Title
Using Social Graph Data to Enhance Expert Selection and News Prediction Performance

Author
Moghbel, Christopher

Publication Date
2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Using Social Graph Data to Enhance Expert Selection and News Prediction Performance

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Computer Science

by

Christopher S Moghbel

2013

ABSTRACT OF THESIS

Using Social Graph Data to Enhance Expert Selection

and News Prediction Performance

by

Christopher S Moghbel

Master of Science in Computer Science

University of California, Los Angeles 2013

Professor Junghoo Cho, Chair

Human intuition leads us to believe in the existence of experts, individuals with knowledge or insight that exceeds that of an average person. Can the idea of experts be harnessed to accurately perform popular news prediction? Can they perform this task better than “the crowd”, a collection of all or large amounts of the entire population? We explore this concept, first introducing various expert selection strategies, and then attempting to improve on them through the use of social graph data. We also examine the possibility of using expert characteristics and social data as parameters for machine learning models. Ultimately, we make two conclusions: it is extremely difficult for expert wisdom to outperform crowd wisdom, but expert selection can be used as a means of resource efficient sampling.

The thesis of Christopher S Moghbel is approved.

Stott Parker

Carlo Zaniolo

Junghoo Cho, Committee Chair

University of California, Los Angeles

2013

Table of Contents

1. Introduction

Human intuition leads us to believe in the existence of experts, individuals with knowledge or insight that exceeds that of an average person. We trust experts in our lives everyday, whether it's trusting our doctor to come up with the correct diagnosis for our symptoms, or trusting that our favorite team's coach will pick the correct strategy for the big game or draft the right player. Collective decision making, until recently, has been largely constrained to the political domain, and most often in the form of representative democracy, which can be viewed as using collective decision making to select the best expert(s) to run the country's government.

However, with the emergence of the internet, mobile devices, "Big Data" and techniques in Crowd Sourcing and Machine Learning, collective decisions are becoming more and more a part of people's everyday lives. People use decisions or information provided by the collective to find out how to get to where they want to go (crowd sourced mapping applications like Google Maps), where to eat (crowd sourced rating applications like Yelp or Foursquare), and what to read (crowd based news ranking sites like Reddit or Twitter Trends). Clearly, aggregating "crowd wisdom" can provide a great tool for harnessing the power of collective decision making.

Is crowd wisdom inherently better than expert wisdom? Should our doctor's be replaced by a crowd based diagnostic algorithm? Should the team's starting line up be chosen by polling the fans? Or is it possible that experts exist and can actually outperform the crowd's wisdom? This is a question that has been examined before in certain circumstances and domains. For example, the "Efficient Market Hypothesis (EMH)" [7] from Economics states that no expert can consistently outperform the market in making stock investments in an informationally efficient

market. Is this the case in every domain? Until recently, studying such a question has been extremely expensive, as it was very difficult to gather reliable data to represent crowd wisdom on a large enough scale. However, with the rise of social media, aggregating the wisdom of the populace is now not only possible, but inexpensive to accomplish.

We examine this question of expert versus crowd wisdom through a study of the news domain as represented on Twitter, one of the world's largest social media, micro-blogging services. To this purpose, we collected tweets from a large number of Twitter users over a long period of time to generate a body of crowd wisdom. We then examine this data in an attempt to determine possible criteria for selecting certain users who are "experts" in predicting popular news within a brief period after it's initial occurrence. We then compare the performance of these experts and the crowd (a polling of all the users in our data set) in a news prediction task.

Ultimately, we conclude a similar result to that of the EMH: expert wisdom cannot outperform crowd wisdom over the long term. However, we do find evidence to suggest that experts do exist, and that while their wisdom may not outperform the crowd in aggregate, it can be used to enhance or augment the crowd's wisdom, or to serve as an effective biased sampling of the crowd in circumstances where limitations on resources do not allow for an effective polling of the entire crowd. We also discover several interesting properties about the crowd, including that the removal of certain noisy members can actually serve to improve crowd wisdom. Finally, we discover that certain users with high influence can bias or sway the opinion of the crowd.

In the remainder of this paper, we will first discuss our data set and collection techniques along with relevant statistics in section 2. In section 3, we will discuss how we define and select

the crowd and various expert groups. In section 4, we will present the results of our experiments. We will then discuss related research in section 5 before providing our final conclusions in section 6.

2. Data Set and Statistics

In order to examine the relative performances of experts and the crowd, we first had to obtain a data set from Twitter that contained data on who tweeted what when, and how interesting it was. Ideally, we would want to obtain all tweets from Twitter relating to news in some way to form out data set. However, this is not feasible for multiple reasons. First, the number of tweets belonging to the news domain on Twitter is far beyond the scope of the computing resources available to us. Second, even if we did have the infrastructure to store and examine tweets on that scale, Twitter sets limits on it's public APIs regarding the number of tweets any one can download at any time. Finally, we would need a clear concept of what a piece of "news" is. In other words, we would need to be able to determine whether two tweets were regarding the same piece of news. Even if we limit our definition to a tweet containing a link to a news article, this remains a difficult problem How do we determine whether a story from The New York Times is about the same piece of news as another story from CNN?

To solve these issues, we limit the scope of tweets collected in our data set to those tweets containing a link to a story from The New York Times website. In such a case, even taking into account differences in URL (possibly through the use of different URL shortening services) or additional meta-content, we can determine if two stories are the same by comparing their titles. Also, in our initial studies, we collected tweets from other major news account on Twitter, such

as CNN, and found that links to stories on The New York Times' website outnumber those links pointing to other services approximately 10 to 1, giving us reasonable satisfaction that our data set would be strongly representative of our ideal data set. To help clarify our discussion, we now introduce a few definitions:

News Tweet - A news tweet is defined as any tweet that contains a link to the New York Times website, http://www.nytimes.com. This definition includes items such as blogs and opinion pieces in addition to traditional news stories.

News Thread - We define a news thread as the set of all news tweets that point to the same news piece. Two tweets are considered to point to the same piece of news if the page they link to contains the same <title> tag in their html markup.

Seed Tweet - We define a seed tweet as the first tweet of any news thread when viewed chronologically.

To collect our data set of news tweets, we used the public Twitter Streaming API. In order to collect only tweets containing a link to The New York Times, we set up a keyword filter through the Streaming API to view only those tweets containing the substring "http nyti" as both the full New York Times URL (http://nytimes.com) and it's shortened version (http://nyti.ms) both contain the string "http nyti". Over a 6 month period starting from August 1st, 2011 through

January 31st, 2012, we downloaded a total of 4,234,899 news tweets via the Twitter streaming API.

However, in order to properly perform experiments upon our data set, we needed to make sure that we could observe the full life span of any news thread. For example, if we encountered the seed tweet of a news thread on January 31st, 2012 (the last day of our data collection), there is no way to fully observe the activities of users tweeting this news story. Likewise, if we encountered the seed tweet of a news thread on August 1st, 2011 (the first day of our data collection), we could also not fully observe the activity of this news thread, as many users may have already been tweeting about this story before we began collecting data. Thus, we decided to perform *censoring*, a common techniques from Statistics, to deal with missing data. *Censoring* involves deleting or ignoring certain data from the tail ends of a data set in order to solve the missing data problem. In order to determine how large of a time period we should use censoring for on our data set, we decided to profile the typical life span of a news thread.

