, without strong grounding in established theory from the human behavioral sciences.

This paper describes an experimental framework to map and validate established models of human behavior with the Twitter network and the information that flows within it. If applied successfully, such a framework has three clear benefits. First, it can serve as a form of validation for existing theoretical models by applying them at scales that were previously unattainable. Second, it can help analysts to constructively reason about observed phenomenon in the real world data. Finally, it can be used to improve design of search and recommendation applications that attempt to relieve the information overload problem.

Mapping of complex theoretical models of human behavior to observed behaviors in Twitter is clearly not a trivial task. The examples shown in the following sections all require a level of interpretation and a common sense reasoning about the links between factors in the model, and features and indicators in the Twitter information network. For the purpose of generalization we highlight the following steps for integrating an arbitrary human behavioral model with the network and associated data from Twitter API, and follow this with an example implementation of the general process.

- Task Identification and Analysis What are the information requirements? What data elements from Twitter API can provide insight?
- Model Selection Is there a model in the behavioral/social science literature that is relevant to the task?
- Feature Selection What are the best features in the social network that may be useful indicators to the model?
• **Interpretation and Mapping** How should the features be related to the model itself

• **Model Building and Validation** Train a prediction model using the mapped feature set and validate against a test set of annotations, or other available ground truth data.

In the remainder of this paper we detail the above mapping procedure using an example task and an established theoretical model over four large current event data sets crawled from Twitter. Since identification of reliable information is such a critical aspect of today’s social web, we have chosen the following as an example task: can we predict that a Twitter user will provide information about a target topic in a competent way. Since Twitter is still a relatively young platform, and many users are still unfamiliar with the full scope of its operation and use, we have borrowed a model of competence from educational psychology known as the “Dreyfus Model of Skill Acquisition” [3] as a working example that to our knowledge has not previously been applied to social web data.

II RELATED WORK

In this section, we introduce the state-of-the-art techniques from literature that identify unique features for social media analytics and building models to predict various facets of human behavior. Twitter has a unique combination of text content and underlying social link structure, in addition to a variety of dynamic or ad-hoc structures, making it ideal for the study of information credibility and competence of an information provider. Common methods for data mining in Twitter can be loosely classified by the type of data that they operate on.

• **Content-based Methods** generally rely on the text and other metadata in a message to make assertions about information or users. For example, trust, credibility, competence of the author etc. These methods can be quite scalable, since they require only a single API query per assertion. Examples include Canini et al. [8] Kang et al. [9] and Castillo et al. [10]

• **Network-based Methods** generally rely on analysis of the underlying network structure to make decisions about information quality. Examples include Zamal et al. [11]. Network based methods can be slower and less scalable since they potentially require many API queries to make assertions about a single user or message. Dynamic network analysis methods, such as retweet analysis can be even more computationally expensive and less scalable, since they focus on information flowing through an already complex network.

• **Hybrid Methods** combine facets from content and network-based approaches. Examples include Sikdar et al. [12], O’Donovan et al. [7] and Kang et al. [9].

Canini et al. [8] present a good example of content-based analysis of messages in Twitter, they concentrate on modeling topic-specific credibility, defining a ranking strategy for users based on their relevance and expertise within a target topic, using Latent Dirichlet Analysis. Based on user evaluations they conclude that there is “a great potential for automatically identifying and ranking credible users for any given topic”. Canini et al. also evaluate the effect of context variance on perceived credibility.

