Predicting Collective Sentiment Dynamics from Time-series Social Media

Le T. Nguyen, Pang Wu, William Chan
Carnegie Mellon University, Silicon Valley Campus
NASA Research Park, Bldg. 23
Moffett Field, CA 94035
{le.nguyen, pang.wu, will.chan}@sv.cmu.edu

Wei Peng
Xerox Innovation Group
Xerox Corporation
Rochester, NY, 14580
wei.peng@xerox.com

Ying Zhang
Carnegie Mellon University, Silicon Valley Campus
NASA Research Park, Bldg. 23
Moffett Field, CA 94035
joy@cs.cmu.edu

ABSTRACT

More and more people express their opinions on social media such as Facebook and Twitter. Predictive analysis on social media time-series allows the stake-holders to leverage this immediate, accessible and vast reachable communication channel to react and proact against the public opinion. In particular, understanding and predicting the sentiment change of the public opinions will allow business and government agencies to react against negative sentiment and design strategies such as dispelling rumors and post balanced messages to revert the public opinion. In this paper, we present a strategy of building statistical models from the social media dynamics to predict collective sentiment dynamics. We model the collective sentiment change without delving into micro analysis of individual tweets or users and their corresponding low level network structures. Experiments on large-scale Twitter data show that the model can achieve above 85% accuracy on directional sentiment prediction.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*human information processing*

General Terms

Human Factors

Keywords

Sentiment prediction, sentiment analysis, social network analysis

WISDOM '12, August 12 2012, Beijing, China

1. INTRODUCTION

The rapid growth of online social network sites (e.g., Facebook, Twitter, LinkedIn, and Google+) and their public available data acquiring API has led the prosperity of social network analysis research in recent years. Since more and more people share their opinions and join various activities on social networks, enterprises and government agencies have sought the opportunities to leverage this data for intelligence applications such as enterprise marketing services, customer relationship management and public opinion management. It has become very important for enterprises to unlock customer sentiment embedded in the huge amount of social media data so that they can quickly respond to complaints and improve their product quality. It is even more critical and ideal to predict customers' sentiment change ahead of time to help enterprises quickly identify the root causes, therefore preventing huge negative message cascades. Similarly for government agencies, it is critical to understand the sentiment in the public's opinion and react promptly to manage crisis such as the accident of the Fukushima Daiichi nuclear disaster by dispelling rumors and clarifying facts to change the dynamics of the social media.

Both academia and industry have been exploring to harness the wealth of social media data not only as a reactive analytics tool but also as a predictive analytics tool. For example, Asur and Huberman from HP Labs successfully forecast box-office revenues from movies using twitter data, which outperforms market-based predictors [1]. Work from [9] shows that even using a relatively simple sentiment detector based on Twitter data, the result can replicate consumer confidence and presidential job approval polls. This result delivers an encouraging message that expensive and time-intensive polling can be supplemented or supplanted with the simple-to-gather text data that is generated from online social networking, which reveals the huge application potential on social media sentiment analysis.

The above mentioned existing works somewhat correlate the collective sentiment/mood mined from large-scale social media data with their prediction objectives. Different from the existing work, this paper aims to predict the sentiment/mood change toward particular products/brands at some point in the future. It does not focus on the senti-

ment analysis of each individual tweet; but rather focused on predicting the aggregated global population sentiment ratio and its transformation through time. Thus it can help enterprises monitor their brand perception and provide effective customer care. It can also be employed to identify and predict critical social sentiment inflection points such as the London riots or the Occupy Wall Street protests. The predicted sentiment resulted from this work can be used as a feature for other social media predictive analysis purpose.

In this paper, we develop a statistical model to predict the sentiment change in the social media and to address the following questions:

- How long back to the tweet history is most appropriate to learn a sentiment prediction model?
- How long does it take for the social media to demonstrate its response (sentiment change) after certain dynamics/events/activities occur?
- How long does the response on social media last?

Additionally, we introduce three parameters: *history window size*, *prediction bandwidth*, and *response time*, and discover how they would influence the sentiment prediction quality. Comprehensive experiments are conducted to evaluate our sentiment prediction model on large-scale twitter data.

The remainder of this paper is organized as follows. Section 2 reviews related work in the area of social media prediction and sentiment analysis. Section 3 defines the research problem and introduces the prediction model. In Section 4, we describe the experiment conducted and evaluate the results, followed by discussion and conclusion in Section 5.