As a news thread is never truly finished (a user may decide to randomly tweet an old story 1 year after it came out), we needed to determine a point at which we would consider a news thread "inactive". For our analysis, we decided to consider a news thread inactive once it achieved 90% of the total tweets that it would accumulate in our data set. We then compared the lifespan of all news threads in our data set, as well as those in the top 2% of popularity, with this definition of inactive. As shown in figure 1, over 80% of all stories and popular stories are inactive after 100 hours (roughly 4 days), and almost 100% of both all and popular stories are inactive after 1000 hours (approximately 42 days). Also interesting to note is that 40% of all news threads are immediately inactive, essentially meaning that 40% of news stories only ever

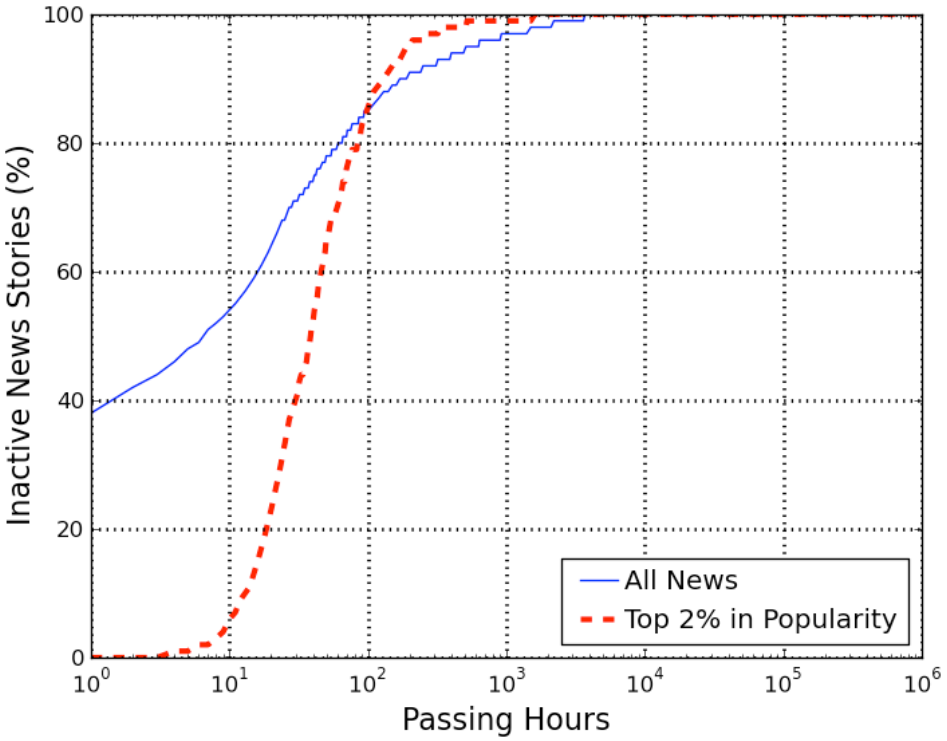

Figure 1: Longevity of News Threads

receive 1 tweet. Based on the results of this analysis, we decide to perform *left-censoring* on our data set for 1 month, and *right-censoring* on our data set for 1 month. In other words, we do not consider any news thread whose seed tweet occurs in the first month or last month of our data set. After censoring our data set in this fashion, we were left with 2,837,026 news tweets from a total of 402,102 unique users.

3. The Crowd and Expert Selection

In this section, we discuss our formal definition of the crowd and our models for expert selection.

Earlier, we mentioned that to perform our experiments, we needed data on who tweeted what when, and how interesting it was. In the previous section, we discussed how we collected a data set telling us the who, what, and when, but so far we have not discussed how we determine the interestingness of a news thread. Obviously, this is a subjective issue: no two people will completely agree on the relative interestingness of a large set of news stories. However, aggregate popularity is often highly correlated with interestingness, and often used as a way of objectively assessing the interestingness of something. In our experiments, we use popularity as a way of objectively measuring the interestingness of a news thread. Thus, if a news thread is more popular (receives more tweets), then we consider it more interesting. To facilitate our experiments and discussion, we now introduce our definition of a "interesting" news thread:

Golden Set - The golden set is defined as the top k% of news threads, when ranked and sorted by the number of tweets they receive in total in our data set.

The golden set serves as our ground truth for the interestingness of news threads in our experiments.

Now that we have defined how we determine the interestingness of a news story, we will discuss the means by which we select experts. When thinking of an expert in a traditional sense, we believe there are two qualities that are generally attributed to experts. First, for someone to be an expert, all or most of the decisions they make should be correct. This concept translates to the metric of *precision*, one of the standard metrics used in the IR community. More formally, precision in our experiments is defined as follows:

$$precision = \frac{\#\ of\ recommended\ news\ in\ golden\ set}{\#\ of\ all\ recommendations\ by\ user}$$

Second, of all the possible decisions to be made, an expert should correctly make most or all of them. In other words, someone cannot be an expert if they make 1 correct decision, and then never attempt to make another, as this correct decision could be attributed to random chance. This concept translates to the metrics of *recall*, another standard metric used in the IR community. More formally, recall in our experiments is defined as follows:

$$recall = \frac{\#\ of\ golden\text{-}set\ recommended\ news\ by\ the\ user}{total\ \#\ of\ news\ in\ the\ golden\ set}$$

Together, precision and recall are the two main criteria for regarding someone as an expert.

However, in our news prediction task on Twitter, we identify two additional criteria that may effect whether someone is selected as an expert. The first of these is *promptness*, as we are concerned with an expert's ability in identifying a news story before it becomes popular. In other words, we want to avoid selecting experts who are "Monday Morning Quarterbacks", or those persons who only make decisions after the outcome is already obvious. Our final possible selection criteria is *influence*, as a user who has high influence can often sway the opinion of others. In the context of Twitter, if a user has an extremely high number of followers, any news they tweet is much more likely to be tweeted by their followers (a very large number), and thus is more likely to become popular. As such, intuition tells us that highly influential users may be good candidates for selection as experts, due to their ability to influence the ultimate ground truth. We will go into this phenomenon, which we refer to as *social bias*, in more depth during the analysis of our experimental results.

With these criteria in mind, we now present 4 distinct models for expert selection: *Precision Frequency*, *F-Score*, *Confidence Interval*, and *Social Bias*. The first three of these models (*Precision Frequency*, *F-Score*, and *Confidence Interval*) take into account only the first three criteria, while the fourth (*Social Bias*) takes into account only the fourth criteria, influence.

The **Precision Frequency** model relies primarily on ranking users based on their precision at the news prediction task during the training set. Then, we select the top k% of users based on this ranking as experts. However, this simple model fails to take into account recall, and allows for those users who tweet once and score a hit to be considered experts. To solve this issue in this model, we take a naive approach and setting a threshold filter, f, or a minimum number of tweets the user must achieve in order to be considered an expert. Any user having less than f tweets during the training period will not be considered as a candidate in this model.