Twitter has been studied extensively from a media perspective as a news distribution mechanism, both for regular news and for emergency situations such as natural disasters, and other high-impact situations [5, 13, 14]. For example, Thomson et al. [14] model the credibility of different tweet sources during the Fukushima Daiichi nuclear disaster in Japan. They found that proximity to the crisis seemed to moderate an increased tendency to share information from highly credible sources, which is further evidence for our earlier argument that credibility models in Twitter need to account for and adapt to changes in context. Castillo et. al. [5] describe a study of information credibility, with a particular focus on news content, which they define as a statistically mined topic based on word co-occurrence from crawled “bursts” (short peaks in tweeting about specific topics). They define a complex set of features over messages, topics, propagation and users, which trained a classifier that predicted at the 70-80% level for precision/recall against manually labeled credibility data. While the three models presented in this paper differ, our evaluation mechanism is similar to that in [5], and we add a brief comparison of findings in our result analysis. Mendoza et. al [13] also evaluate trust in news dissemination on Twitter, focusing on the Chilean earthquake of 2010. They statistically evaluate data from the emergency situation and show that rumors can be successfully detected using aggregate analysis of Tweets.
Table 1: Common demographic attributes used in Twitter mining algorithms.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Feature</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>language use (stylistic features:</td>
<td>traditional text [15,16], blog [17], email [18], user search query</td>
</tr>
<tr>
<td></td>
<td>pronouns, determiners, prepositions,</td>
<td>[19, 20], review [21], Twitter [11, 22], Facebook [23]</td>
</tr>
<tr>
<td></td>
<td>quantifiers, conjunctions, etc.)</td>
<td></td>
</tr>
<tr>
<td>message location</td>
<td>message/web content, search query,</td>
<td>[19,24,25]</td>
</tr>
<tr>
<td>regional origin</td>
<td>message text, user behavior, network</td>
<td>[22]</td>
</tr>
<tr>
<td></td>
<td>structure</td>
<td></td>
</tr>
<tr>
<td>profile age</td>
<td>search query, profile description</td>
<td>[11,19,22]</td>
</tr>
<tr>
<td>political</td>
<td>message text</td>
<td>[11,22,26]</td>
</tr>
<tr>
<td>orientation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

While identification of indicators of human-behavioral features such as competence and credibility is an important task, it is also important to consider the end-user’s perception of them. Morris et al. [27] performed a study to address users perceptions of the credibility of individual tweets in a variety of contexts, for example, from socially connected and unconnected sources, e.g., in blogs [17], email [18] and search [19,20]. From the results, Morris et al. derive a set of design recommendations for the visual representation of social search results.

Demographics play an important role in understanding information quality in Twitter. Table 1 presents an overview of key user-based attributes that researchers tend to rely on. In this table, attributes are shown on the left, example features for each are shown in the middle column, and the research papers that employ the features/attributes are given in the right column. For example, [28] conducted a simple survey on the application of features which can be used for analyzing people’s profiles on the style, patterns and content of their communication streams. Herring [15] investigate the language/gender/genre relationship in web blogs and show gender-related stylistic features from diary and filter entries. Incorporating occurrence of words and special characters based on pre-defined corpora is another type of feature selection. For example, [29] use simple nominal or binary binary features to classify tweets into different categories such as news, temporal events, opinions, deals or conversations. [24] propose a probabilistic framework for content-based location estimation using microblog messages. The framework estimates each user’s city-level location based purely on the message text without any geospatial coordinates, while [22] apply stacked-SVM-based classification algorithms for their classification task on a Twitter dataset. Since we are interested in creating mappings between existing models of human behavior and the Twitter network, understanding these different features, methods and their performances is a critical first-step.

III SETUP AND DATA COLLECTION

In this section, we will describe the experimental setup for our evaluation, particularly the crawling process and the collected data. Table 2 shows a summary of all data used in our evaluation, and Figure 1 shows an overview of the crawling process for users and topics. The larger circle denotes a set of messages gathered during a retroactive crawl using keywords that emerged after a period of time had elapsed since the initial crawl, but were still deemed to be a part of the core topic.

Figure 1: Overview of the crawled set of users and topics. Set $S_{seed}$ represents the initial seed crawl from a key hashtag. Set $S$ represents an expanded topic crawl to incorporate additional hashtags that evolve over the course of the event. Set $u$ represents the set of all tweets from users who exist in $S$.

1 DATA COLLECTION

To allow for comparison of feature and model behavior, three different data sets are used in our evaluation. The first data collection is centered around the 2014 winter olympic games in Sochi, Russia. Data was crawled for approximately three weeks using a variety of keywords shown in Table 2. Sochi was chosen as a potentially interesting data set because of the diversity of cultures involved, and because of the associated excitement, politics and availability of concrete ground truth data in the form of event results.
Table 2: Overview of three data collections used to evaluate the mapping framework.