2. RELATED WORK

There are many research studies on using social media to predict values in real-world and on analyzing sentiment in social media. To our best knowledge, we are not aware of any work studying how and why sentiment changes over time in social media and if social media dynamics can predict such changes.

2.1 Predictive analysis on social media

Many works tried to use social media content to predict real-world outcomes [6, 1]. As one of the early works to leverage social media for future prediction, Gruhl et al [6] explored the correlation of mentioning rate of products in online chatter posts and its sales spikes. Their analyze shows that volume of blog postings can be used to predict spikes in actual consumer purchase decisions at online retailer Amazon. They also constructed a simple predictor based on the mention volume in blog posts of certain products to predict sales rate. However, they didn't analyze how sentiment in blog post can impact the consumer's decision.

Asur et al [1] demonstrated how social media content can be used to predict real-world outcomes. They use the chatter from Twitter.com and constructed a linear regression model for predicting box-office revenues of movies in advance of their release. Their experiments also showed that the results outperformed in accuracy those of the Hollywood Stock Exchange and that there is a strong correlation between the amount of attention a given topic has (in this case a forthcoming movie) and its ranking in the future. They also compare the prediction performance by using only tweet-rate and both tweet-rate and sentiment ratio as features, which shows sentiment can further improve the prediction.

Researchers also studied the sentiment distribution in social network. Tan et al [15] studied the user-level sentiment distribution in social network based on user's connection. Their result shows that connected users tend to hold the same sentiment, and, two users with the same sentiment are more likely to have at least one link to the other than two users with different sentiment. Their model demonstrated that user-level sentiment analysis can be significantly improved by incorporating link information from a social network.

Additionally, researchers have found that there does not seem to be any features of a user or environment that can reliably predict *virality* of social media at birth, although early growth patterns relative to peers might [2]. [14] suggested that the best predictor of virality is virality itself; the rich get richer, even if the early riches are completely accidental.

2.2 Sentiment analysis

There are many works trying to address the problem of automatic sentiment analysis using machine learning or other techniques. There are two major tasks in the sentiment analysis. The first one is called sentiment detection, which classifies the text into subjective or objective. The second task is called polarity classification: given a piece of text with opinion, the goal is to classify the sentiment to one of two opposite polarities, i.e., positive or negative. These two tasks can be done on different levels, and there are different techniques for different level as well: n-gram [16] and lexicon are usually used on term level while Part-of-Speech [11] works for sentence and phase analysis.

Fundamentally, sentiment classification starts with identifying the semantic orientation of words, then goes to higher level text structure like the semantic orientation of sentences and documents. Several techniques are used to achieve this task: Words were directly weighted by lexicons of semantic words which were manually or automatically constructed. Most of the manually constructed lexicons are extensions of the general purpose ones [4]. Statistical analysis such as word co-occurrence also can provide efficient approaches to infer semantic orientation of words. In addition, a variety of training data labeled manually can help to perform sentiment classification. Therefore, the popular algorithms in machine learning, such as support vector machines [12] and Naive Bayes [17], are used to train the sentiment classifier of words and sentences.

Meanwhile, by simply combining the polarities of all words [4], a document can only have two possible polarities, and no extreme opinion exists. Beside the positive and negative category, mixed opinions are classified by introducing threshold values in identification [8]. Only considering traditional bag-of-words features, some misleading texts may drop down the performance of polarity classification. A minimum cut formulation that integrates cross-sentence contextual information, is applied to just the subjective portions of the document. By utilizing contextual information, the accuracy of sentiment analysis can be significantly improved [10].

3. SENTIMENT TIME-SERIES PREDICTION

There are two main objectives of the sentiment time series

prediction, namely, 1) To *predict* the change of sentiment of a given topic over time; 2) To *identify* key features that contribute to the change of sentiment.

In this section we first define the "change of sentiment" and describe the statistical model to predict the change based on features extracted from the time-series social media dynamics.

