Polarity Trend Analysis of Public Sentiment on YouTube

Amar Krishna†, Joseph Zambreno and Sandeep Krishnan
Dept. of Electrical and Computer Engineering, †Dept. of Computer Science
Iowa State University, Ames, IA 50011
{a16, zambreno, sandeepk}@iastate.edu

ABSTRACT

For the past several years YouTube has been by far the largest user-driven online video provider. While many of these videos contain a significant number of user comments, little work has been done to date in extracting trends from these comments because of their low information consistency and quality. In this paper we perform sentiment analysis of the YouTube comments related to popular topics using machine learning techniques. We demonstrate that an analysis of the sentiments to identify their trends, seasonality and forecasts can provide a clear picture of the influence of real-world events on user sentiments.

1. INTRODUCTION

With the rapid growth of social networking and the Internet in general, YouTube has become by far the most widely used video-sharing service. Current YouTube usage statistics indicate the approximate scale of the site: at the time of this writing there are more than 1 billion unique users viewing video content, watching over 6 billion hours of video each month [1]. It is important to note that YouTube provides more than just video sharing; beyond uploading and viewing videos, users can subscribe to video channels, and can interact with other users through comments. This user-to-user social aspect of YouTube (the YouTube social network [2]) has been cited as one key differentiating factor compared to other traditional content providers.

While the sheer scope of YouTube has motivated researchers to perform data-driven studies [3, 4, 5], to our knowledge there has been no significant work related to identifying trends in users' sentiments. We attempt to bridge this gap by mining the corpus of YouTube comments. We claim that in the analysis of the sentiment contained in these comments can act as the basis of an effective prediction model, with additional application to correlation between web "buzz" and stock market trends, box office results, and political elections.

We collect more than 4 million total YouTube comments to provide a representative dataset, allowing us to shed light on the following questions:

The 19th International Conference on Management of Data (COMAD), 19th-21st Dec, 2013 at Ahmedabad, India.

- *How does the sentiment for a particular keyword (video) trend over a particular window of time?*
- *How well can we forecast the users' sentiments for the next 26 weeks following the last timestamp of each dataset?*
- *Are the sentiments associated with the comments a good indicator of the correlation between the web buzz and real world events in politics, sports, etc.?*

Analyzing the sentiment of each comment and their trending patterns can give us a clear picture of overall user sentiment over a particular span of time. In this paper we calculate the polarity of each comment using the Naïve Bayes classification technique and observe the sentiment trends associated with the aggregated comments. Our approach is promising in identifying correlations between the trends in users' sentiments and real-world events.

The rest of this paper is organized as follows: Section 2 discusses the related work sentiment analysis. Section 3 describes the data collection process and the characteristics of dataset. The sentiment analysis process using Naïve Bayes is also explained. Section 4 discusses the trends in the users' sentiments, the results of forecasting the sentiment values for future, and the dependency between the comment polarity and the real-world events in the field of sports, stock market and politics. Concluding remarks and future directions are presented in Section 5.

2. RELATED WORK

Several researchers have performed sentiment analysis of social networks such as Twitter and YouTube [6], [4], [7]. Among these, the work most closely related to ours is by Siersdorfer et al. [4]. They analyzed more than 6 million comments collected from 67,000 YouTube videos to identify the relationship between comments, views, comment ratings and topic categories. The authors show promising results in predicting the comment ratings of new unrated comments by building prediction models using the already rated comments.

Pang, Lee and Vaithyanathan [8] perform sentiment analysis on 2053 movie reviews collected from the Internet Movie Database (IMDb). Their work depicted that standard machine learning techniques such as Naïve Bayes or Support Vector Machines (SVMs) outperform manual classification techniques that involve human intervention. However, the accuracy of sentiment classification falls short of the accuracy of standard topic-based text categorization that use such machine learning techniques. Another prominent work

in sentiment analysis is by Bollen et al. [9]. The authors analyzed the Twitter feeds of users using two sentiment tracking tools to accurately predict the daily changes to the closing values of the Dow Jones Industrial Average (DJIA) Index. The authors report an accuracy of 86.7% and a reduction of more than 6% in the mean average percentage error.