The **F-Score** model attempts to solve this issue in a less naive way. Instead of only calculating how precise a user is in the training period, we instead calculate each users *F-Score* metric (a well known metric in the IR community), which combines both precision and recall. The precise formula for computing a users F-Score is as follows:

$$F_\beta = (1+\beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall}$$

In the formula, B serves as a tuning parameter that adjusts the relative weight given to each precision and recall. When B = 1, precision and recall are given equal weight. Once users are ranked by their F-Score, we then select the top k% of users as experts. Because F-Score takes recall into account, user's who make one lucky tweet will not be selected.

Another way of tackling this issue is to model the uncertainty in our expert selection through a **Confidence Interval** model. In such a model, we can view our certainty that a user really is an expert as increasing with the number of correct selections (tweets) they make. That is, our certainty that a user is an expert is much higher if a user makes 10 correct tweets than if they make one. In our model, we use a Wald 1-sided confidence interval [1, 4] at a 95% confidence level. To take an example, if a user recommends one article and does so correctly, the confidence level of their precision is between 0.24 and 1. This large gap indicates a large level of uncertainty, and would lead to that user not being selected as an expert.

Our final expert model takes into account solely a user's influence or social bias factor, and thus we call it the **Social Bias** model. Our approach with this model is to select experts based solely by their number of followers. Thus, once users are ranked by number of followers, we select the top k% as experts.

Given these definitions, we define the **Crowd** as being all users in the data set, except for those selected in any one of the expert groups (i.e. the union of all experts). Defining the Crowd in such a way allows us to easily compare and contrast the wisdom of the crowd versus the wisdom of the various expert groups. With an expert selection level set at the top 2% in our experiments, and with some users being selected as experts by multiple models, the Crowd contains 96.1% of all users in our data set.

4. Results

4.1 Experiment Set Up

In order to select our experts, we first need a *Training Set*, in which we will evaluate the performance of all users, and use our models to select our various experts groups. We then need a *Testing Set*, in which the performance of our expert groups selected from the Testing Set are compared against the results of the crowd. To create these sets, we take data from the dates of September 1st to October 31st (the first two months of our data set after cleaning) as the Training Set, and data from the dates of November 1st to December 31st (the second two months of our data set after cleaning) as the Testing Set.

A means of evaluating the performance of each of our various groups was also needed before we could begin our experiments. To do this, we ask each group (experts and the crowd) to provide a recommendation list of news articles. This is done by aggregating all tweets from every member in the group that occurs within a certain promptness threshold (for example, within 4 hours of the seed tweet). All news articles are then sorted by number of tweets, which can be viewed as votes, and the top *n* articles are selected. The recommendation list for each group is then compared against the Golden Set to create a Precision-Recall curve. This curve is generated by increasing the number of articles selected from a group's recommendation list from 1 to n (with a step size of 1), until all articles in the list are selected. With the Precision-Recall metric, we expert to see the first few selections from each group be a hit (ie we expect the top ranked stories from each group to exist in the Golden Set). Then, as the number of stories selected from the recommendation list increases, we expect recall to increase from 0% towards 100% (as more stories are being selected), but precision to gradually decline as the group starts to select stories incorrect (i.e. recommend stories not in the Golden Set), as stories lower down in their recommendation lists are selected. Note the Precision-Recall curves are a metrics widely

used to evaluate the performance of recommendation or information retrieval systems in the literature [10, 11].

Finally, values for the various parameters in our experiments needed to be chosen. After repeated experimentation, the top news level was set at 5% (ie the Golden Set consisted of the top 5% of news when ranked by popularity), the expert group size was set at 2% of all users, and the promptness threshold was set at 4 hours. Also, the B value for the F-Score model was set as 2. Unless otherwise noted, our experiments used this combination of parameters. We also tried various other parameter levels: top news set to 1%, 2%, 5%, and 10%; expert group size set to 1%, 2%, 5%, and 10%; and the promptness threshold set to 1, 2, 4 and 8 hours. However, unless otherwise specified, any change made in the parameter settings for our experiments did not significantly alter the results of the experiments, or change our evaluation of the results.

4.2 Expert Selection Model Performance

We now present the results of the three past-performance based expert selection strategies (Precision-Frequency, F-Score, and Confidence Interval), and how they perform against the Crowd.

As we can see from the graph in Figure 2, none of these three expert selection strategies were able to outperform the Crowd. We also note that, of the three strategies, the Confidence Interval strategy performed the best. Both these findings were replicated in our experiments with different parameter combinations.

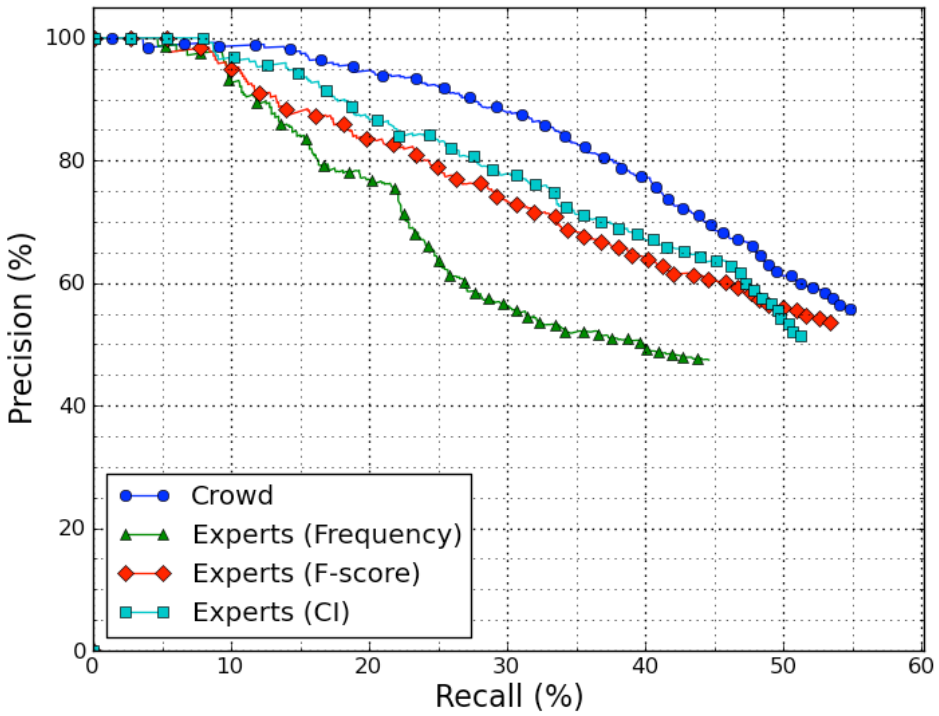

Figure 2: Wisdom Comparison (Promptness: 4hrs, Top News Size: 5%, Expert Size: 2%)

What makes Crowd Wisdom so hard to beat? To examine this question, we performed the same experiment, but this time comparing the performance of crowds of varying sizes, along with our best expert selection strategy, the Confidence Interval strategy. These smaller crowds were created by random sampling from the larger crowd. We see in Figure 3 that, as crowd size increases, so does performance at predicting popular news. We also note that this increase does not happen in a linear fashion. Increasing crowd wisdom from 33% to 100% leads to only a small increase in performance, whereas increasing crowd size from 10% to 33% has a much larger performance increase. Our best expert selection strategy, the Confidence Interval strategy, performs roughly equivalent to a random sampling of 33% of the crowd.