3.1 Sentiment Change

The goal of this work is to predict the sentiment change over time rather than the absolute sentiment values (e.g., the change of number of positive tweets at a certain time). We quantify the dynamics of the sentiment in social media through measuring the ratio r between positive tweets and tweets with either positive or negative polarity for a particular time interval. r is defined as:

$$r = \frac{\#tweets^{+}}{\#tweets^{+} + \#tweets^{-}} \quad (1)$$

r ranges between 0 and 1. It is 0 when there are no positive tweets and it is 1 when there are no negative tweets at a certain time slice. Positive and negative tweets are classified by the sentiment analysis described in more details in Section 4. r is a ratio and it does not depend on the absolute number of tweets. This is important as we are comparing the sentiment changes over multiple topics (namely, iPhone, Android and Blackberry) and different topics have different number of tweets (Figure 3), trying to create a model predicting absolute values for iPhone might for example not work for Android or for other domains such as politics, etc. Figure 1 shows the r ratio of iPhone over 7 days (168 hours).

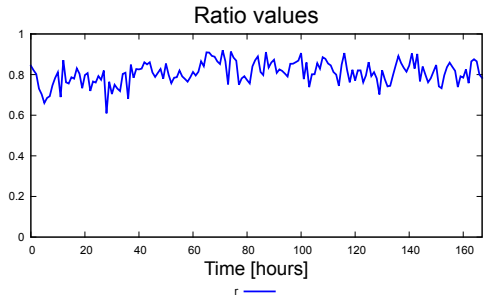

Figure 1: Ratio between positive tweets and tweets with either positive or negative polarity on the "iPhone" topic over 7 days (168 hours). X-axis is time (by hours) and Y-axis shows the r value.

We model the sentiment change prediction problem as a *classification* problem. By modeling the social media dynamics, the prediction model predicts whether the positive ratio r will "go up", "go down" or "remain relatively unchanged" at a future time.

3.2 Social Media Dynamics

We extract about 80 features from the social media to model its dynamics. Table 1 lists major features used in this study.

3.3 Prediction

Feature Type	Feature Example
Tweets	Sentiment of the tweet Number of being marked as *favorite* Number of retweets
User	Number of followers Number of friends Number of posted statuses Number of lists a user belongs to
Sentiment Ratio	#positive : #negative tweets #positive : #(positive+negative) tweets #negative : #(positive+negative) tweets #neutral : #(positive+negative) tweets #(positive+negative) : #all tweets #neutral : #all tweets
Dynamics	First and second order derivatives of all above features

Table 1: Features extracted from the social media time series to model the dynamics of sentiment.

The goal of this research is to predict the sentiment dynamics in social media in the future. The prediction process is conditioned on three random variables, namely, **history window size** α, **prediction bandwidth** β and **response time** γ (as shown in Figure 2). Our prediction model uses features extracted from the history window α and predict the sentiment changes in a future window β which is after "response time" γ from now.

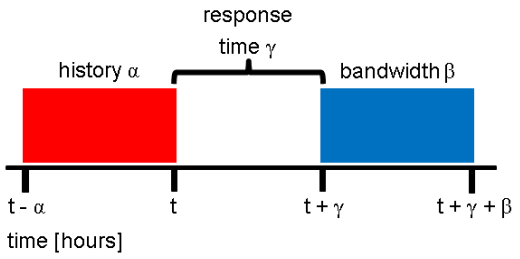

Figure 2: Parameters of the prediction: history window size α, prediction bandwidth β and response time γ. Prediction model extracts features from history window and predict the sentiment change of the social media in a future window of size β which is γ hours after the current time t.

- *History window size*: History window size α indicates the length of history data sequence that is used for the prediction. For example, if $t = 12\ pm$ and $\alpha = 1\ hour$ then we would take all the data collected from 11 am to 12 pm, extract the features and use them for the prediction.

- *Prediction bandwidth*: From Figure 1 we can observe that the ratio value contains a significant portion of noise. With the prediction bandwidth β we can adjust the smoothing process, which is used for extracting the noise. For example, by specifying $\beta = 6\ hour$, we are basically calculating the ratio value over the next 6 hour instead of using only 1 hour time interval. Thus,

we can extract the outliers from the data and avoid overfitting while training our machine learning model.

- *Response time*: The response time γ indicates the time interval between a certain action and its observable effect. For example, if we would have a certain critical information that might cause the change of the collective sentiment and we would tweet it at 1 am, it is most likely that people we read our tweet in the morning. In that case the response time is about 8 hours.