Our work differs significantly from the above works. Previous work on the YouTube has focused on analyzing the number of likes and dislikes for a particular comment and have performed sentiment analysis at a given point of time. Whereas we focus on the polarity of each comment rather than the number of likes and dislikes. Also, we focus on the changes in the trends of users' (commenters') sentiments over a period of time based on the contents of the videos (queries based on different keywords).

3. OUR APPROACH

This section presents the data collection steps and describes the sentiment analysis methodology.

3.1 Data collection and integration

We modeled the data by automating queries and keyword based searches to gather videos and their corresponding comments. Python scripts using the YouTube APIs were used to extract information about each video (comments and their timestamps). We collected 1000 comments per video (YouTube allows a maximum of 1000 comments per video to be accessed through the APIs), and used keywords like "Federer", "Nadal", "Obama", etc., to collect the data for specific keywords. The timestamp and author name of each video were also collected. The final dataset used for the sentiment analysis had more than 3000 videos and more than 4 million comments. YouTube comments comprise of several languages depending on the demography of the commenter. Data pre-processing involved filtering the comments only in English language. Due to limited space, we do not provide the details of the data collection algorithm here.

3.2 Sentiment Analysis

We follow the standard sentiment classification approach given in [10]. We use the Naïve Bayes classification technique for sentiment analysis. The classifier is trained on the IMDb database (based on the analysis done in the research paper [8]). The training set consists of 5000 positive and 5000 negative movie reviews respectively. The comments we collected for each keyword is used as the test data for classification. The size of the test data varies for each keyword and ranges from as low as 10000 (Dow Jones data) to as high as 300,000 (Obama). The Naïve Bayes classifier is trained on the comments from the training set and is then used to calculate the overall sentiment for each comment in the test set.

A comment is considered as a bag of independent words (i.e., the ordering of the words is not considered). The positive and negative comments in the train dataset are stored in two separate dictionaries, which we refer to as positive dictionary (positive comments) and negative dictionary (negative comments). For each comment, the polarity/sentiment of each word is calculated by calculating the number of times the word appears in the positive and negative dictionaries. For each word, the positive polarity is number of times the word appears in the positive dictionary divided by the total number times it appears in both the positive and the negative dictionaries. We calculate the negative polarity in a similar manner. Figure 1 shows an example of how the overall sentiment of a comment in the test set is calculated. The word "like" in the figure appears 292 times in the positive dictionary and 460 times in the negative dictionary. Thus, the positive polarity of the word "like" is 0.39 and negative polarity is 0.61. Similarly the positive and negative polarity of each word is calculated. The comment is then classified as positive if the positive polarity of the comment is greater than the negative polarity, and negative otherwise. For this example, since the negative polarity is greater than the positive polarity, the sentence is classified as negative.

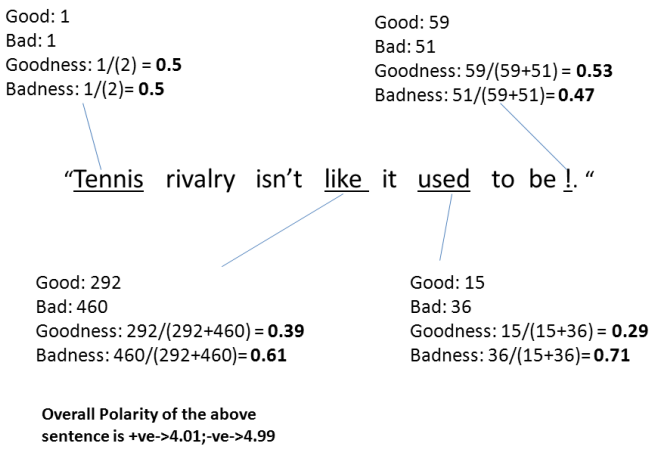


Figure 1: **Example polarity calculation [10]**

4. RESULTS

After identifying the sentiments for each comment using the Naïve Bayes classifier, we performed the analysis of the trends in sentiments. For analyzing the sentiment trends, we aggregated the sentiments for comments on a weekly basis and calculated the mean sentiment for each week. We expressed the data as a time series model and used the statistics tool R for finding the trends in the comments for each keyword.

