Using Prediction Markets and Twitter to Predict a Swine Flu Pandemic

Joshua Ritterman
School of Informatics
University of Edinburgh
j.ritterman@sms.ed.ac.uk

Miles Osborne
School of Informatics
University of Edinburgh
miles@inf.ed.ac.uk

Ewan Klein
School of Informatics
University of Edinburgh
ewan@inf.ed.ac.uk

July 30, 2009

Abstract

We explore the hypothesis that social media such as Twitter encodes the belief of a large number of people about some concrete statement about the world. Here, these beliefs are aggregated using a Prediction Market specifically concerning the possibility of a Swine Flu Pandemic in 2009. Using a regression framework, we are able to show that simple features extracted from Tweets can reduce the error associated with modelling these beliefs. Our approach is also shown to outperform some baseline methods based purely on time-series information from the Market.

1 Introduction

Prediction markets are mechanisms for aggregating beliefs about concrete outcomes in the world. They are structured as a betting exchange or peer-to-peer gambling system, where participants bet amongst themselves as to the outcome of specific world events—such as who will win the 2009 US election, or which Hollywood movie will achieve highest box office totals. The prediction market organizer creates 'shares' in an event occurring. People can then buy and sell these shares at a price determined by the market. In this way, the market drives the price of a share to the mean belief of traders, interpreted as the probability of the event occurring (Gjerstad, 2006; Wolfers and Zitzewitz, 2004). In fact, it has been shown that this type of market system is able to generate an optimal global solution to the prediction problem better then any individual expert (Watkins, 2007). It is also a considerably less expensive method than alternative methods, such as hiring analysts for an expert opinion on the outcome of an event, or conducting a poll. For this reason, prediction markets have become a major area of interest to governments, corporations and academics over the last few years.

In this paper we explore the hypothesis that we can extract useful information from social media and that modeling this information will yield better results then a model constructed with information from the prediction market in isolation. We have collected almost 50 million Twitter posts (Tweets) over roughly a two month period. We will present a method using this Twitter data to forecast the closing price of a prediction market, thus showing that we can explictly model changes in the belief, as represented in a Prediction Market, from beliefs implictly represented in Tweets.

2 The Task

Since prediction markets are considered to be overall analogous to public opinion, a model that is able to forecast such markets would be a valuable supplement to opinion polling and market research. We will focus on using the Hubdub online prediction market[1] to model public belief about the possibility that H1N1 (Swine Flu) virus will become a pandemic. On April 10th 2009, just after news about the virus became public, Hubdub posted the following question:

> *Will Influenza A (H1N1) (aka "swine flu") grow into a pandemic in 2009 as feared?*

By modelling this market, we can thereby model public belief about the event in question.

Unlike newswire, Twitter goes beyond factual information in providing a wealth of information about public opinion on a topic. Tweets contain rumor, commentary, opinion, and even jokes; cf. Table 2. When news about H1N1 first broke, Twitter was highly active with posts about the spread of the flu, and in fact was considered by CNN to be overreacting (CNN, 2009). However, in the weeks that followed, mainstream news coverage and Twitter activity relating to the flu subsided until a pandemic was declared on May 11th 2009. During this same time period, we collected and stored Twitter posts, and will make use of this source of data for our forecasts.

3 Related work

There have been a number of studies of the effects of news on financial markets. Koppel and Shtrimberg (2004) and Devitt and Ahmad (2007) attempted to use the movement of stock markets as training data to automatically label the sentiment of news articles, implying a relation between news

[1] www.hubdub.com

Table 1: Sample of H1N1-related Tweets.