4. EXPERIMENTS AND EVALUATION

We collected Twitter data from its 10% Gardenhose API over a period of 5 months and ran sentiment analysis over tweets that contain keywords: "android", "blackberry" and "iphone". After data preprocessing, we apply the state of the art sentiment analysis tools over the filtered data and measure the aggregated sentiment and investigate different classification models to predict the sentiment changes in the future.

4.1 Data

Twitter provides Streaming APIs which allows high-throughput near-realtime access to various subsets of Twitter data. It samples the statuses (including the tweets and the authors) from the Firehose stream of public statuses which is the full feed of all public tweets. Our paper uses Twitter Gardenhose streaming API, which is said to sample 10% of all public tweets. In our work, Twitter data are collected from January 2011 to May 2011, from all languages and regions; however only English tweets were analyzed — in total there were 12 million tweets collected of which 7 million were English. Of the English Tweet corpus; we focused on only tweets that contained at least one of these words: "android" (1 million tweets), "blackberry" (0.8 million tweets), or "iphone" (2.5 million tweets) or one of their inflected forms such as plural. Each Tweet included basic information such as timestamp and the anonymized posting user id.

Figure 3 shows the distribution of tweets across different topics in our data. X-axis indicates the time in hours starting from January 2011. Y-axis shows the flow percentage of each individual topic. Overall, the most popular topic during this period of time was iPhone followed by Android and Blackberry yet from time to time there are spikes in Twitter on Android. For example, around April 2011 (hour 3200 in Figure 3), there was a big discussion about Android. This could be caused by the Google I/O event or the rumor of the Apple's iPad2 which generated heated debate among Apple fans and Android fans.

The raw tweets acquired from Twitter API include many languages such as English, Chinese and Japanese. However, the goal of our project is to focus only English sentiment analysis. We extracted and removed tweets that were not English. We use the language detection tools powered by *Cybozu Labs* [3], which employs Naive Bayes to classify documents into different language categories and with an accuracy of approximately 99%.

4.2 Feature extraction

The goal of this research is to avoid micro-analysis of individual tweets or users and their corresponding complex low level network structures. Rather, we wish to utilize an antagonistic approach; we desire to perform "aggregated statistical analysis" at the time series level. The smallest level of granularity is defined by a timeslice, which we set the length of to be one hour. For a random variable (e.g., user follower count), its value is aggregated over all users' activities in a particular timeslice. For each time slice we extracted a feature vector by aggregating all data of the tweets collected in that particular time slice. Table 1 lists all types of features extracted from the Twitter data.

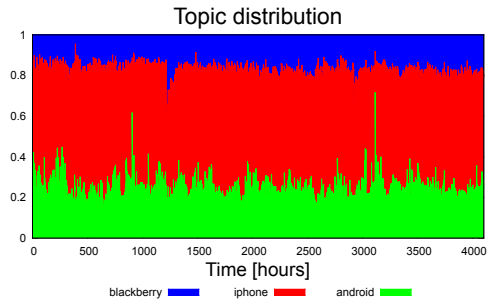

Figure 3: Distribution of tweets across different topics

4.3 Sentiment analysis

We evaluated two sentiment analysis algorithms: machine-learning-based Dynamic Language Model (DynamicLM) [10] and lexicon-based Constrained Symmetric Nonnegative Matrix Factorization (CSNMF) [13]. DynamicLM uses a labeled training data set to train a language model, which is then used for estimating a sentiment label of unseen data. In contrast, CSNMF constructs the adjective relation graph by combining both WordNet and conjunction rules directly extracted from social media data. It can be considered as a "semi-supervised" clustering, to take advantages of both 'attraction' and 'repulsion' between adjectives to better assign sentiment strength scores to the adjectives.

For the evaluation we used the data set of labeled Digg comments [18], which includes 890 positive and 1,065 negative comments. In the experiments, CSNMF achieved 79% accuracy while the DynamicLM model achieved only an accuracy of 60%. DynamicLM is typically used for analyzing the sentiment of "long" documents. Therefore, it is less suitable for social media sentiment analysis, where the documents have usually a short length and where the words are rather informal. On the other hand, CSNMF's aim is to generate lexicon specifically for social media data. The Digg comments have similar properties (such as length, language style, etc.) as the tweets used in our project, we use CSNMF for obtaining the sentiment labels of the tweet data set.