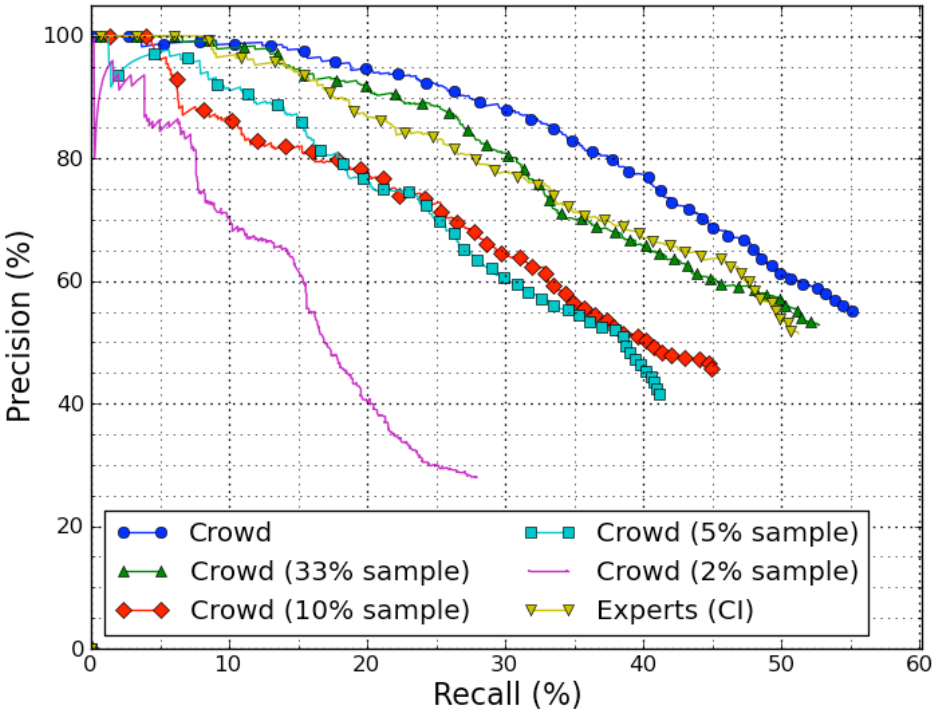

Figure 3: Crowd Size Comparisons (Promptness: 4hrs, Top News Size: 5%, Expert Size: 2%)

Since our expert selection strategies contain a number of users equal to about 2% of the full crowd, we can see that, per user, our strategies outperform pure random sampling. As such, these strategies could provide the basis for a news prediction engine that faces resource limitation. Indeed, it seems that a resource efficient means of creating a news prediction engine would be to use these expert selection strategies to rank users, and then sample as many users as possible, in order of their expert rank.

4.2 Leveraging Expert Wisdom to Boost News Prediction

Despite the fact that crowd wisdom consistently beat expert wisdom overall, the expert groups often exhibited wisdom on certain articles that outperformed the crowd's predictions.

This can be seen when comparing an article's ultimate ground truth rank against its crowd ranking and its ranking from each of the expert groups. Occasionally, one or more expert groups would rank a popular story highly but the crowd would rank that same story much lower. For example, the article "http://www.nytimes.com/interactive/2011/12/06/magazine/13villains.html" had a ground truth ranking of 4. The crowd ranked this article in 1970th position, but both the expert groups chosen by the Precision model and the expert group chosen by the Confidence Interval model ranked this article as the most popular story. Inspired by this trend, we designed experiments to see if these moments of expert insight could be harnessed to improve news prediction accuracy.

Our first attempt was based on the observation that, at lower recall rates, each expert group performed the news prediction task with 100% precision. Thus, we determined for each expert group in the training set the maximum recall level at which they still performed the task with 100% precision. For ease of reference, we call articles chosen by a group within this recall level as the group's "Strongly Recommended Set". To generate our boosted results, we started with the crowd predictions. Then, whenever we came across a story that at least one expert group had ranked within their Strongly Recommended Set but was not contained within the crowd's predictions, we boosted that story into the news prediction set given by the crowd. However, the results of this experiment showed that our boosted model was still outperformed by the crowd.

Examining the results in more detail, it seemed that a positive signal from one expert group was not strong enough to justify altering the crowd's news prediction set. Re-examining the data, we noticed that occasionally two or more expert groups would highly rank an article that the crowd had mistakenly ranked as low popularity. This is the case in the previous example,

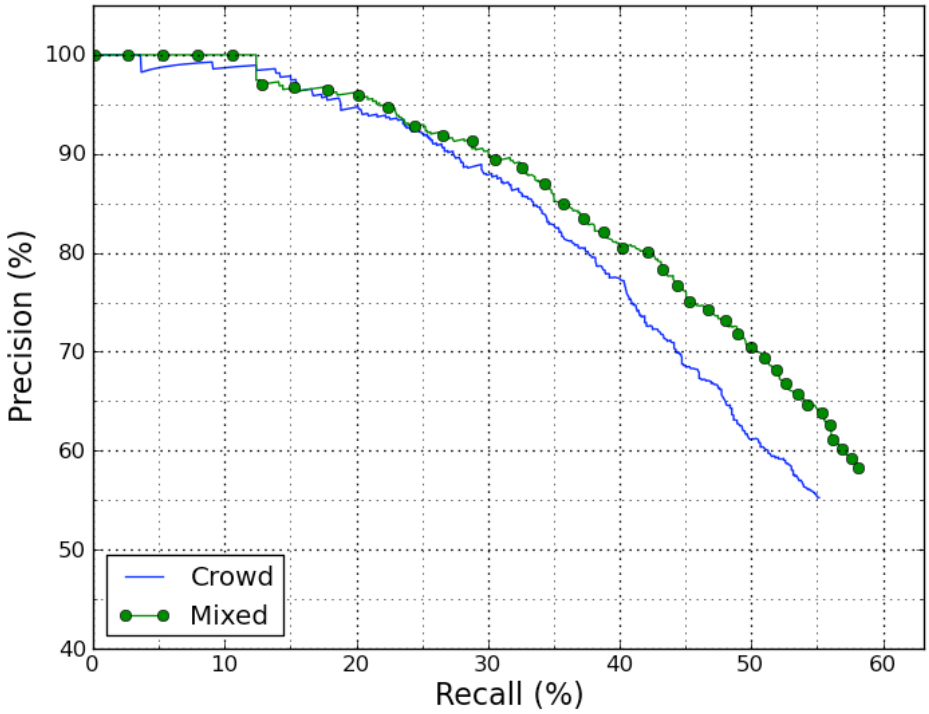

Figure 4: Results of boosting experiment, second attempt
(Promptness: 4hrs, Top News Size: 5%, Expert Size: 5%)

where both the Precision based expert group and the Confidence Interval based expert groups both ranked the given article within their Strongly Recommended Set. Attempting to take advantage of this observation, we modified the previous experiment. In this iteration, we boosted an article only if at least two expert groups ranked it within their Strongly Recommended Set. The results of this experiment, shown in figure 4, were much more promising, with our boosted model reliably outperforming the crowd at the news prediction task at all recall levels.