For forecasting the future sentiment values, we used the *forecasting* module of the Weka data mining tool [11]. The forecasting module uses the existing dataset to forecast the sentiments for 26 weeks into the future.

This section addresses the first question described in Section 1. We use the *decompose()* function in R to model the comment sentiments as time series data, and to give the overall trend, the seasonality (repeated pattern), and the random trend in the data. Figure 2 shows the sentiment trends for the keyword "Roger Federer". The figure is divided into four layers. The uppermost layer (observed) gives the observed mean values per week. The second layer (trend) gives the overall trend. The third layer gives the seasonal component of the trend, which is the repeated pattern in the data. The lowermost layer gives the random component in the trend. We see that during the initial stages of the graph the sentiment was mostly positive (because those were the peak years of Roger Federer's career, winning as many as 6 Grand Slams in a span of 3 years). During 2011, the trends have hit their lowest values because Federer did not win a single Grand Slam in 2011.

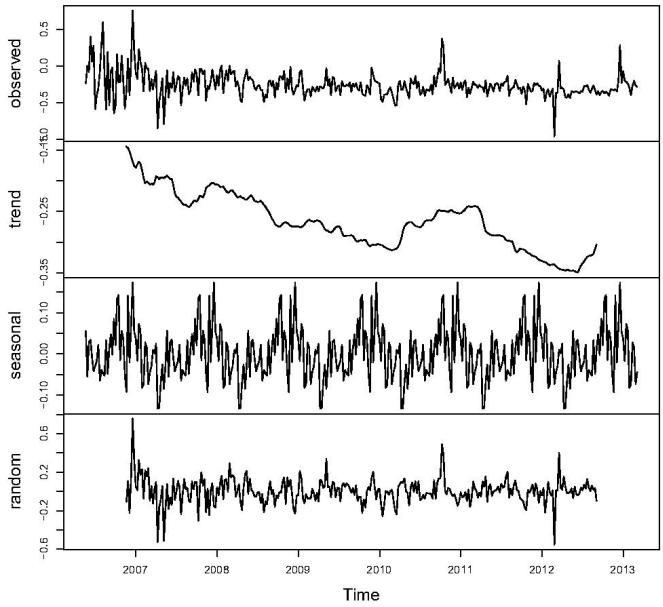


Figure 2: **Trend decomposition for Federer data**

4.1 26 weeks forecasting using Weka

This section discusses the results for the second question.

We perform the forecasts for the Federer dataset using SMO regression, a support vector machine approach. Each dataset consists of the comments, sentiment values and their corresponding timestamps.

Weka allows the forecast module to use the entire training data to build a model, and using it to forecast the values for a specified time period into the future. Since the data is aggregated over a week, we forecast the sentiments for 26 weeks into the future. We use the standard settings of the forecast module except for the base machine learner, which we set to SMO regression. Weka allows evaluation of the forecast using several metrics. We select the mean Absolute Error (MSE) and the Root Mean Square Error (RMSE) metrics. The default confidence level of 95% is selected. The system uses the known target values in the training data to set the confidence bounds. A 95% confidence level means that 95% of the true target values fall within the interval.

The 26 week forecast result for the Federer data is shown in Figures 3. We average the MAE and RMSE values for the 26 weeks (26 values). The average MAE value for the 26 weeks is approximately 0.08 and the average RMSE value is 0.13. The low RMSE value indicate that using SMO regression enables us to forecast the future sentiment values accurately.

4.2 Comparing the trends (Real World Dependencies)

This section discusses the results for the third question.