Figure 4 shows the distribution of the tweet sentiment across different topics. In each topic most of the tweets are neutral. The number of positive tweets tend to be much higher than number of negative tweets.

While analyzing the sentiment classification algorithms in detail, we observe many challenges related to the sentiment classification of the tweets. Since tweets have only a limited number of characters, they do not typically contain the user's contextual information, which are essential for classifying the sentiment. For example, the tweet "does anyone like the new iPhone 4?" might be classified either as nega-

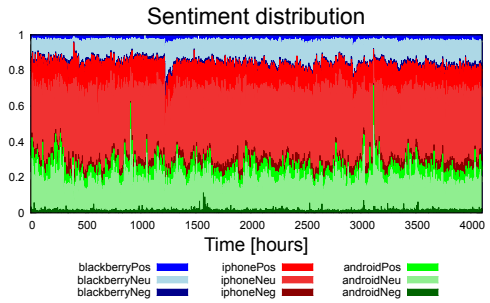

Figure 4: Sentiment distribution across different topics

tive or neutral, depending on the context of the Twitter user. The user might be an owner of the new phone device and expresses a negative opinion through this tweet. The user could be also a person who is considering buying the new phone device, but wants to survey a public opinion about the product first.

Another tweet example shows the complexity of the classification process: "maybe some day when i'm as cool as you and have an iphone so i can spend all day following people". Clearly, this tweet expresses sarcasm and should have a negative label. However, what the system observes is a sentence which contains a positive word "cool" and does not contain any negative expressions. The sentiment classifier would very likely mark this tweet as positive or neutral, but not as negative.

Another challenge would be identifying the relationship between the sentiment of the expression relatively to the actual subject of interest. Assuming our system could recognize the sarcasm, it should not mark the sentence as negative if the subject of interest is the keyword "iphone". Obviously, the Twitter user wanted to express a negative statement about the other person, rather than about the mobile device. Many tweets such as "i love playing angry birds on my iphone" are relatively difficult even for human annotators to label their sentiment. In this example, is the Twitter user positive about the mobile phone or about the game he or she is playing?

In order to determine the correct sentiment label for each individual tweet, the system would need to "understand" the language as human do, which is still not possible with the state-of-the-art sentiment classification. In this work, we are interested in the aggregated global sentiment ratio. The sentiment classifier might have relatively high error rate on the individual tweet level. However, on the global level with a large data set the errors tend to cancel out as pointed out by O'Connor et al. [9]

4.4 Prediction

We experimented with various combinations of parameters to determine the optimal set of parameters. These values provided us with insightful information about the data that we used in our research project.

As the baseline for the comparison we used a simple approach of predicting the change of future sentiment ratio using heuristics. The approach is based on the idea that the growth or decline of the ratio values has a certain momentum that forces it to that keep the direction of the sentiment change. For example, when the number of positive tweets is currently growing faster than the number of negative tweets, then the baseline approach assumes that due to the momentum the ratio values will be also growing in the next hour. In order to generalize this baseline approach we also integrated all the three prediction parameters (described in Section 3.3) into our baseline algorithm. Thus, similarly to our machine learning model, when we set the response time to 8 hours for the baseline, our goal would be to predict whether the ratio value in 8 hours is higher or lower than the current ratio value.

In the following experiments we sorted the tweets based on time and divided them into two timely non-overlapping data sets A and B with an equal number of tweets. The dataset A contains tweets collected from January 2011 to March 2012 and dataset B contains tweet collected from March 2012 to May 2011. The dataset A was used for training a Support Vector Machine and the data set B was used to evaluate the trained model. In the following, we report F1 scores of experiments with various parameter configurations.

4.4.1 *Impact of history length*

Figure 5 shows the results of the configuration [α, β, γ] = [x,1,0] where we set the prediction bandwidth to 1 hour ($\beta = 1$) with an immediate response time ($\gamma = 0$) and test different the history window sizes ($\alpha = x$, where x represents a set of tested values). By increasing the history window size the size of our feature vector also increases, since we are considering the history as a sequence of data and we do not aggregate feature values at different timeslices.

In Figure 5, the x-axis represents the increasing value of the history window size and y-axis represents the F1 score. We hypothesize that with the longer history window, there should be more data for the prediction and therefore the prediction accuracy should be higher given abundant evidence in history. However, the results show that the prediction accuracy increases with the history window size only to a certain threshold. While using a larger history window the F1 score decreases. This indicates that events occurred in the past have decayed impact on future sentiment. Using features extracted from too large history window will suppress important features that happened immediately before the prediction time.

