Professor: Maro Vlahopoulou

Academic co-worker: Nikolaos Misirlis

Predictive Analytics in Social Media:
Considering Twitter data for future predictions and research

ARTICLE INFO

Published in February of 2016

Keywords:
Social Networks
Predictive Analytics
Twitter
Sentiment Analysis

ABSTRACT

Twitter is a microblogging website where users read and write millions of short messages on a variety of topics every day. Given this, we consider Twitter and data mined from it for research and prediction purposes [13]. More specifically, we focus in the predictive power of Twitter and demonstrate its influence in the fields of politics and box-office revenues. Lastly we develop a theoretical case study examining if Social media can interfere with artificial intelligence improvement. In this paper, we also present some of the principal tools for data mining, trying to understand they work.

Students involved in research and writing

Athanasoudi Eirini, Georgalidis George, Mikrouli Paraskevi Alkistis, Oikonomou Styliani Anna, Papa Anna, Prokova Konstantina

Table of Contents

Introduction

The last decade, there is a remarkable growth of the World Wide Web (commonly known as the Internet) that causes it to emerge into everyone's everyday life and to be considered as indispensable. Being part of the constantly spreading Web, Social Media has become a term equivalent to the Internet. Providing the users with the ability to communicate, create, exchange or share knowledge and opinions and generally mimic face-to-face discourses through their computers, those platforms seem to appeal to everyone despite their gender, age or heritage. As a result, millions of people are daily logging into Facebook, Twitter, YouTube etc. (W. Fan and MD Gordon, 2014).

This massive participation on the Social Media Platforms, has led to their profound influence on the way people interact with each other as well as on their behavior in various fields of human life. This is giving rise to the emerging discipline of Social Media Analytics (SMA), which draws from Social Network Analysis, Machine Learning, Data Mining, Information Retrieval (IR), and Natural Language Processing (NLP). A more specific definition of the Social Media Analytics (SMA) refers to the approach of collecting data from social media sites and blogs and evaluating that data to make business decisions (Zeng *et al*, 2010). This process goes beyond the usual monitoring or a basic analysis of retweets or "likes" to develop an in-depth idea of the social consumer.

A significant branch of the SMA which provides the ability of future predictions for consumer actions is the one called Predictive Analytics and is the method that this article mostly revolves around. Predictive analytics is the field of Data Mining concerned with the prediction of future probabilities and trends. The central element of predictive analytics is the *predictor*, a variable that can be measured for an individual or other entity to predict future behavior.

In this paper, we are going to focus on the predictive power of Twitter. Being one of the most worldwide popular social media websites, in the list of the top 10 most visited websites in the world, Twitter has garnered a significant amount of attention as a rich data source for many important forecasting problems. Having a more political nature than Facebook, it has been concerned by many as a useful tool to predict elections' outcomes. The formal character of this social network and the variety of opinions on its political predictive accuracy has gained our interest; hence we decided to dedicate this article on it.

Our paper is structured as follows:

- First, we show how Twitter help us gain important information about a large amount of different issues. The millions of interactive users gives us a reliable sample to extract predictions.
- Second, we examine if Twitter is a way for online political debating and political behavior investigation by analyzing users' interactions. Then we compare Twitter messages' prediction results with offline political sentiment.

- Third, we examine the effect of Twitter posts and conversations on the prediction of box – office revenue.
- Eventually, we propose a theoretical foundation on how sentiment analysis and other extracting intelligence tools can shape multidimensional AI behaviors.

Twitter

Since its creation in July 13th of 2006, Twitter became one of the most popular microblogging services online. Thus, in 2010, twitter gains notability reaching over 105M registered users and over 50M tweets in every-day raw. Nowadays, Twitter is considered one of the fastest growing social networks in the Internet, hosting about 500M tweets per day.
Defining the sociability, every user has a set of subscribers known as followers, while everyone is submitting periodic status updates called tweets, consisting of short messages of maximum size 140 characters. Each tweet might contain several features, such as links (images, videos and articles), hashtags, pointing into a certain issue, or /even user mentions (@username). These days, postings cover every imaginable topic from political news to product information in a variety of formats, e.g. short sentences, link to a website or even direct messages to other users. All those features mentioned, including retweets and replies are proceedings leading to interaction between users and communities [13].