Tweets	
26 Apr	ya, Im over the Swine flu Tweets. Eat, drink and be merry cuz tomorrow itll be something else killing us.
29 Apr	Whuh oh, the swine flu's Patient Zero in Mexico was flanked by U.S. owned pig farms.
29 Apr	No Americans have died from the swine flu, yet, but every yr 36K Americans die from the regular flu.
11 May	22 confirmed cases of H1N1 Flu in Pima County. The flu appears to be similar to seasonal flu in its impact. Take regular precautions.
12 May	Free bottle of hand sanitizer at work today! No swine flu for me!
19 May	Health UN to discuss swine flu vaccine: UN chief Ban Ki-moon is to meet top pharmaceutical firms to discus..
19 May	New T-Shirt in Harajuku: ”For Beautiful H1N1 Pandemic Life.” I'm off work with sore throat, fever... - shld I buy one?
29 May	thanks to tylenol for reducing my fever... now i'm shedding layers and turning on the AC

sentiment and stock price movements. They had some limited success with this approach, finding that it was easer to detect and label negative stories then positive ones.

Pennock et al. (2000) discussed the relation between artificial markets such as prediction markets and external events, looking at whether the Hollywood Stock Exchange could accurately predict how movies would fare in the real market place. They concluded that in this case, prediction markets were a good indicator of real world events.

Lerman et al. (2008) analysed newswire text to forecast the values of the Iowa Electronic Market for the 2004 US elections. They chose four sets of features: bag-of-words features based on unigram counts; 'news focus' features that track the relative change in unigram feature counts over the preceding 3 days; features for counting sentences that mention predefined named entities such as “Bush”, “Kerry” and “Iraq”; and finally features that label named entities according to their dependency relations. Each of the resulting models were combined with a simple internal market feature. For each day, a logistic regression classifier was trained on the features extracted from about 20 newspaper articles to label the day as closing up or down. If a person were to buy and sell on the recommendation of this system using the best feature combination (news focus + dependency), they would have on average profited about 12 dollars per share over the course of the elections. The Lerman et al. (2008) study differs from the current work in two critical ways. First, it only attempts to classify a day as being up or down, whereas we forecast the closing price for the day. Secondly we are using much more data to make our models, Lerman et al. (2008) uses 20 newspapers per day we are using almost 1 million tweets.

Closely related to the question we are examining, Google developed a system to predict seasonal flu activity based on search queries (Ginsberg et al., 2008). The Google Flu Trend system counts search terms that indicate influenza-like illness activity. They found that there is a strong correlation between these types of search terms and actual influenza infection rates. This correlation was actually a more timely indicator of influenza activity then the traditional surveillance systems used by the US Center for Disease Control and Prevention (CDC) and the European Influenza Surveillance Scheme (EISS). The CDC and EISS both use virological and clinical data as well as physician visits to make influenza forecasts. Using online query data, the Google system was able to predict influenza rates 1–2 weeks ahead of the publication of CDC's US Influenza Sentinel Provider Surveillance Network. Our study looks at this same topic but from a different point of view; while Google Flu Trends is forecasting influenza infection rates, we are forecasting public perceptions of a single influenza outbreak. It is important to forecast the actual infection rate, but it is also useful to forecast public perception of the outbreak, since this gives policy makers insight into the public's mood and fears, as well as valuable marketing data to companies making healthcare products. Also the Google system makes use of a proprietary corpus of search terms, whereas we are using publicly available social media to make our forecast.

4 Approach

4.1 Data

Our corpus consists of Twitter posts that were collected on a daily basis by a crawler from the beginning of April 2009. The data for this experiment is a subset of the corpus, consisting of all Tweets collected during the period April 10th–June 11th. This subcorpus contains 48 million Tweets, on average of 1 million Tweets per day; see table 4.1.

4.2 Classification System

In order to forecast the future prices of the prediction market, we decided to use the Support Vector Machine algorithm to carry out regression. This algorithm was chosen since it can be trained rapidly and can interpret a large feature vector; libSVM Chang and Lin (2001) was chosen as the implementation of the Support Vector Machine regression (SVR) algorithm. In order to make a market forecast for

Table 2: Twitter Corpus Statistics.