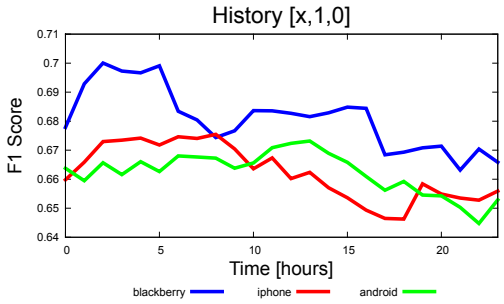

Figure 5: Testing different values of the *history window size*.

Figure 6 compares the results of SVM with the results of the baseline approach. We were able to almost double our

prediction accuracy from 35% with baseline to almost 70% F1 score with our approach.

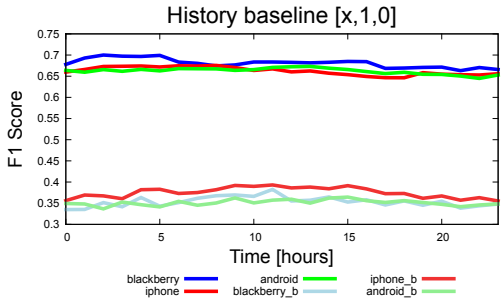

Figure 6: Prediction accuracy vs. *history window size* comparing SVM with the naive baseline approach (where *_b indicates the result of the baseline)

4.4.2 Bandwidth of prediction

The result of the configuration [1,x,0] is shown in the Figure 7. In this experiment we set the history window size to 1 with an immediate response time and tested different values of the prediction bandwidth. As mentioned in the previous section by setting the bandwidth window size we specify the range of prediction target. In other words, through this parameter specify the granularity of our predictions. The higher bandwidth, the more coarse-grained is our prediction. E.g. by setting prediction bandwidth to 6 hours, our model will try to predict the ratio value over the next 6 hours. Thus, our predictions are less sensitive to the noise in the dataset.

As expected, with the increasing bandwidth we get better prediction results. When consider an extreme case and set the bandwidth to 1 year then our model will try to predict a ratio value over the next 1 year. Since the bandwidth is large the ratio value remains almost a constant. This makes it easy for the machine learning model to predict the correct results with a high certainty.

From the Figure 7 we can observe two significant values of the bandwidth window sizes: 12 and 24 hours. The F1 score increases significantly from 0 to 12 hours. Between 12 and 24 hours we can still observe a certain increase of F1 score. However, by setting the bandwidth with a value higher than 24 hours we can acchieve only an insignificant improvements. From the results we can infer that our model performs the best at predicting the sentiment dynamics occurring in the next 12 to 24 hours.

With the baseline approach the highest F1 score is around 65% (as shown in Figure 8), which is about 20% lower than what can be achieved by applying our machine learning model.

4.4.3 Response time

In the next experiment we tested different values of the response time γ. Response time measures how fast can social media respond to certain events happened in history. Figure 9 shows the results of the configuration [1,1,x] where we set both history and bandwidth to 1 hour and increased the response time. From the results we can infer that the

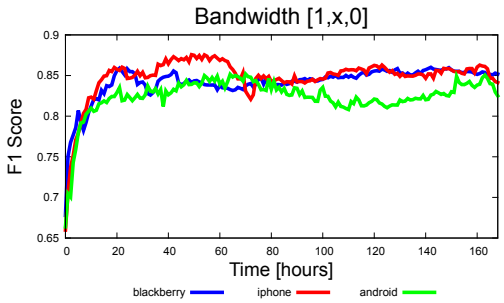

Figure 7: Testing different values of the *prediction bandwidth*.