Twitter and its significance

The exponential growth of Twitter has started to draw the attention of researchers from various domains. In the case of Twitter, the enormity and variance of the information, presents an interesting opportunity of transforming the data into a form that allows for specific predictions and future forecasting of multiple outcomes [12]. Another important measure is that Twitter offers its data to researchers, practitioners and organizations through API (Application Programming Interface) so as to collect and analyze them for free and without limits.
Twitter data is attached to a great deal of intelligence extraction processes. An analytical framework encompassing such research methods and metrics for extracting intelligence from twitter data consists of three methodologies from different intellectual backgrounds: descriptive analytics (DA), content analytics (CA) and network analytics (NA). DA focuses on descriptive statistics, such as the number of tweets, and the number of hashtags. CA refers to a broad set of natural language processing (NLP) and text mining methods. NA compares given data with samples from certain factors.
Data mining processes among Twitter data has given substance into a large amount of researches in every field, including finance [24, 14], healthcare [25], journalism [26] information systems [27] politics [28] marketing [29] and psychology [30].

Expectations

A key characteristic of Twitter other than its predictive power is the real time broadcasting of diverse events and accidents. Over 270 million users worldwide are potentially reporters or sensors shaping the future.

It is expected that there would be a rapid growth of Twitter data use for many other applications, including public safety and humanitarian assistance, to name but a few [15].

Twitter Predicting Election Outcomes

Introduction to Related Work

Twitter's extensive user growth in the last years, as its rankings and numbers of active users indicate, has changed the social media landscape in an interesting way. This has respectively increased the volume of information that is being available, intriguing researchers' interest in the data mining community. Thus, forecasting and predicting events using Twitter data has recently turned, according to Gayo-Avello (2012), into a popular fad. But what seems to be the researchers' foremost concern on this field is Twitter's prediction power on election results.

There is considerable debate about the accuracy of such predictions and significant analysis from various researchers in order to prove their point. One the one hand there are those who support the idea of using social media and Twitter in particular to accurately predict political alignments. DiGrazia *et al* (2013) after accounting for an array of potentially confounding variables on two U.S. congressional election cycles (2010 and 2012) come to the conclusion that there is a significant relationship between tweets and electoral outcomes. Similar are the results of Brendan O' Connor's *et al* (2010) research, where it is noted that text stream could potentially be used "*as a substitute and supplement for traditional polling"* (p.122). Bermingham and Smeaton (2011), after examining various mining techniques on the recent Irish General Election, confirm Twitter's predictive quality, highlighting the significance of sentiment analysis. Based on a rigorously constructed dataset extracted from the popular social media platform, Conover *et al* (2011) seem to share the same opinion, as they argue that political-active users' behavior on Twitter can turn out to be a treasure of political alignment prediction. A rather demonstrative research supporting Twitter's data ability to reflect the election result comes from Tumasjan *et al* (2010). Using the 2009 German National Election as a case study and analyzing prior and post election data, the team endorses the predictive accuracy of the reported platform, recognizing however the minority's dominance on Twitter's political discussion. Inspired by their work, Erik Tjong Kim Sang and Johan Bos (2012) tested their methods on the Dutch Senate Elections of 2011. Their results and conclusions though seem to be much more different, as in the latter case evaluation of tweets turns out to be insufficient.

On the other hand, Jungherr 's *et al* (2012) paper as a response to Tumasjan *et al* comes to intensify this opinion, highlighting their method's flaws and