Twitter data	
Data Size	10 GB
Word Tokens	703 Million
Total Tweets	48 Million
Tweets per day	1 Million

the upcoming day, the SVR was trained using all extracted features on the prices of the market for all days minus the current one, then the SVR was used to make a regression forecast for the current day see Algorithm 1.

Algorithm 1 Run a Prediction Market Forecast

```
Input: Feature Vector: $X$, Market Prices: $P$
Output: Forecast Vector: $F$, Mean Square Error: $mse$
$ErrorSum \leftarrow 0.0$
for $i \in MarketDates$ do
    for $j \in MarketDates_{i-1}$ do
        SVMTrain($X_j, P_j$);
    end for
    $F_i \leftarrow SVMPredict(X_i)$
    $ErrorSum \leftarrow ErrorSum + (F_i - P_i)^2$
end for
$mse \leftarrow ErrorSum/Count(MarketDates)$
return $F, mse$
```

4.3 Modeling

Prediction Markets can be modelled in two non-exclusive ways:

- *Internal Market.* This treats the task as a time-series problem and models the evolving price just using previous price movements. For example, we might predict today's price as being the average of the previous two days' price. The extent to which this is possible is related to the efficiency of the market.

- *External Market.* This treats the task as price movements being caused by measurable events happening in the world. For example, if a cure for Swine Flu was announced then this event might cause traders to buy (or sell). This event might be mentioned in Tweets.

Within the securities trading domain, market movements are analyzed with two different methodologies. One looks at the fundamentals of the company and the world events that are occurring that will effect the companies share price. The other looks at the historical prices of the security and attempts to forecast the upcoming price based entirely on this historical market data. The latter approach is often called *technical analysis* in the securities trading and was the inspiration for what we are calling 'internal market features'.

Internal Market

For the baseline, we use only data that is internal to the market system; this allows us to assess whether adding external information from the social media leads to measurable improvements in forecasting accuracy.

A simple approach is simply to take a moving average as the forecast value F_n for a given day n. Here, this is the average price of the last 5 days, as shown in equation (1): the last day's average price, $AvgP_{n-1}$, is divided by the sum of the average prices for the last 5 days.

$$F_n = \frac{AvgP_{n-1}}{\sum_{i=2}^{6} AvgP_{n-i}} \tag{1}$$

Using only the moving average is however a fairly poor model for the market, since the moving average will always have a delay in reacting to a change in market price, and will be unable to move proactively with spikes in the price. Furthermore, the moving average by definition always dampendowns movements.

An improvement over a static moving average is instead to have a function which encodes the price history at various levels. In particular, we trained the SVM regression model using extra features:

- The first feature is simply the last day's average price, $AvgP_{n-1}$ for day n, as shown in equation 4.3. This feature give us the short-term history of the market, and helps to capture the quicker market movements and detect local patterns. It is however not suitable for providing evidence for the longer-term and global trends.

$$F_n = AvgP_{n-1}$$

- In order to capture the mid-term trends, the value of the 5 day moving average calculated from the previous day, as shown in equation 1, is used as a feature. This gives the prediction system a longer-term context. In technical analysis, when a price moves above or below the moving average it is often taken as evidence that a new market trend is about to start.

- The final feature provides an indication of the long-term direction of the market. This feature, which we call market momentum and is shown in equation 4.3, is the sum of a vector of binary values M, indicating if the market is above or below the previous day's value. The larger the positive value of this feature, the stronger the upward trend, conversely a large negative value is indicative of a strong downward trend, while values close to zero provides evidence for flat long-term growth.

$$F_n = \sum_{i=0}^{n-1} M_i, M_i = \begin{cases} M_{i-1} + 1 \text{ if } AvgP_i \geq AvgP_{i-1} \\ M_{i-1} - 1 \text{ if } AvgP_i < AvgP_{i-1} \end{cases}$$

This internal market system provides us with a useful baseline against which we assess the added value of analysing data from social media. The baseline is shown in Figure 1.