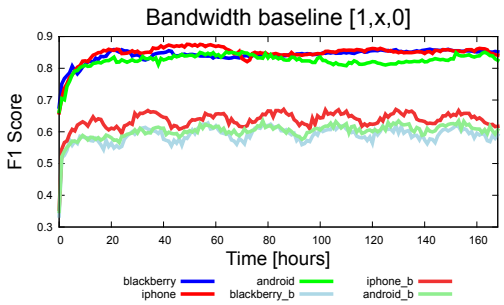

Figure 8: Testing different values of the *prediction bandwidth* and comparing them with the baseline approach (where *_b indicates the result of the baseline).

best prediction can be achieved with the response time of 12 hours. Moreover, we observe an interesting pattern in the graph that shows that the F1 scores have local maxima at 12, 36, 60, 84 hours, etc., which corresponds to 12 hours and N days where N start from 0. This makes us to believe that there is a certain underlying sentiment daytime pattern our Twitter data. Golder et al. [5] discovered that people tend to be more positive in the morning hours and change their sentiment towards the end of the day. This might be the explanation the phenomenon observed in Figure 9. Since the sentiment follows a certain daytime pattern our model was able to capture this pattern and use it in order to achieve a more reliable prediction.

Figure 10 depicts results of the baseline approach compared to the results of our machine learning model. It is obvious from the figure that the heuristical baseline approach was not able to achieve results better than 40%. On the other hand, our approach was able to almost double to prediction accuracy, reaching the F1 score of 75%.

4.4.4 Classification models

Besides the experiments with different prediction parameters we also tested the various machine learning techniques. We selected the topic iPhone and used the above configuration [1,1,x] in this experiment. In Figure 11 we compared the F1 score of SVM, logistic regression and decision tree. SVM and logistic regression have a similar result and outperform

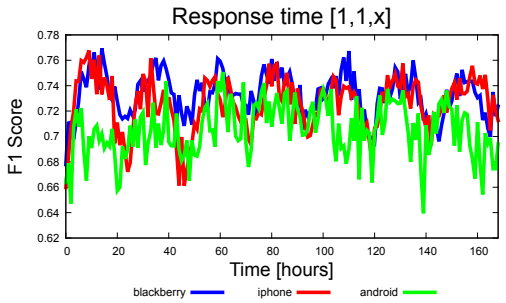

Figure 9: Testing different values of the *response time*.

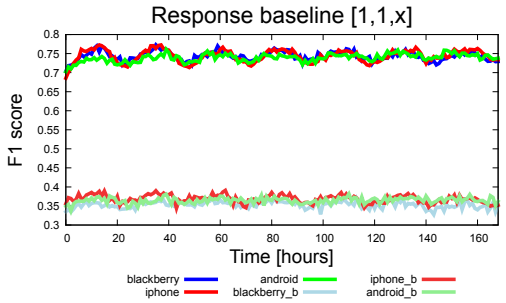

Figure 10: Testing different values of the *response time* and comparing them with the baseline approach (where *_b indicates the result of the baseline).

the decision tree.

4.5 Significant features

The second objective in our research is to identify key features that contribute to the change of sentiment. We used Weka's InfoGainAttributeEval algorithm to evaluate the importance of an attribute by measuring the information gain with respect to the output class [7]. The most significant features are: 1) Φ_1: Ratio between positive and negative tweets. 2) Φ_2: Ratio between positive and positive+negative tweets. 3) Φ_3: Ratio between negative and positive+negative tweets. 4) Φ_4: First order derivative of Φ_1. 5) Φ_5: First order derivative of Φ_2. 6) Φ_6: First order derivative of Φ_3.

Table 2 shows the weight of each feature in various configurations of [history, bandwidth, response time] parameters. The higher the weight the more important is the specific feature. From the results we can observe that the 3 ratio values with their corresponding first order derivatives contribute the most to the change of sentiment. This also explains, why the baseline described in Section 4.4 performed poorly. The baseline approach assumes that only the first order derivate of the ratio between positive and positive+negative tweets are essential for the prediction. However, in practice the change of sentiment is influenced by a set of multiple features and the sentiment evolution cannot be predicted by using one individual feature.