finally considering it as not valid. Daniel Gayo-Avello (2012) expresses as well definitely a negative opinion towards the topic, referring to Twitter's predictive power as exaggerated. Likewise on his cooperative paper with Panagiotis T. Metaxas and Eni Mustafaraj (2011), in which it is argued that one should not rely completely on data mined from such sources, questioning their feasibility. Further research of the latter justifies the above, proving that already implemented techniques have no better results than chance and thus explaining their limits. Stieglitz and Dang-Xuan (2012), although they acknowledge the augmentation of political involvement through the discussed network, conclude that this may not be sufficient for predictions. Following Gayo-Avelo's recommendations, Burnap *et al* (2015) used Twitter to predict the UK 2015 General Election, revealing limitations on forecasting multi-party systems and remarking the need of geolocating tweets. Akin to this act Sheng Yu and Subhash Kak (2012) refer on their paper to attributes that increase the error of predictions, an issue concerning Mahendiran on one of his latest studies (2014). In an effort to improve the performance of traditional election forecasting algorithms using Twitter data, such those applied on the above studies, Mahendiran introduces a new approach and methodology based on query expansion that promises to reduce the prediction error.

Used Methods

The above authors have used notable methods and techniques to come to their conclusions. These are explained as detailed below.

Di Grazia *et al* , using the Twitter ''Gardenhose'' streaming API, which provides a random sample of approximately 10% of the entire Twitter stream, retrieved a random sample of 547,231,508 tweets posted between August 1 and November 1,2010 and 3,032,823,110 posted between August 1 and November 5, 2012. Of this random sample, they extracted 113,985 tweets in 2010 and 428,984 in 2012 that contained the name of the Republican or Democratic candidate for Congress from each district.
Next, they collected data on election outcomes from the 2010 and 2012 U.S. congressional elections from the Federal Election Commission. Additionally, for 2010, they collected socio-demographic and electoral control variables commonly used in other research on electoral politics for all 435 U.S. They estimated the effect of Twitter activity on electoral outcomes using three ordinary least squares regression (OLS) models.

Brendan O'Connor *et al* applied two techniques on their text data; message retrieval in order to identify messages relating to the topic and opinion estimation to determine whether these messages express positive or negative opinions or news about the topic. Particularly, message retrieval focuses only on messages containing a topic keyword. For elections, they use *obama* and

mccain. On the other hand, through Opinion estimation they derive day-to-day sentiment scores by counting positive and negative messages. Positive and negative words are defined by the subjectivity lexicon from OpinionFinder. A message is defined as positive if it contains any positive word and negative if it contains any negative word.

Bermingham and Smeaton developed a system called "#GE11 Twitter Tracker". Between the 8th of February and the 25th they collected 32,578 tweets relevant to the five main parties, identifying relevant tweets by searching for the party names and their abbreviations, along with the election hashtag, #ge11. They use MAE (Mean Absolute Error) defined to compare Twitter-based predictions with polls as well as with the results of the election. As predictive measures they define their volume-based measure as the proportional share of party mentions in a set of tweets for a given time period: 2010. On sentiment analysis they decided to use classifiers specifically trained on data for this election.They trained nine annotators to annotate sentiment in tweets related to parties and candidates for the election. The tweets in each annotation session were taken from different time periods in order to develop as diverse a training corpus as possible. They also used Freund and Schapire's Adaboost M1 method with 10 training iterations as implemented in the Weka toolkit3 (Freund and Schapire, 1996). Following from this, they use an Adaboost MNB classifier which achieves 65.09% classification accuracy in 10-fold cross-validation for 3 classes. On incorporatin sentiment they use two novel measures; inter-party sentiment, (volume-based measure, SoV) and intra-party sentiment (log ratio of sentiment, Sent(x)).

Conover *et al* extract their data from Twitter's 'gardenhose' streaming API which provides a sample of about 10% of the entire Twitter feed in a machine-readable JSON format. Among all tweets, they consider as political communication any tweet that contained at least one politically relevant hashtag. From the set of political tweets they also construct two networks: one based on mention edges and one based on retweet edges. One of the central goals of their paper is to establish effective features for discriminating politically left- and right-leaning individuals. For content-based classifications they use linear support vector machines (SVMs) They analyse these data by Full-Text, Hashtags and Latent Semantic Analysis of Hashtags. Furthermore, on Network Analysis they focus on relationships between users.