<table>
<thead>
<tr>
<th>Collection</th>
<th># Users</th>
<th># Msgs</th>
<th>Keywords</th>
<th>Hashtags</th>
<th>Example tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>boston</td>
<td>357,152</td>
<td>460,945</td>
<td>marathon, pray, suspect, victims, bomb, police, hit, shards, doctor, pellet, running, die, affected, rip, explosion, swat, blood, bombings, FBI, tragedy, donate, watertown, arrest, kill, injured, runner, hurt, donors, dead, identified</td>
<td>#bostonmarathon, #prayforboston, #boston, #prayersforboston, #water-town, #bruins</td>
<td>RT @Channel4News: There have been no arrests made yet after the bombings at the #BostonMarathon - US sources. #c4news</td>
</tr>
<tr>
<td>bostonstrong</td>
<td>62,461</td>
<td>120,442</td>
<td>affected, bostonishack, bostonstrong, boybston, charitymiles, donate, FBI, flyers, fund, help, honor, hope, marathon, memorial, oneboston, onefundboston, police, silence, spell, strength, strong, support, donors, tribute, victims, blood, bomb, doctor, tragedy, dead, rip, pray, hurt</td>
<td>#bostonstrong, #oneboston, #copley, #bostonishack, #prayforboston</td>
<td>@Nicolette_O Thank you for your support of the original #BOSTONSTRONG campaign, Nicolette! Nearing $400K raised for The One Fund Boston! xxx</td>
</tr>
<tr>
<td>sochi</td>
<td>4,305,508</td>
<td>9,521,089</td>
<td>sochi, olympic, winter, female-olympians, games, gold, team, russia, hockey, medal, opening, usa, athletes, figure, canada, win, men’s, ceremony, skating, ice, stray, putin, women’s, gay, sport, won, ski, live, slope, skater, world</td>
<td>#sochi, #olympics, #sochi2014, #sochiproblems, #wearewinter, #sougofollow, #olympics2014</td>
<td>RT @Bobby_Brown1: In air shot on the #Olympic slope course. Jumps are huge. Gonna be fun <a href="http://t.co/XCQz90k1Eb">http://t.co/XCQz90k1Eb</a></td>
</tr>
</tbody>
</table>

Figure 2: Overview of the Dreyfus model of skill acquisition. A component mental function is represented on each row and associated skill levels are shown on the columns. The horizontal arrows on each row represent the change in an observed mental function that facilitates an increase in the skill level represented in the model.

Our second and third data sets are related to the terrorist attack that occurred during the 2013 Boston Marathon. The larger of the two collections was collected about the event itself, using the popular hashtag “#boston”. In this case, the data crawling began an hour after the event occurred and continued for two weeks. The second data collection was about the aftermath and recovery movement, crawled using the keyword “#bostonstrong” This was also crawled for approximately two weeks.

2 THEORETICAL FOUNDATION

To exemplify the mapping process, we have chosen to borrow a model from the field of educational psychology known as “the Dreyfus model of skill acquisition” [3]. Since Twitter is a relatively new phenomenon, many of its users are still learning about the complex information, information flow, and network structure that Twitter supports, so we deemed this competence-based model of skill acquisition to be a reasonable example. Ideally, the generalizable framework we are describing will support many other established models of credibility, competence, trust or
other factors that influence human decision-making, provided that appropriate mapping steps can be performed.

2.1 DREYFUS MODEL OF SKILL ACQUISITION

The Dreyfus model of skill acquisition describes the process of human skill acquisition in 5 different levels. This model was first introduced by the brothers Stuart and Hubert Dreyfus [3], and is established in the fields of education and operations research. The model is based on the four different transitions that define boundaries between five binary states of mental function during human learning. The original model, as can be seen in Table 3, is based on the three scenarios that show progression of a through each of the transitions, respectively. Table 3 suggests one of many possible mappings to a set of observable features in the Twitter based on expert interpretation of both.

### 3 MAPPING

Now that we have selected a model, the next step is to study the meaning of each component within it, and formulate a reasonable analog in the behavior of an available set of Twitter features. A discussion of all such features is not possible here. The feature sets that we consider are discussed in Sikdar et. al [12], especially in Tables I and II of [12].