External market

4.3.1 n-gram Models

A simple model using social media data trains the SVR classifier with unigram and bigrams and their frequencies. After removing stop words and low frequency n-grams, we ended up with a vocabulary of 1,431 unigrams and 347 bigrams.

In the most simple version of the n-gram model, we use the daily counts of the unigrams and bigrams as features for the classifier. No extra processing of the counts is carried out, and no internal market data is given to the classifier.

4.3.2 Combining n-grams with Historical context

We observed that when the SVR was trained with unigram and bigram features over the entirety of the data, lower performance was obtained in comparison with training over a proper subset of the data. For example, immediately after the news about H1N1 broke-out, the volume of Tweets on the topic surged dramatically. Then in the weeks that followed, H1N1-related Tweets declined up until the point when a pandemic was declared on May 11th. This type of pattern was problematic since the initial burst of activity about the topic caused further activity to be 'drowned out'. Consequently when a local spike in the Twitter activity occurred after the declaration of a pandemic, the model was unable to predict it because it was still much lower volume then the global maximum seen when the story first broke-out.

Using a window of only 10 days of Twitter data helped limit the training to data that was relevant to the current news cycle. But windowing suffers from the disadvantage that potentially relevant information from more distant events will be ignored. As a more principled alternative, we added a historical context to each n-gram, so that the system should be able to gauge how the current day's n-gram volume fits into the current news cycle and to the overall history.

To capture historical context, we use four features:

- the n-gram count for the current day;
- n-gram counts for the preceding three days;
- n-gram counts for the preceding week; and
- the total n-gram counts from the start of the collection point to the current day.

For example, the bigram *swine flu* might occur 59 times on May 7th, 242 times from May 4th to May 7th, 389 timse from May 1st to May 7th, and a total of 2,843 timea since the beginning of the data till May 7th. This pattern would be encoded aa the following feature in the classification system: `("swine flu", 59, 242, 389, 2843)`.

The ranges for one, three and seven days were chosen on the basis of empirical trial, and motivated by observation of the duration that news events resonate with the online community.

Mixed Internal and External Market

The internal market model successfully captures the general shape of the curve, as seen in Figure 1, but fails to handle external events that will affect the market in the near future. We attempted to combine the internal market features with the n-gram model, since this should lead to a system were the internal market data can model the general curve and the n-grams can provide cues to external factors that may influence the market. It is hoped that this will help the model to better account for rapid spikes in the data cased by breaking news.

5 Evaluation

We measure the error in the system using *mean square error* (MSE). This is the absolute difference between the actual price and the predicted price, averaged over the lifetime of the market.[2]

This provides a readily understood measure of how far the model's forecast deviates from the actual market price. We also present the actual predicted curves themselves.

6 Results

In Table 6.2, we summarize our results in terms of the total mean square error for each model. It is clear that the moving average model yields the highest error. Training the SVR classifier on the internal market features provides a rather strong baseline; this set of features is simple to extract and requires no external data yet gives a model with a low MSE. The simple unigram and bigram models were the overall worst in terms of error, producing an error that was 29% worse then even the moving average model. The addition of the internal market data to the n-grams results in an improvement in the overall error, reducing the error

[2] The Mean Squared Error encodes just one *scoring rule*; we could equally use some other scoring rule.