It can be seen from these results that the expert groups do indeed capture some wisdom that the crowd misses. It can also be seen from the failure of our first experiment that often a

single expert group can be biased towards certain articles, due to the desired aspects emphasized by that expert selection model. However, when two or more expert groups, picked via two different models, agree on the popularity of a news article, this provides a very strong signal for an accurate prediction. This insight can then be harnessed to improve overall accuracy in the news prediction task.

4.3 Super Experts: Attempting to Combine Expert Wisdom

In light of the results of the boosting model, we wondered if there were other ways in which expert wisdom could be harnessed to enhance news prediction. Since each expert selection model puts emphasis on different desired expert attributes, it seemed that a single expert group could be biased towards certain articles. However, operating under the assumption that combinations of expert groups could provide strong signals, we hypothesized that an expert picked by multiple models might indeed be a true expert, or "super expert", and that a group of these super experts might be able to outperform the crowd at the news prediction task.

To test our hypothesis, we re-ran our previous experiments with the addition of this new super experts group. More formally, we define the super expert group as the intersection between the Precision model based expert group, the Confidence Interval model based expert group, and the F-score model based expert group. However, the super expert group performed poorly compared to both the crowd, and the individual expert groups (for ease of viewing, in the figure we show the super expert's performance in comparison to the Crowd and the Confidence Interval expert group). Looking into the data, the reason for this seems to be the small size of the super

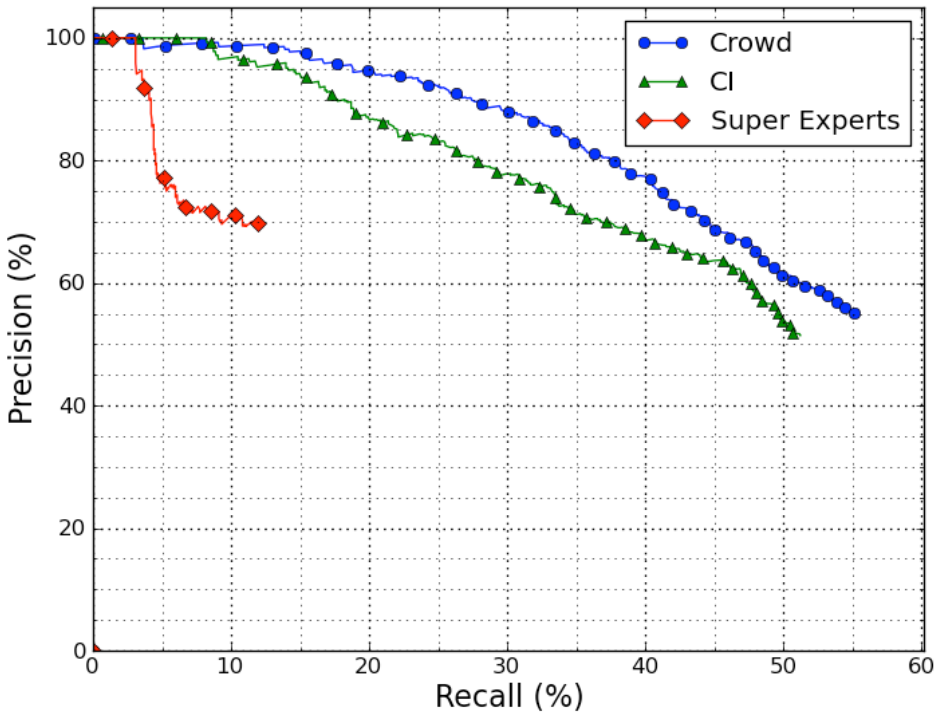

Figure 5: Wisdom Comparison with Super Experts (Promptness: 4 hrs, Top News Size: 5%, Expert Size: 2%)

expert group. Compared to the individual expert groups, which each had 2836 users, the super expert group contained only 191 users. Furthermore, only 376 tweets in the testing set were from

To test our hypothesis, we re-ran our previous experiments with the addition of this new super experts group. More formally, we define the super expert group as the intersection between the Precision model based expert group, the Confidence Interval model based expert group, and the F-score model based expert group. However, as seen in figure 5, the super expert group performed poorly compared to both the crowd, and the individual expert groups (for ease of viewing, in the figure we show the super expert's performance in comparison to the Crowd and the Confidence Interval expert group). Looking into the data, the reason for this seems to be the small size of the super expert group. Compared to the individual expert groups, which each had

2836 users, the super expert group contained only 191 users. Furthermore, only 376 tweets in the testing set were from super expert users, compared to 2555 for the precision based experts, 6719 for the confidence interval based experts, and 222,984 for the F-score based experts (note that the F-score model is biased towards selecting experts based who achieve a very high level of recall). With such a small set of users and tweets, the super expert group does not provide enough information to provide accurate predictions.

Despite the poor results of this particular experiment, this seems like a promising area for further research. We can see from the boosting experiments that combined expert wisdom can be valuable for performing news prediction. If a method for combining expert groups can be devised that produces a large enough group and sufficient information for the news prediction task, it seems likely that we will see promising results.

4.4 Utilizing Social Graph Data for Expert Selection

Thus far in our investigations, we have ignored social graph data when performing expert selection. However, as shown by the work of other studies, we know that users with high levels of influence can disproportionally influence whether or not a story becomes popular on Twitter. Figure 6 shows the rate at which two separate stories, one eventually popular and one not, accumulate tweets over time. The unpopular story is one where the @nytimes Twitter account, the official Twitter account of The New York times with over 5 million followers (and thus highly influential) never tweets a link to that story. In fact, no users with significant levels of influence take part in tweeting that news story. As such, we see a gradual ascent in the tweets accumulated followed by a gentle tapering off as time passes and the story is no longer