First it is necessary to define the network, topic, users and associated features more concretely: For the following discussion, we view the Twitter domain as a triple $(S, U, T)$, where $S = (s_1, s_2, ..., s_n)$ is a set of tweets crawled about a target topic. $U$ is the set of users ($u_1, u_2, ..., u_m$) who have at least one tweet in $S$. Additionally we define $T$ a vector of event timestamps representing when messages in $S$ were posted. This is given by $T = (t_1, t_2, ..., t_n)$. Furthermore, each topic $S$ can be represented by its component hashtags, $S_{hash} = (h_1, h_2, ..., h_n)$. A notable property of $S_{hash}$ is that the vector emerges over the values in $T$. Last, we define $S_{seed}$ as the subset of $S$, gathered from the earliest emergent hashtags in $S_{hash}$.

Importantly, the mapping procedure we discuss here is simply an example to demonstrate the process. Mappings between a complex network and a complex behavioral model obviously require a degree of manual interpretation. Figure 2 illustrates a general form of the Dreyfus model, highlighting four key mental functions and the related competence levels. Table 3 shows the mental function on the leftmost column, followed by the associated indicators of competence or non-competence. The third row is the critical component, showing the analog feature combinations in Twitter. This is followed by other notable analogs and a text description of each feature. Our approach first looks at behavioral features in Twitter that could potentially serve as an indicator of each state. First we will describe the reasoning behind each mapping, and in the following section we present an evaluation of the behavior of each mapped feature, further indicating its potential to measure competence.

To recap, we are interested in evaluating the competence of information providers in Twitter with respect to a target topic. This covers both authorship and information propagation alike. Within this context, we interpret recollection in a topic as the ability to think back into the topic history, in the sense of maximizing ones posterity in the target topic. To approach this computationally, we consider the sequence of event times $T$ of topic $S$ from our earlier definition, and attempt to gauge where individual users reside with respect to the normal for the topic. For example, if Alice’s history goes farther back than Bob’s, she has a greater degree of posterity, and perhaps this can be an indicator of competence. We compute this for every user simply as the earliest timestamp of a tweet that they have made in topic $S$. This is com-

<table>
<thead>
<tr>
<th>Function</th>
<th>Non-competent State</th>
<th>Competent State</th>
<th>Corresponding features</th>
<th>Other possibilities for features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recollection</td>
<td>Non-situational</td>
<td>Situational</td>
<td>$S([u, t_0]) \rightarrow \text{avg}(S)$</td>
<td>specific #ht $\leftrightarrow$ non-specific. writing of content</td>
<td>Adaptation to context (time specific)</td>
</tr>
<tr>
<td>Recognition</td>
<td>Decomposed</td>
<td>Holistic</td>
<td>Fraction of $T$ that is in $u$</td>
<td>-</td>
<td>Coverage of topic $T$ by user $u$</td>
</tr>
<tr>
<td>Decision</td>
<td>Analytical</td>
<td>Intuitive</td>
<td>$\text{Opin}(u, T)/U_{\text{Opin}}$</td>
<td>-</td>
<td>Opinion and Sentiment of $u$ in $T$</td>
</tr>
<tr>
<td>Awareness</td>
<td>Monitoring</td>
<td>Absorbed</td>
<td>Fraction of $u$ that is in $t$</td>
<td>-</td>
<td>Involvement/Immersion in a topic $T$</td>
</tr>
</tbody>
</table>

Table 3: Interpreted mappings between the Dreyfus model and a set of Twitter features
pared against the average timestamp of all users’ first
tweets \((s_0)\) within the topic, as shown in Equation 1
below. In a perfect mapping, we could simply ex-
amine the distribution graph of this feature over all
users and segment it using a threshold value to deter-
mine the boundary between the competent and non-
competent state. In this case, the boundary between
non-situational (general) and situational (specific, de-
tailed) recollection. The following section describes
evaluations of this type for all features on all three
data collections.

\[
\text{recollection}(u, S) = T[(s_0, u)] - \frac{\sum_{i=1}^{n} T[(s_0, u_i)]}{n} \quad (1)
\]

The next function of the Dreyfus model in Table 3
is “recognition”. Assessing whether a human’s recog-
nition of a topic is in a decomposed or holistic state
is very difficult, depending on the complexity of
the topic being analyzed. For our simple compu-
tational model, we treat recognition of a topic \(S\) by
user \(u\) as the degree of coverage of \(S\) by \(u\). This
could be simply computed as the sum of all messages in \(u\)
that are related to \(S\), divided by the total number of
messages in \(S\). However, sparsity, irrelevant messages
and other noise in the topic can weaken the link to
the user profile. A better way to approach this map-
ing could leverage a) the set of hashtags in \(S_{\text{hash}}\)
that describe the topic, or b) the set most frequently
occurring terms as a more well-defined descriptor of
the topic. We compute the hashtag-based coverage
as Equation 3 below.