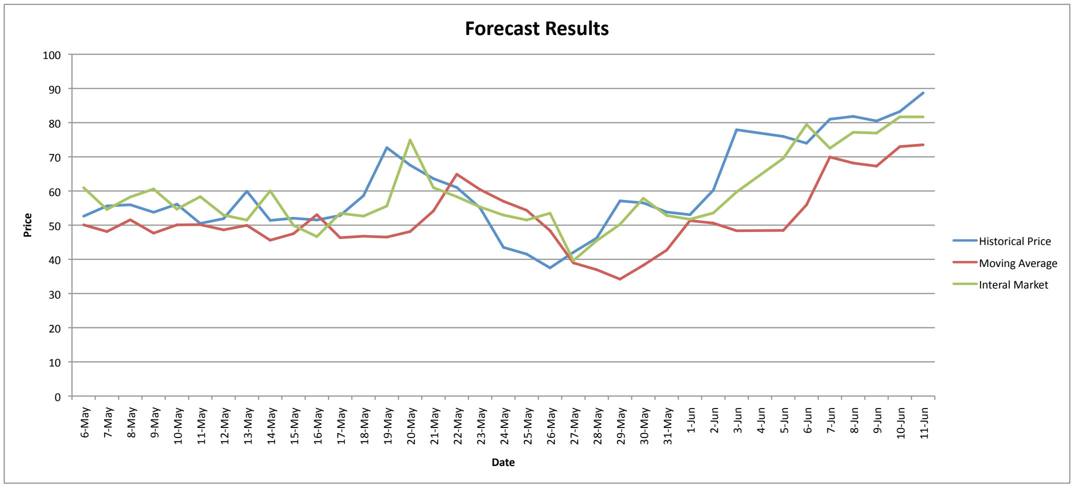

Figure 1: Predicted Prices using Internal Market Indicators

rate to below that of the moving average model but still significantly higher than the baseline. Finally we can see that the error is dramatically reduced by using historical context. The bigram model with historical context features has a 20% lower total error than the baseline. Finally all the features were combined in to a model that used unigrams and bigrams with historical and internal data. This model actually performed slightlly worse. We suspect this is due to overfitting.

6.1 Internal Market

Moving Average

Figure 1 shows results for the model that uses internal market data. It can be seen that using only the moving average as a model performed rather poorly, and was unable to model any of the spikes in the data. Thus, there is a significant spike in the market price on May 19th, yet the moving average is flat till May 23rd and then only shows a slower upward trend. Throughout June the moving average consistently underestimates the price and misses another spike on June 4th.

Internal feature model

The SVR trained on internal market features performed much better and captures many of the spikes. Unfortunately, this model exhibits a lag; thus, the May 19th spike only shows up two days later, on May 21st. We can see this lag again clearly when the small spike on May 14th only appears in the model on May 15th. In general, we can regard the SVR trained on internal market features as providing an acceptable baseline.

6.2 External Market

The n-gram Model

In Figure 2 we show the results of using unigram counts without any internal market data or historical context. This model performed least well, with a mean square error of 313.76 We see that it fails to capture any of the spikes, and does not even reflect the general trend of the market. The model was unstable until May 14th, after which it started to flatten out until June 8th when, counter to the actual price, it shows the market falling sharply. We believe that the model failed to predict the rise from June 1st onwards due to the fact that it had no context for the observed unigram counts. Although the Twitter community at this point was becoming increasingly convinced that H1N1 would indeed become a pandemic, it was now ‘old news’ and less likely to excite comment.

The n-gram Model with History

The best performing model is shown in Figure 4, which results from combining bigrams with the historical context model. The total mean square error of 40.67 (see Table 6.2) beats the best internal market baseline by 20%. We see a sharp reduction in error compared to the n-gram based approaches that lack historical context features. Almost all the spikes and dips in the market were captured, with the curve becoming slightly too flat after June 2nd. As with the internal market model, there is still a lag of about one day lag in the model. For example, the May 19th spike is accurately reflected, but only on the followind day. Reducing this lag is an task that we plan to address in future work.

6.3 Mixed Internal and External Markets

The model that combined the n-gram counts with the internal market features resulted in little improvement. Figure 3

Table 3: Results for H1N1 Pandemic Forecast.