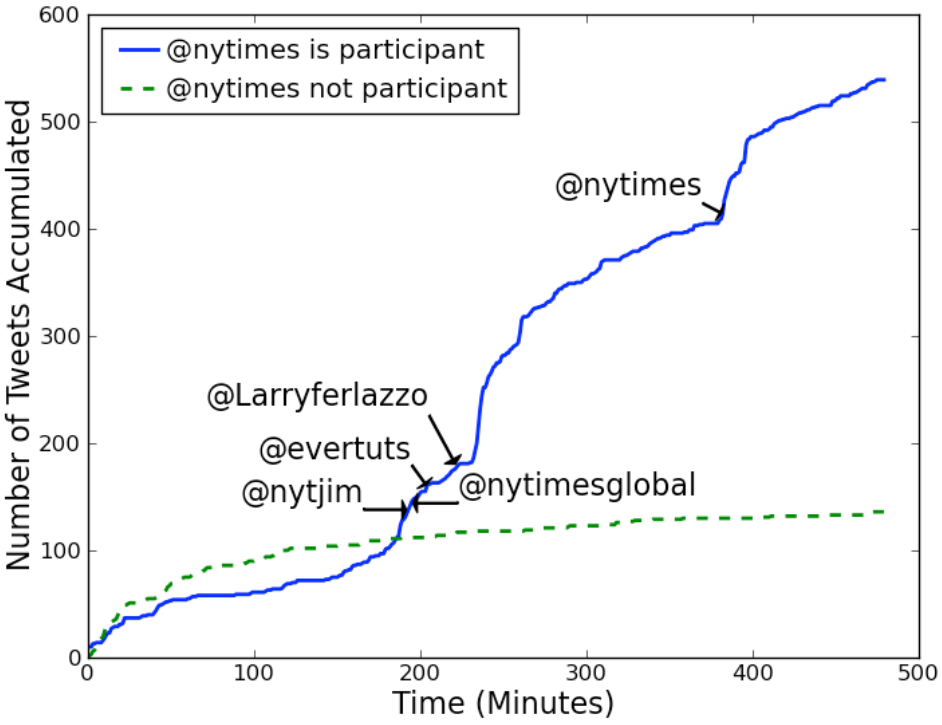

Figure 6: Tweet Rate Spikes cause by Highly Influential Users

interesting. The popular story is one in which the @nytimes account, along with several other highly influential users, participate by producing tweets with a link to that article. At first, we see that is has a growth rate similar to that of the unpopular story. However, about 200 minutes after the article is first seen on Twitter, a number of highly influential users tweet that news article. Despite the fact that most news stories accumulate a large percentage of their tweets in the first few hours, immediately following these tweets by influential users, we see a large spike in the rate of tweet accumulation, which is sustained for multiple hours. Then again, after about 400 minutes, the @nytimes account tweets the news article, and we see another large spike owing to the immense influence of the @nytimes account, which is one of the most followed Twitter

accounts. From this case study, we can see that it does indeed seem that highly influential users can affect whether a story becomes popular or not.

Seeing this, we believed it would make sense to leverage social graph data to pick experts. The first question we asked ourselves was whether a group of experts selected solely by their influence (their number of followers on Twitter), would perform well against our other expert selection models. To test our hypothesis, we repeated our experiment with the addition of an expert group picked solely by their number of followers.

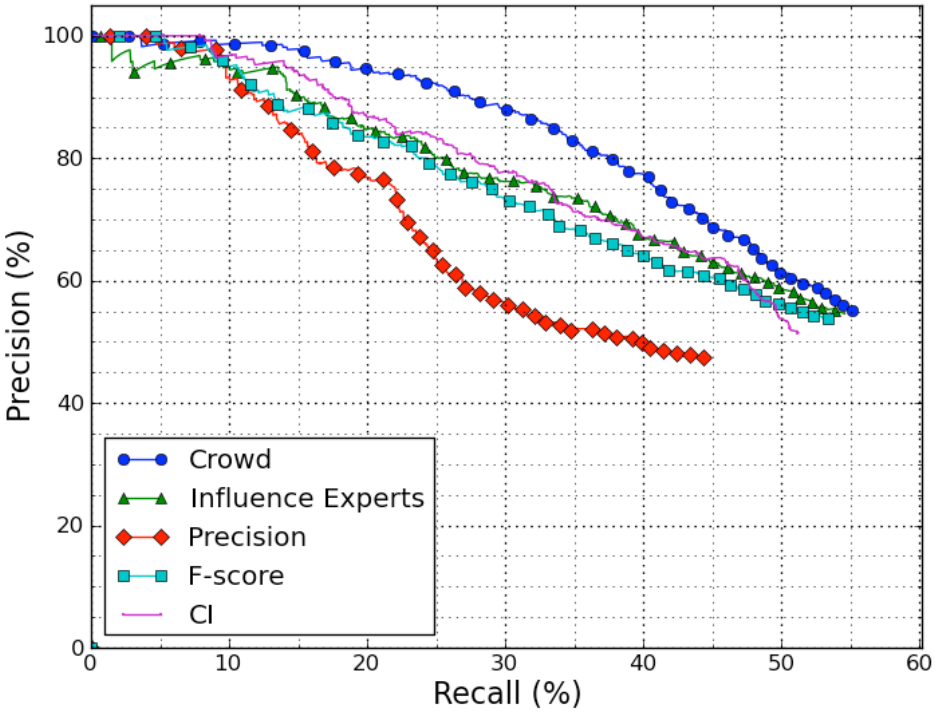

Figure 7: Wisdom Comparison with Influence Experts (Promptness: 4hrs, Top News Size: 5%, Expert Size: 2%)

Looking at the results of the experiment in figure 7, we can see that even an expert selection model as simple as highest influence performs reasonable well, outperforming the F-score model at recall levels of about 10%, and even equalling the precision-based model at recall

levels of about 20%. It seems that social influence is a significant factor that can be harnessed for expert selection.

To further demonstrate this effect, we performed another experiment in which we divided the Confidence Interval expert groups into two groups: those with high social influence, and those with low social influence. To perform this split, we simply divide the group by number of followers. If a user has a number of followers greater than 50% of the experts, they are placed in the "high" group, otherwise, they are placed in the "low" group.

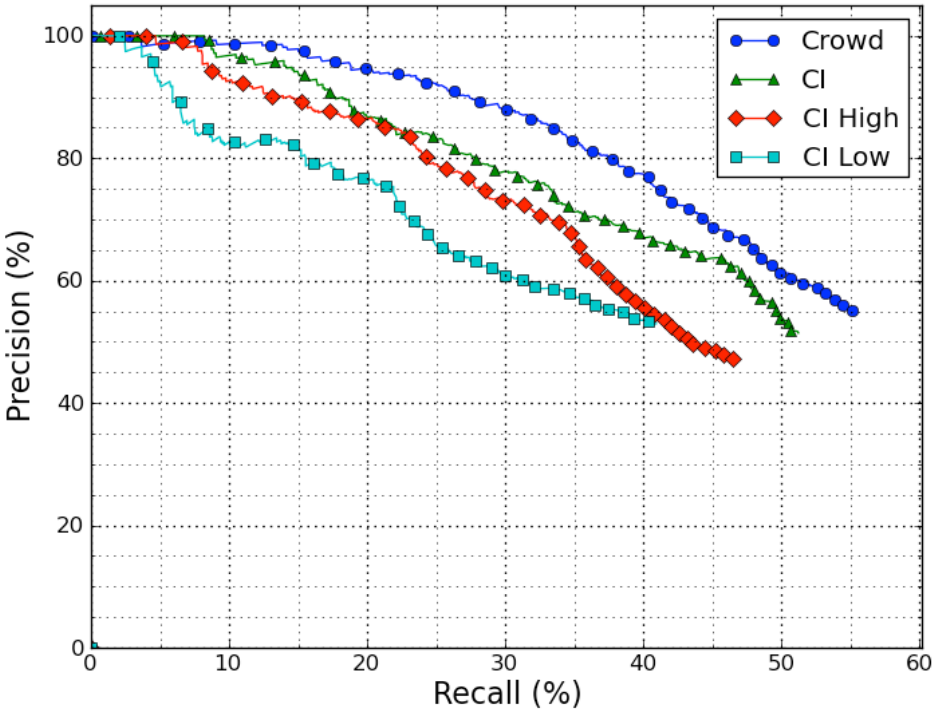

Figure 8: Comparison of high and low social influence experts
(Promptness: 4hrs, Top News Size: 5%, Expert Group Size: 2%)