\[
\text{recognition}(u, S) = \frac{S_{\text{hash}}(u)}{S_{\text{hash}}(\text{all})} \quad (2)
\]

The “decision” function in the Dreyfus model is
treated differently in our mapping. Dreyfus catego-
rizes this into analytical decision-making and intuitive
decision-making, with the latter being an indi-
cator of expertise within the topic (see Figure 2).
Deciding whether an individual is making analytical
or intuitive choices has been the subject of many re-
search papers in itself, e.g. [30], so again, we will need
to simplify here for the purposes of discussion. Our
computational model looks to sentiment as an indi-
cator of decision making potential. This approach
has been studied and validated by many researchers.
For example, O’Connor et al [31] found that decisions
to purchase products (consumer confidence) and de-
cisions about elections [31, 32] can be predicted by
examining frequency of sentiment-related word usage
in Twitter posts.

In particular, we examine three aspects of sentiment:

- **Degree of Subjectivity** If a user demonstrates
  the ability to form subjective opinion on a given
  topic, it *may* point towards a higher level of
  competence. To assess this, we borrow a sub-
  jectivity lexicon from the Opinion Finder tool
  described by Wilson et al. in [33]. Each user \(u\)
is represented as a bag of terms and a count is
performed for terms that occur in the lexicon.
The resulting value is our subjectivity score for
that user. At a finer grained level, we focus
on words that imply personal preference (e.g.
cool, excellent, awesome, etc.), and on expres-
sions / idioms that imply opinion (e.g. I think,
I suppose, I believe etc.).

- **Sentiment Intensity** Intensity of sentiment is a
  good indicator of knowledge about a topic [31].
  In our model, this is measured as a simple count
  against the sentiment lexicon from NLTK [34].

- **Sentiment Polarity** Our third sentiment met-
  ric examines sentiment of user \(u\) as a polar-
  ized scalar \(sp = [-1 1]\) by comparison against
  negative and positive sentiment lexicons from
  NLTK.

While the Dreyfus model from Figure 2 shows a single
factor for “Decision”, we choose to analyze the three
sentiment factors separately in the analysis that fol-
low, in case varying behaviors can be observed. Af-
ter the initial feature behavior analysis they can be
pruned or combined in some way to produce a single
attribute.

The final function listed in Table 3 is the concept of
awareness. According to the model shown in Fig-
ure 2, when a human’s awareness transitions from
persistent monitoring to an absorbed level, it is an
indication of mastery of a particular skill. Put an-
other way, this transition occurs when actions become
“second nature” instead of as a result of careful fine-
grained analysis of rules and inputs. Again, this is
a potentially difficult concept to map onto a simple
computational model, since one essentially needs to
be at the mastery level in a given topic to recognize
such intuitive actions. In this example, our goal is
to evaluate competence of an information provider
in a target topic. As a simple proxy for detecting
the transition in awareness between monitoring and
absorbed, our computational model focuses on the degree of *immersion* of a user in a topic. That is, the percentage of the user’s profile that is dedicated to a topic $S$. One problem with this proxy is that it does not facilitate fair comparison between users—a property that is required for the feature behavior analysis that follows. Consider our Sochi Olympics dataset for example: If the official winter olympic feed has 1,000 tweets all about the event, and a random user (Joe) has 10 tweets that are also about the event, this metric would produce the same score for both profiles. To control for this, we introduce a weight $w$ based on the number of tweets in the profile, shown here as Equation 3:

$$
\text{awareness}(u, S) = \frac{u_{(\text{hash})}}{u_{(\text{all})}} \times w. \quad (3)
$$

This concludes the interpretation and mapping phase of the framework. Now, we arrive at a computational model in the form of a set of observable features that maps, albeit loosely, to the theoretical model in Figure 2. The next step in the procedure is to evaluate the behavior of these features to determine distribution curves and see if we can identify reasonable thresholds that can correspond with the phase transitions of the Dreyfus model, shown in Figure 2.