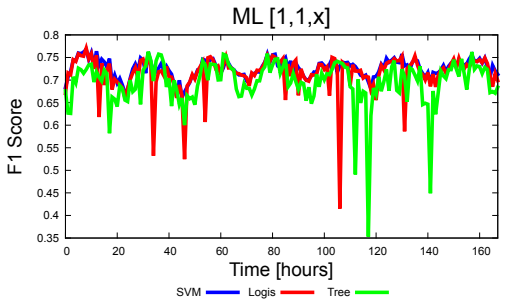

Figure 11: Testing the prediction results for different machine learning algorithms: SVM, Logistic Regression and Decision Tree

features	[1,1,0]	[1,1,12]	[1,12,0]	[12,1,0]
Φ_1	0.119	0.273	0.393	0.119
Φ_2	0.119	0.273	0.393	0.119
Φ_3	0.119	0.271	0.393	0.119
Φ_4	0.082	0.087	0.129	0.086
Φ_5	0.085	0.057	0.104	0.086
Φ_6	0.085	0.057	0.104	0.086

Table 2: Importance of each feature in various configuration parameters

4.6 Multi-class classification

We use a multi-class classification approach to predict the future sentiment direction and ratio quantity. Instead of vanilla binary classification where we predict the sentiment ration r will go up or down, we predict the range of sentiment change. Define sentiment change ratio $X = \delta r/r$.

For example, we can quantize the sentiment change ratio into 5-class by:

Class	
-10	if $-0.05 > X$
-5	$0.00 > X > -0.05$
0	$X \approx 0$
5	$0.05 > X > 0.00$
10	$X > 0.05$

Multi-class classification experimentation was done with SVM with a feature history window of 2 hours, a response time of 12 hours and a response bandwidth of 12 hours: Table 3 gives experimental accuracy results. Multi-class results were obtained via the classical SVM majority voting system.

The most noticeable phenomenon from Table 3 is that as the number of classes increases the accuracy decreases. This is expected; as the number of classes increases the com-

Classes	Accuracy
-3,0,3	72.33%
-5,0,5	74.93%
-5,-3,0,3,5	63.66%
-10,-5,0,5,10	63.34%
-7,-5,-3,0,3,5,7	55.65%

Table 3: Multiclass SVM Experimental Results

Classified As ⇒	−10	−5	0	5	10
-10	284	73	142	0	4
-5	85	95	409	0	3
0	15	41	1802	14	41
5	0	0	387	8	113
10	0	0	143	17	398

Table 4: Multiclass SVM Experimental Results – Classification Matrix for -10,-5,0,5,10

plexity of the classification problem increases and subject to more errors in the SVM voting mechanism.

However, it is interesting if we look at an in-depth analysis of the SVM classification results. Table 4 gives the actual classification for the $-10, -5, 0, 5, 10$ scenario. Take a closer look at Table 4 and we can see that although the SVM multi-class classification is not very good at determining the quantity of sentiment change; it is still very good at determining the direction of sentiment change. A slight note of positive is that although the classifier may quantify the amount of sentiment change wrong; rarely does it classify in the opposite direction.

5. CONCLUSION

In this paper, we analyzed a large collection of Twitter data in order to gain a deep understanding about the sentiment evolution in the social network. We developed a machine learning model to predict the change of sentiment of a given topic over time. Additionally, we identified the key features that contribute to the change of sentiment. The results of our evaluation show that we are able to predict directional sentiment ratio change with accuracy above 85% using SVM. Using multiclass SVM we can achieve an accuracy around 55% to 70% depending on the granularity of sentiment quantity classification desired.

6. REFERENCES

[1] S. Asur and B. A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 492–499, Washington, DC, USA, 2010. IEEE Computer Society.

[2] R. Colbaugh, K. Glass, and P. Ormerod. Predictability and prediction for an experimental cultural market. *Advances in Social Computing*, 6007, 2010.

[3] Cybozu Labs. Language detection library for java. http://www.slideshare.net/shuyo/language-detection-library-for-java.

[4] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. pages 519–528, 2003.

[5] S. A. Golder and M. W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

[6] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 78–87, New York, NY, USA, 2005. ACM.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, November 2009.

[8] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC-2007 Blog Track. In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings*, 2007.

[9] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010.

[10] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, pages 271–278, 2004.

[11] B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, 2:1–135, Jan. 2008.

[12] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *IN PROCEEDINGS OF EMNLP*, pages 79–86, 2002.

[13] W. Peng and D. H. Park. Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2011.

[14] M. Salganik, P. Dodds, and D. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762), 2006.

[15] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1397–1405, New York, NY, USA, 2011. ACM.

[16] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. *Comput. Linguist.*, 30:277–308, Sept. 2004.

[17] J. M. Wiebe, R. F. Bruce, and T. P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications, 1999.

[18] R. Zafarani and H. Liu. Social computing data repository at ASU, 2009.