Feature	**Mean Square Error**
Moving Average	276.33
Internal Market (baseline)	51.81
Unigram	313.76
Bigram	388.54
Unigram + Internal	273.45
Bigram + Internal	210.89
Unigram + History	62.84
Bigram + History	40.67
Unigram + Bigram + History + Internal	54.03

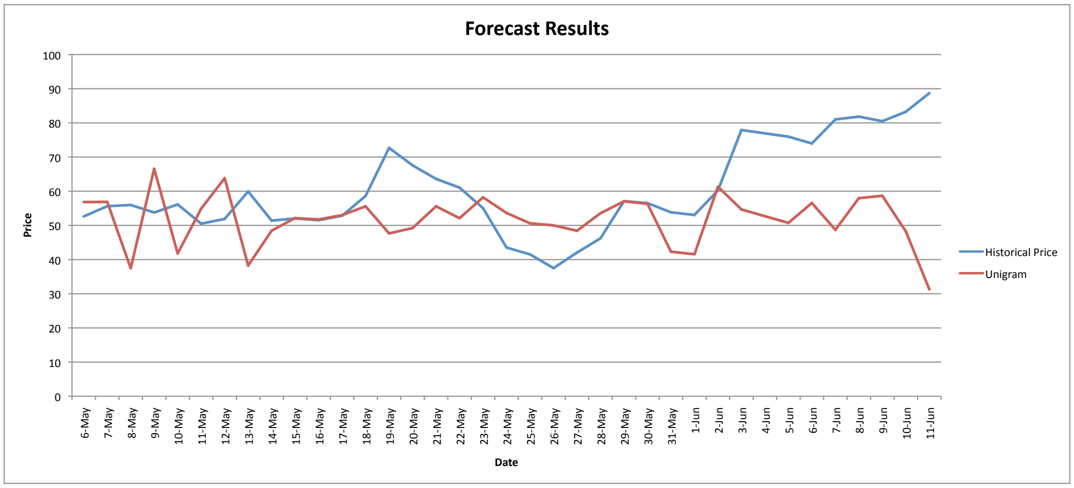

Figure 2: Predicted Prices using Unigram Counts

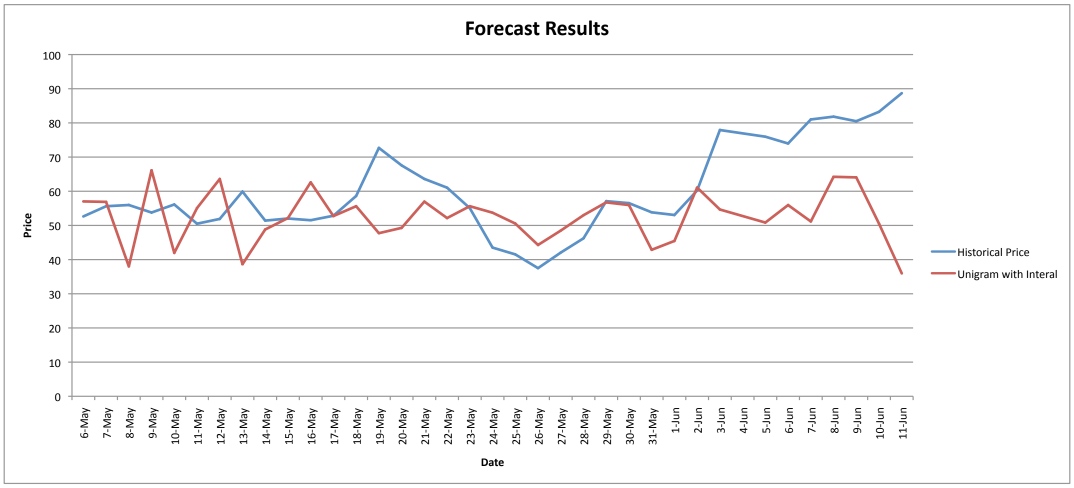

Figure 3: Predicted Prices using Unigram Counts with Internal Market Indicators

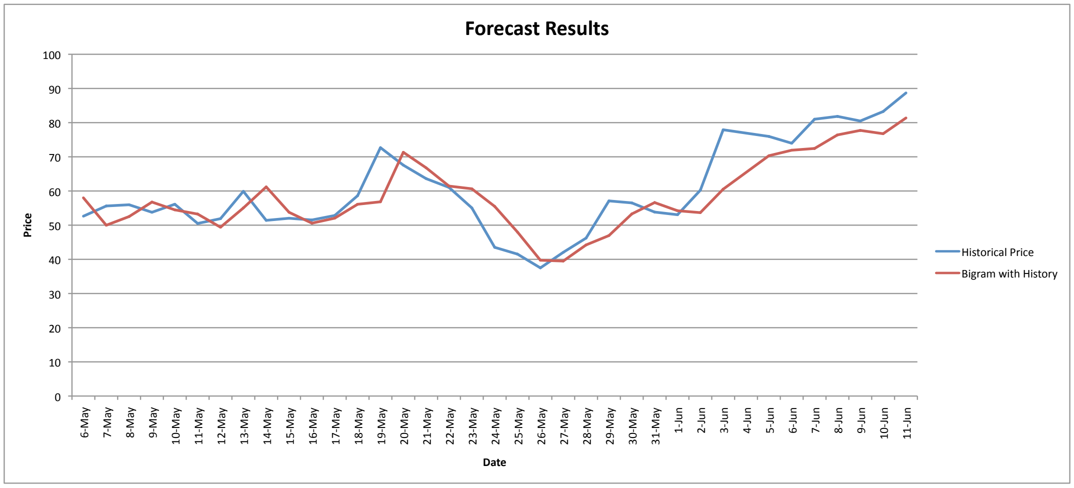

Figure 4: Predicted prices for Bigram Counts with Historical Context Features

closely resembles Figure 2, suggesting that the unigram feature outweighed other information. We see that adding the internal market features to the model lowers the MSE by 40.31, but this is due to the curve being shifted closer to the mean price of the market, rather than any improvement in modeling the spikes or decrease in the lag.

7 Conclusions

We have demonstrated that adding features concerning the historical context of the current day's feature counts has a markedly beneficial effect on the forecast accuracy. By adding mid- and long-range context features, we are better able to exploit all the available data. This also achieved better results than combining internal and external features. These initial results are encouraging and suggest that information present in noisy social media such as Twitter can be used as a proxy for public opinions. Future work will look more closely at the relationship between social media, prediction markets and time series modeling.