### 4 FEATURE ANALYSIS

Now that we have described the computational model we must assess its potential to predict human behavior in real world Twitter data. To achieve this we compute the 6 individual features described in the previous section on each of the three data collections (Boston, BostonStrong and Sochi). All of the features described can be considered user-based features, that is, they are attached to a single user, as opposed to a single message (see [7, 9, 12]) for a discussion on user and message-based features). In order to examine potential of a feature for predicting competence of a user as a provider of information about a topic, we take the following approach: First we compute the individual feature value $f \in F$ for each user $u \in U$ on each data set $S$. Next we plot a distribution $\text{dist}(f, U, S)$ for all features in $F$ and all three of our topics. Results of this analysis are shown in Figure 3, and arranged as follows: each row represents a computed feature, identified by the title on the left side. Each column represents a data collection, identified by the seed hashtag in the header row. This arrangement of distributions is useful since allows us to quickly compare across data collections and across features. All values are shown in percentages with the exception of the first row, which is a time-based value (seconds).

Let us first discuss the behavior of individual features, with a view to locating thresholds that may yield information about competence of users as information providers about the topic. The recollection feature shows distribution of users as a deviation from the mean time that the topic was discussed on Twitter, meaning that the leftmost group are early adopters, those at the peak are discussing the event as it is happening, or close to it in time, while the users to the right are talking about it after-the-fact. The users on the right of the peaks have the important benefit of hindsight. Note that for the Sochi data set, the gaussian curve is cut off because the data runs up to the time of writing of this article. Table 2 shows the crawl times for each plot. Both Sochi and Boston-Strong data sets show clusters of early adopters on the negative slope—an interesting subset for further analysis.

For the recognition/coverage feature all three collections show clusters of accounts with relatively high coverage. Manual inspection of these showed that they were official, government, media or other dedicated accounts to monitor the event during the crawling time, and are therefore a potentially useful information source. The decision feature shows the most interesting result across the three collections. Clearly there is a large amount of sentiment and opinion expressed about the Boston and BostonStrong collections, and the dedicated account clusters are clearly visible on the right. Looking at the sentiment polarity shows a more detailed account of the public feeling at the time. During the event time, the sentiment was all negative relating to the bombing incident, but when we look at the polarity score for the aftermath movement BostonStrong, we see clear signs of positive sentiment relating to the topic. These are likely tributes and other encouraging, hopeful messages stemming from the tragic event. For the olympics data, there is a more even distribution, which is intuitive given the winners and losers at the games.

Last, the awareness metric examined the immersion of a user in a topic, but weighted the score based on the number of tweets in $T$. These plots (bottom row of Figure 3) show a few accounts that are far more dedicated than the others. These accounts are again, likely to be dedicated to covering the topic for one reason or another.
IV EVALUATION

Thus far have described a mapping process between an abstract behavioral model from the field of educational psychology, and a measurable set of features in the Twitter network. We have performed an analysis of the behavior of each individual feature. The next step in our general framework is to evaluate data samples from the distributions in an effort to find useful thresholds for building a prediction model. Figure 4 illustrates the process on a sample distribution. \( m \) messages were sampled from \( n \) users from the extremities of each distribution plot. In this experiment, we chose \( m = 2 \) and \( n = 3 \) for each of the 6 features on each of the Sochi data collection and gauged perceived levels of competence, newsworthiness and topic-relevance in a crowd-sourced study.
1 FEATURE-BASED COMPETENCE ASSESSMENT

A study was run using Amazon’s Mechanical Turk crowdsourcing tool. In total, 150 participants completed the study. Participants were 62% Male, 38% Female, ranged in age from 18 to 58 and took an average of 12 minutes to complete the study. Most participants reported that they had strong reading ability and had at least a Bachelor level college education. A small payment of 50 cents was provided for completed studies. Sampled messages were presented to AMT evaluators in a simple web form. Participants were asked to read groups of three messages (coming from an individual user), and evaluate that user’s competence as an information provider in the target topic. Competence ratings were provided on the 5-point Dreyfus Scale from Novice to Expert. In addition to competence, newsworthiness and topic-relevance was also assessed. Table 4 lists all of the metrics that were recorded in the study. Here we focus only on the competence annotations ($COMP^+$ and $COMP^-$). Figure 5 shows the mean competence score (y-axis) on the Sochi data set for each feature in our mapped model (x-axis). The x-axis is grouped by $COMP^+$ and $COMP^-$, reflecting the users and messages sampled from the right and left sides of each feature distribution curve in Figure 3 and also illustrated in Fig 4.