Tumasjan *et al* examined 104,003 political tweets, which were published on Twitter's public message board prior to the German national election, with volume increasing as the election drew nearer. They collected all tweets that contained the names of either the 6 parties represented in the German parliament. or selected prominent politicians of these parties. To extract the sentiment of these tweets automatically, they used LIWC2007, a text analysis software developed to assess emotional, cognitive, and structural components of text samples using a psychometrically validated internal dictionary. This software calculates the

degree to which a text sample contains words belonging to empirically defined psychological and structural categories. Specifically, it determines the rate at which certain cognitions and emotions (e.g., future orientation, positive or negative emotions) are present in the text. For each psychological dimension the software calculates the relative frequency with which words related to that dimension occur in a given text sample. Following the methodology used by Yu, Kaufmann, and Diermeier (2008) they concatenated all tweets published over the relevant timeframe into one text sample to be evaluated by LIWC.

Erik Tjong Kim Sang and Johan Bos collected Dutch Twitter messages (tweets) with the filter stream provided by Twitter. In order to get rid of false positives: tweets that contain apparent Dutch words but are actually written in another language they applied a language guesser developed by Thomas Mangin (Mangin, 2007). Their data collection process contains two filters: one is based on a word list and the other is the language guesser. They started with examining the Dutch tweets of two weeks prior to the Provincial elections and extracted the tweets containing names of political parties. In the data, they searched for two variants of each party: the abbreviated version and the full name, allowing for minor punctuation and capitalization variation. They then converted the counts of the party names on Twitter to Senate seats by counting every tweet mentioning a party name as a vote for that party. The predicted number of seats was compared with the results of two polls of the same week. To normalize party counts they used incorporating sentiment analysis improving this way the results of the prediction.

Jungherr *et al* to test whether the results of TSSW's method depend upon which parties are included, repeated their analysis including a seventh party, namely the Pirate Party. In the run-up of the 2009 German election, they collected a data set comprising all twitter messages of users whose tweets at least once contained the name of a German party or a politically loaded keyword. As a result, adding a single political party to the analysis decreases the predictive power tremendously. The results show that the absolute errors between predictions based on mentions of party names in their replication data are far from stable and vary for each party depending heavily on the time frame.

For Daniel Gayo-Avello's study, two data sets related to elections in the US during 2010 were employed. Predictions were computed according to Twitter chatter volume as in (Tumasjan *et al*. 2010) -tweets mentioning both competing candidates were not included- and sentiment analysis as in (O'Connor *et al*. 2010). While they allow a tweet to be both positive and negative, Gayo-Avello considered it to be only one of the three options (positive, negative, or neutral) depending on the sum of labeled words. Then, the predictions were compared against the actual election results.

Burnap *et al* began by collecting data from the Twitter streaming API (Burnap *et al*. 2014).

Tweets were selected if they included party and/or leader names. The search was not case sensitive. They then applied automated sentiment analysis using software developed by Thelwall et. al (2010), which allocates a string of text a positive and negative score ranging from -5 (extreme negative) to +5 (extreme positive), where each score is produced based on words in the string that are known to carry such emotive meaning (e.g. 'love'=5, hate='---4'). Where a tweet contained more than one of the search terms, it was removed from the sample to avoid misallocating the positivity in the tweet.

They first calculated sentiment scores for each tweet and produced a list of all tweets with associated positive and negative sentiment scores. Applying a rationale that positive tweets containing party or leader names can be treated as vote intentions, all tweets where sentiment scores were below -1 were removed , and those between -1 and +5 were kept . The value of the remaining sentiment scores was summed to produce a party sentiment score and a leader sentiment score. Scores for leaders representing the same party were combined, as were party mentions. The summed sentiment scores for all parties and their leaders were then combined to produce a single positive party sentiment sum for each party. All positive party sentiment sums were combined to calculate the total sentiment, which was used to normalize the positive party sentiment sum for each party, with respect to all other parties, thus producing a party-specific Twitter positive sentiment proportion.

Using 3-way human annotation, where three individuals manually annotated a random sample of 1,000 tweets including these terms according to whether each tweet was actually related to the UK Labour or Green Parties, it was identified that 78.9% of tweets containing the word "Labour" were actually about the Labour Party and only 19.4% of the tweets containing the term "Greens" were actually about the Green Party. This weighting was applied when calculating positive Twitter proportions and had the effect of reducing the overall representation of these party mentions in the relative proportions.