References

C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

CNN. Swine flu creates controversy on Twitter, 2009. URL http://edition.cnn.com/2009/TECH/04/27/swine.flu.twitter/index.html.

A. Devitt and K. Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 984–991, Prague, Czech Republic, 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P07-1124.

J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.

S. Gjerstad. Risk aversion, beliefs, and prediction market equilibrium. In *Annual Meeting of the Allied Social Science Associations*, Boston, MA, Jan 2006. URL http://www.aeaweb.org/annual_mtg_papers/2006/0106_1015_0701.pdf.

M. Koppel and I. Shtrimberg. Good news or bad news? Let the market decide. *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pages 86–88, 2004.

K. Lerman, A. Gilder, M. Dredze, and F. Pereira. Reading the markets: Forecasting public opinion of political candidates by news analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 473–480, Manchester, UK, August 2008.

D. Pennock, S. Lawrence, C. Giles, and F. Nielsen. The power of play: Efficiency and forecast accuracy in web market games. Technical report, NEC Research Institute, Jan 2000.

J. H. Watkins. Prediction markets as an aggregation mechanism for collective intelligence. Technical report, Human Complex Systems, University of California, Los Angeles, 2007. URL http://repositories.cdlib.org/hcs/WorkingPapers2/JHW2007.

J. Wolfers and E. Zitzewitz. Interpreting prediction market prices as probabilities. Working paper 12200, National Bureau of Economic Research, 2004.