Figure 3 shows some interesting results for each feature. The only instance where $COMP^+$ is lower than $COMP^-$ is on the recollection feature. In other words, the users selected from the left side of this feature distribution, i.e. the early adopters of the topic, received higher competence scores than those who began tweeting about the topic later in its evolution. This is a good indication that recollection is a useful feature for measuring competence in Twitter. The second group in Figure 3 (recognition) shows us that those users who covered a greater portion of the topic were considered to be more credible. The largest difference between competence ratings is for the opinionatedness feature. Here we can see that users in $COMP^+$ (right side of distribution curve, and highly opinionated) were rated as more competent than those in the $COMP^-$ group (left side of distribution, less opinionated), with a relative increase of 35.5%. The smallest difference was shown for the sentiment polarity group (12% relative increase for $COMP^+$ group), meaning that polarity of sentiment was less correlated with the competence annotations than intensity of sentiment, coverage of a topic or opinionatedness.

Figure 5 shows the general distribution of the ratings from the study, for each of the metrics in Table 4. This trend was evident across all data sets and features evaluated in the study, with mean ratings between 3 and 4 on the 5 point rating scale. Figure 7 shows a different perspective on the AMT data. Here, we focus on the trend in the difference between $COMP^+$ and $COMP^-$ across the rating bins from novice to expert. The upper chart shows the differences for the recollection feature. This tells us that there are far more early adopters of the topic in the proficient and expert bins than in the the novice and beginner bins. Interestingly, this was a significant
trend for the competence annotations, but not for the news worthiness annotations. The lower chart in Figure 7 shows the opposite trend for the opinionatedness feature: more highly opinionated users exist in the proficient and expert bins than the beginner and novice bins. These trends show that opinion and adoption-time (time of first tweet about the topic) are strong indicators of competence, but less so of news worthiness.

![Figure 7: Differences between AMT competence ratings for the Recollection and Opinionatedness features.](image)

<table>
<thead>
<tr>
<th>Series</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMP+</td>
<td>Competence score for tweets on right side of feature distribution</td>
</tr>
<tr>
<td>COMP−</td>
<td>Competence score for tweets on left side of feature distribution</td>
</tr>
<tr>
<td>NEWS+</td>
<td>Newsworthiness score for tweets on right side of feature distribution</td>
</tr>
<tr>
<td>NEWS−</td>
<td>Newsworthiness score for tweets on left side of feature distribution</td>
</tr>
<tr>
<td>REL+</td>
<td>Relevance score for tweets on right side of feature distribution</td>
</tr>
<tr>
<td>REL−</td>
<td>Relevance score for tweets on left side of feature distribution</td>
</tr>
</tbody>
</table>

Table 4: Description of recorded results from AMT study.

V CONCLUSIONS AND FUTURE WORK

This paper has presented a step-by-step generalizable framework for linking existing models of human behavior from the social and cognitive sciences with real world measurable features from the Twitter social network. Specifically the research proposed 5 integration steps and provided a worked example using the Dreyfus model of skill acquisition as a representative model. Features were mapped to a computational model over the Twitter network and behavior of each feature was analyzed over three large data collections. A study of 150 participants evaluated the competence levels of users sourced from both poles of the feature distributions. Results and manual analysis indicate that there is potential in the distribution plots to identify useful (competent) information sources related to a particular topic. A feature-by-feature comparison outlined a range of interesting effects between competence ratings for users selected from the poles of the feature distribution plots for the Sochi data collection. As a follow up study the authors propose to compare against a range of other models from the behavioral sciences, and to combine the resultant features into a predictive model and run accuracy-based evaluations over multiple ground-truth metrics. In conclusion, while there are many assumptions in the mapping stages of the approach, the authors believe that the methodology can help both algorithm designers for the social web and researchers in the behavioral sciences to better understand complex data interactions in Twitter.

Acknowledgments. This work was partially supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References