We can see from figure 8 that the experts with high social influence do perform significantly better on their own when compared to those experts with low social influence. However, it does

seem that, especially due to its poor performance as lower recall levels, social influence is not a strong enough signal on it's own to support an expert selection strategy. Instead, we believe it is best combined with a precision based expert selection strategy to achieve the best results.

4.5 Augmenting Precision Based Models with Social Influence

As discussed in the previous section, the social influence a user holds seems to be a signal as to their performance in the news prediction task. In this section, we attempt to create new expert selection strategies by augmenting our previously discussed precision based expert selection strategies with social influence data. In particular, we will introduce a new expert selection strategy called Precision Weighted Followers.

The Precision Weighted Followers strategy augments the Confidence Interval expert selection strategy by giving higher weight to votes cast by those experts deemed to have a "high" social influence. Again, an expert is considered to have high social influence if they have a number of followers greater than 50% of the experts selected by the initial strategy. The first step in the Precision Weighted Followers strategy is to pick an expert group by using the Confidence Interval expert selection strategy. Then, when creating the ranking for the group's News Selection Set, instead of giving each vote the same weight (a simple count of number of tweets for each story), each vote is weighted according to the following formula, where S_i is the score for story i, Vl_{ow} is the number of tweets by experts with low influence, V_{high} is the number of tweets by experts with a high influence, and w is an assigned weighting :

$$S_i = w * V_{low} + (1 - w) * V_{high}$$

For example, if a story X is tweeted by expert A, who has high influence, and experts B and C, who both have low influence, and we set w to be .35, then the total score for story X will be 1.4 (2 * .35 + 1 * .75), as opposed to the score of 3 that would be given by the unweighted ranking strategy. To evaluate the performance of this model, we ran the same precision-recall experiment, setting this strategy against the pure Confidence Interval strategy as well as the crowd.

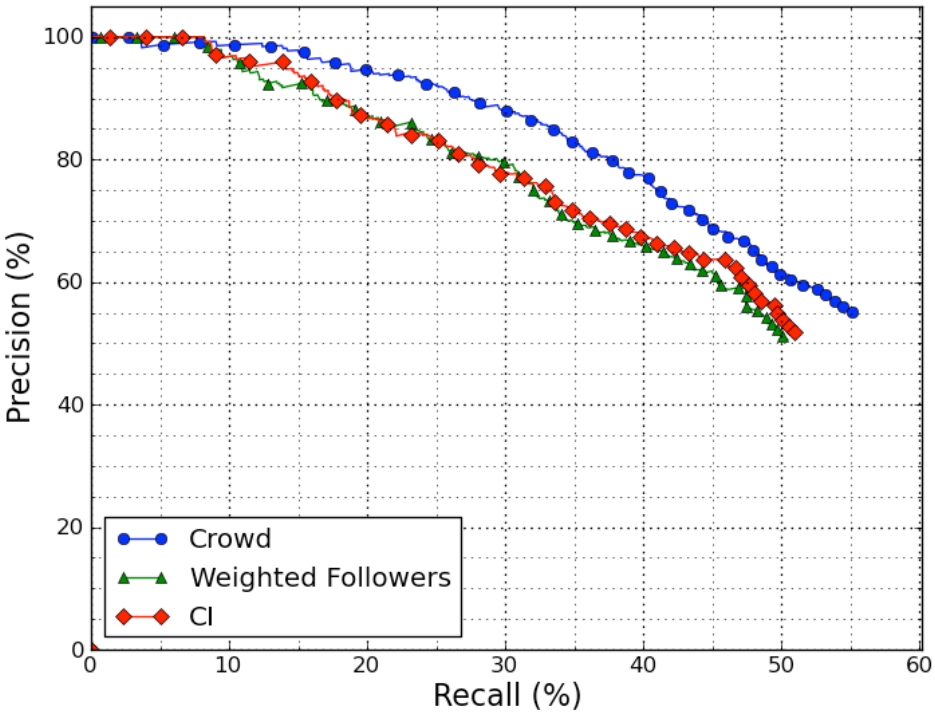

Figure 9: Precision Weighted Followers Model (w1 = .35)

However, despite experimenting with multiple weights for w, we were unable to find a weight in which the Precision Weighted Followers model significantly outperformed the standard Confidence Interval model. Despite being able to show that social influence is indeed a factor in an expert groups performance at the news selection task, using this signal to create a new expert

selection strategy, or augment an existing one, is a challenging problem. This subject is one we envision as an area for future work.

4.6 Using Expert Characteristics and Social Data with Machine Learning

In the process of defining our various expert selection strategies, we noticed that we had several metrics with which we determined the quality of a user and their ability to predict popular news. These were precision, recall, F-Score, and Confidence Interval score. In addition to this, we had the additional metric of number of followers as an indicator of a user's social influence. This led us to wonder if these characteristics that fueled our expert selection strategies could be used as parameters to machine learning algorithms, and whether such a model would be able to beat the crowd.

To perform this experiment, we transformed out data such that each tweet was accompanied by these 5 parameters. This data was then fed to the Support Vector Machine (SVM) module implemented by Chang et. al. [3]. We kept the same training and testing periods as in our previous experiments. Once a model was learned from the training set, we then used the libsvm library to make predictions on which news from the testing period would appear in the golden set. This allowed us to generate the same precision-recall curves as in our previous experiments to compare the performance of SVM against both the crowd and our expert selection strategies.