Figure 4 shows how the sentiment trends vary when it comes to comparison between two opponents in the same field of interest. Figure 4a illustrates one of the greatest rivalries in Tennis i.e., Federer vs. Nadal. The graphs are used as an example to illustrate that user sentiments were complementing in the case of both these players. We can see that during 2006-2007, which is point "a" in the plot, users had positive sentiments for Federer as he was on top of his career. Nadal became his greatest opponent by defeating him in French Open Finals for two years in a row (which depicts more positive sentiment for Nadal with respect to Federer).

Between 2007-2009 (point "b"), users' sentiments for Federer tend to decrease but stay consistent in case of Nadal (he defeated Federer in Australian Open and Wimbledon in the year 2009). Nadal's trend shows better consistency than Federer's trend. During the year 2011-2012 (point "c") (Dec 2011), Nadal's trend shows a steep decrease as he was out of action for almost 8 months to injury. This observation suggests that user/commenter sentiment is highly correlated to the performance of the respective keywords.

Similar observations can be seen for the competition between Obama and Romney for the presidential elections. Figure 4b shows the trends for both parties. Point "a" on the plot shows that during the year 2008, users' sentiments for Obama were highly positive as compared to that of Romney's as Romney was not in contention. Point "b" depicts the a low sentiment for Obama because of the economic recession. Point "c" shows Obama is way ahead of Romney after he started his campaign for a second term as a President. Point "d" shows a narrow increase in Romney's sentiment over Obama's as a result of his performance in First presidential debate.

Figure 4c presents the sentiment trends for keywords like Dow Jones and how they are closely related with the real-time fluctuations in the Dow Jones index. The figure depicts the trends in users' sentiments with respect to how the stock market behaved from 2008 to 2012.

5. CONCLUSION

In this paper we investigate the comments associated with YouTube videos and perform sentiment analysis of each comment for keywords from several domains. We identify whether the trends, seasonality and forecasts of the collected videos provide a clear picture of the influence of real-world events on users' sentiments.

We perform sentiment analysis using the Naïve Bayes approach to identify the sentiments associated with more than 3 million comments. Analyzing the sentiments over a window of time gives us the trends associated with the sentiments. Results show that the trends in users' sentiments is well correlated to the real-world events associated with the respective keywords.

Using the Weka forecasting tool, we are able to predict the possible sentiment scores for 26 weeks into the future with a confidence interval of 95%. While previous studies have focused on the comment ratings and their dependencies to topics, to the best of our knowledge, our work is the first to study the sentiment trends in YouTube comments, with focus on the popular/trending keywords.

Our trend analysis and prediction results are promising, and data from other prominent social networking sites such as Twitter, Facebook, Pinterest, etc. will help to identify shared trend patterns across these sites.

6. REFERENCES

[1] Youtube statistics. `http://www.youtube.com/yt/press/statistics.html`, 2013.

[2] Mirjam Wattenhofer, Roger Wattenhofer, and Zack Zhu. The Youtube social network. In *Proceedings of*