We see in figure 10 that using SVM with the aforementioned parameters, we were able to roughly match but not outperform the crowd in the news prediction task. From this, we gather further evidence supporting our conclusion that it is extremely difficult to outperform the crowd at the news prediction task, as even very sophisticated machine learning algorithms can only

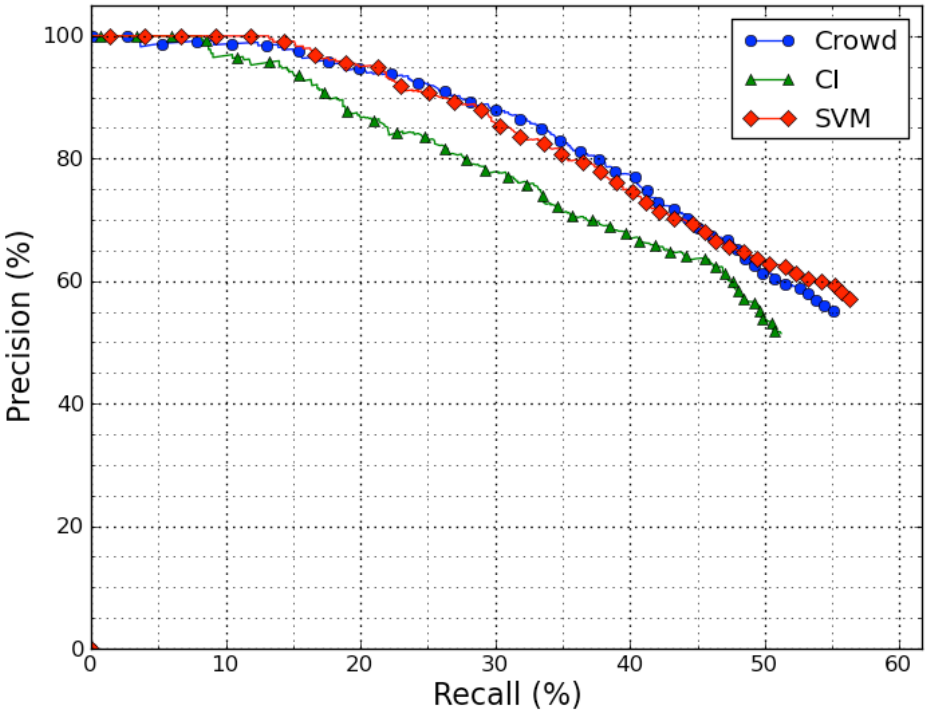

Figure 10: SVM News Prediction Comparison (Promptness: 4hrs, Top News Size: 5%)

match the Crowd's performance. In our experiments SVM learning and prediction was done with the libsvm default settings. It may be possible that, with additional parameter, and through experimentation with SVM and libsvm's different settings, it may be possible to train a model that can outperform the Crowd. However, we leave this as future work for those with more experience with SVM and machine learning.

5. Related Research

News recommendation on Twitter: In [14], Kwak. et al. give a overview of information on Twitter, including the fact that 85% of hot topics on Twitter are headline news. In [18], Petrovic et al. show how to detect the birth of a news story on Twitter, while Phelan et al. discuss

how to recommend real-time topic news in [19]. Yin et al. also examine this question in [23]. In contrast to these works, we view news recommendation through the lenses of group decision making, and also future popularity.

Crowd vs expert wisdom: In [8, 13], researchers argue that the crowd will make better decisions than the expert. Also, in [7], the EMH states that no expert can continually outperform an efficient market. However, Hill et al. argue that the wisdom of a group may not always exceed that of a larger group in [12]. We explore this same problem in this paper, on a significantly larger scale, and applied to the news prediction domain through Twitter.

Social Influence and Bias: In [2], Bakshy et al. found that highly influential users may in certain cases be cost-effective for making predictions, while in other circumstances those with less influence may actually perform better. Ma et al. shows means for augmenting recommender systems with social data in [16]. In [17], Mishra and Bhattacharya show how to calculate bias and prestige scores for nodes in a network based on trust score. In [18], Wu et al. explain the TwitterRank algorithm for computing the influence of user's on Twitter, and find that 72.4% of users on Twitter follow back at least 80% of their own followers. In [22], Wu et al. find that roughly 50% of URLs consumed on Twitter are generated by 20K "elite" users, and find a high degree of homophily within categories of users. In this paper, we attempt to take advantage of the social influence characteristics discussed in these papers to create a system that can perform better at predicting future popular news.

6. Conclusion

In this thesis, we explored whether it is possible to discover experts who can outperform the crowd in predicting popular news. To do this, we introduced three expert selection strategies: Precision-Frequency, F-score, and Confidence Interval. We then explored ways in which social graph data could be used to improve the performance of these expert selection strategies. Finally, we explored whether the characteristics and signals we used to select experts could be used as input to machine learning algorithms in another attempt to outperform the Crowd in predicting popular news. Ultimately, none of our strategies, even the sophisticated machine learning algorithm SVM, could outperform the Crowd, forcing us to conclude that doing so would be extremely difficult, if not impossible. However, we also conclude that the characteristics and strategies we identified do help find users who outperform the average when it comes to predicting popular news. We also propose that this knowledge can be used to create a resource-efficient news prediction engine.

7. References

[1] A. Agresti and B. A. Coull. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, May 1998.

[2] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.

[3] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 3, Article 27 (May 2011), 27 pages.

[4] L. D. Brown, T. T. Cai, and A. Dasgupta. Interval estimation for a binomial proportion, July 30 1999.

[5] 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*, MDMKDD ’10, pages 4:1–4:10, New York, NY, USA, 2010. ACM.

[6] M. Demirbas, M. A. Bayir, C. G. Akcora, Y. S. Yilmaz, and H. Ferhatosmanoglu. Crowd sourced sensing and collaboration using twitter. In *WOWMOM*, pages 1–9. IEEE, 2010.

[7] E. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, May 1970.

[8] R. Frederking and S. Nirenburg. Three heads are better than one. In *Proceedings of the fourth conference on Applied natural language processing*, ANLC ’94, pages 95–100, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[9] Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. 2010. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining* (WSDM '10). ACM, New York, NY, USA, 241-250.

[10] A. Gunawardana and G. Shani. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10:2935–2962, 2009.

[11] J.L. Herlocker, J.A. Konstan, L.Terveen, and J.T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5-53, 2004.

[12] G. W. Hill. Group versus individual performance: Are N+1 heads better than one? *Psychological Bulletin*, 91(3):517–539, May 1982.

[13] A. Kittur, B. A. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World wide web*, 1(2):19, 2007.

[14] 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*, WWW ’10, pages 591–600, New York, NY, USA, 2010. ACM.

[15] K. Lerman and T. Hogg. Using a model of social dynamics to predict popularity of news. CoRR, abs/1004.5354, 2010.

[16] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender systems with social regularization. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM ’11, pages 287–296, New York, NY, USA, 2011. ACM.

[17] Abhinav Mishra and Arnab Bhattacharya. 2011. Finding the bias and prestige of nodes in networks based on trust scores. In *Proceedings of the 20th international conference on World wide web* (WWW '11). ACM, New York, NY, USA, 567-576.

[18] S. Petrovic, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In HLT-NAACL, pages 181–189. *The Association for Computational Linguistics*, 2010.

[19] O. Phelan, K. McCarthy, and B. Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems*, RecSys ’09, pages 385–388, New York, NY, USA, 2009. ACM.

[20] V. V. Raghavan, G. S. Jung, and P. Bollmann. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems*, 7(3):205–229, 1989.

[21] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. TwitterRank: finding topic sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining* (WSDM '10). ACM, New York, NY, USA, 261-270.

[22] Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web* (WWW '11). ACM, New York, NY, USA, 705-714.

[23] P. Yin, P. Luo, M. Wang, and W.-C. Lee. A straw shows which way the wind blows: ranking potentially popular items from early votes. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM ’12, pages 623–632, New York, NY, USA, 2012. ACM.