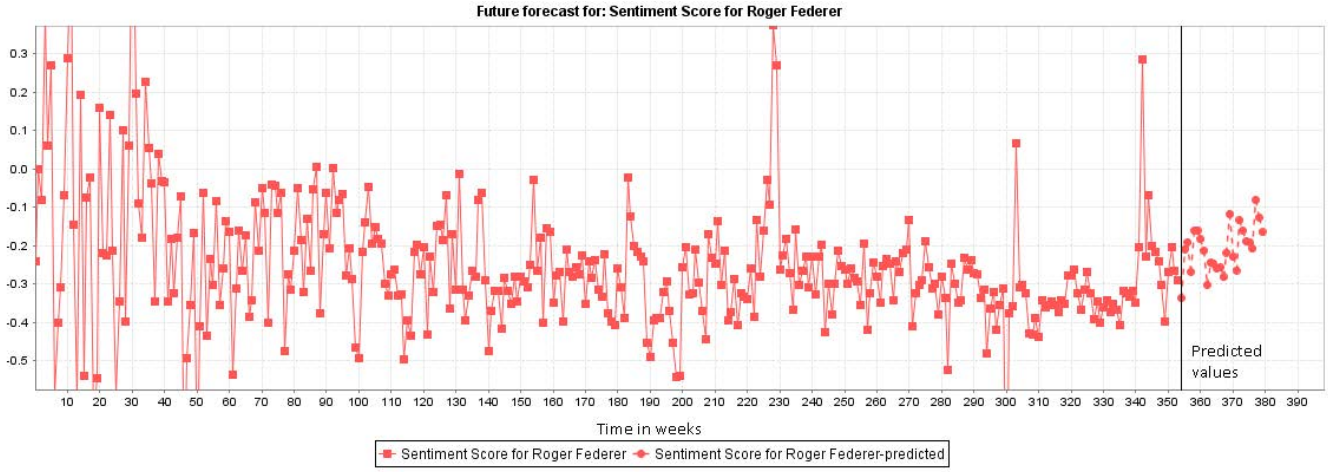
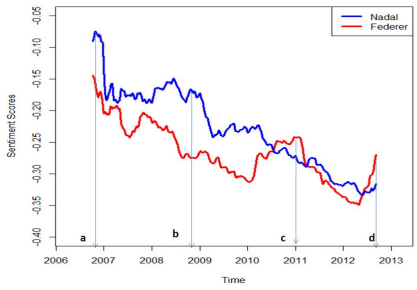
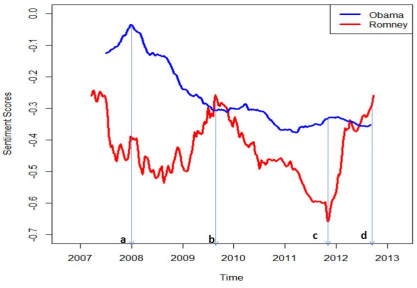


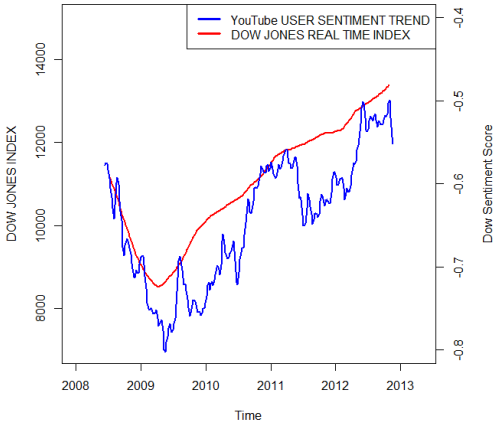
Figure 3: **26 week forecast plot for Federer dataset**



(a) Federer vs. Nadal



(b) Obama vs. Romney



(c) DJIA vs. Sentiment

Figure 4: Trend Comparisons

the Sixth International AAAI Conference on Weblogs and Social Media, 2012.

[3] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, 2007.

[4] Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. How useful are your comments?: analyzing and predicting Youtube comments and comment ratings. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

[5] Fabricio Benevenuto, Tiago Rodrigues, Virgilio Almeida, Jussara Almeida, Chao Zhang, and Keith Ross. Identifying video spammers in online social networks. In *Proceedings of the Intl. Workshop on Adversarial Information Retrieval on the Web*, 2008.

[6] Ashish Sureka, Ponnurangam Kumaraguru, Atul Goyal, and Sidharth Chhabra. Mining Youtube to discover extremist videos, users and hidden communities. In *6th Asia Information Retrieval Societies Conferences*, 2010.

[7] France Cheong and Christopher Cheong. Social Media Data Mining: A social network analysis of tweets during the 2010-2011 australian floods. In *Pacific Asia Conference on Information Systems*, 2011.

[8] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2002.

[9] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *Journal of Computer Science*, 2(1), 2011.

[10] Andrej Karpathy. Sentiment analysis example. `http://karpathy.ca/mlsite/lecture2.php`, 2011.

[11] Time series analysis and forecasting with weka. `http://wiki.pentaho.com/display/DATAMINING`, 2013.