In a final step they converted the vote shares into a seat forecast .To do so they applied the vote share to the UK 2010 results and calculated a measure of national swing which was then applied on a constituency by constituency basis to produce an estimate of which party would win a given seat. The final number of seats won was calculated for each party by selecting the maximum value for each seat.

Stefan Stieglitz and Linh Dang-Xuan consider in their framework methodological approaches from various disciplines such as computer science, statistics, computational linguistics as well as communication studies and sociology. The framework consists of two major parts: data tracking and monitoring, and data analysis. Regarding Twitter, the data to be tracked and monitored are in the form of public ''tweets'' to which access can be easily obtained. In their framework, they consider three major approaches:

(1) topic/issue-related, (2) opinion/sentiment-related, and (3) structural.

Outcomes

Judging from the methodology and results of each article, we extrapolate some considerable outcomes on their validity and further use.

On the analysis of Di Grazia *et al* we could say that, despite the collection of socio-demographic and electoral control variables as well as the inclusion of a measure of how frequently a candidate is mentioned in transcripts of broadcasts on the cable news network CNN during the same time period, their methodology could not be very considerable. It does not require information about the physical location of Twitter users and the results are found in a random sample of all tweets during the first three months before the two election cycles, despite the fact that Twitter has been well-studied as a biased sample of the general population. Similarly, O'Connor *et al.* described mixed results: simple sentiment analysis methods on Twitter data exhibited a rather high correlation with the index of Presidential Job Approval, but the correlation was not significant when compared with pre-electoral polls for the US 2008 presidential race. Pete Burnap *et al*, however, use a model that meets at least three of the core criteria set out in the literature to reach a minimal acceptable standard for forecasting, making it replicable. Contrariwise, the method described in Bermingham and Smeaton's paper is trained using polling data for the elections it aims to predict; therefore, as stated in (Gayo-Avello, 2012), it is debatable the point of a predictive method underperforming the results obtained from the training data. On the case of Conover *et al*, we notice that one of the paper's central goals was achieved by establishing effective features for discriminating politically left- and right-leaning individuals. Using a combination of network clustering algorithms and manually annotated data, the authors demonstrate that the network of political retweets exhibits a highly segregated partisan structure, with extremely limited connectivity between left- and right-leaning users. Hence, their method could a useful tool on further investigation on Twitter's political forecasting power. We also draw some serious conclusions about the paper of Tumasjan *et al*. Their work focuses directly on whether Twitter can serve as a predictor of electoral results. In 2009 German federal election, even though 4% of all users are responsible for more than 40% of the contents, the number of messages on Twitter still could have predicted the election result, and it even came close to the tradition election poll's accuracy with a reported mean average error (MAE) of only 1.65%. Moreover, these researchers found that co-occurrence of political party mentions accurately reflected close political positions between political parties and plausible coalitions. As it constitutes the first published attempt to connect Twitter with election outcomes, its results have gained the attention of other researchers, such as Tjong *et al*, Jungherr *et al* and Gayo-Avello *et al* and thus we devote a significant part of this chapter to its annotation.

The outcomes of this paper have been criticized because they depend upon arbitrary choices made

by the authors in their analysis. Specifically, Jungherr *et al* found some important faults, concerning the procedure of data collection, the included parties and the chosen period. According to them, TSSW are somewhat vague on their procedure of data collection as they do not specify their source. Furthermore, the performance of the TSSW method depends critically upon the exclusion of certain parties for whom the instrument performs poorly – without specifying any criteria. They do not give any account of why they decided to choose this period. Moreover, a closer look at the period does not give any hint at which reasoning led the authors to choose it. The results thus depend considerably on the period under scrutiny.

Although they do not invent new techniques of analysis, Gayo-Avello *et al* present as well important conclusions about the work of Tumasjan *et al*. By repeating their methods they found that the results were not repeatable and revealed that data from Twitter did no better than chance in predicting results in the last US congressional elections.

Tjong's *et al* work provides indirect support to the conclusions by Jungherr *et al* and Gayo-Avello *et al*. After replicating Tumasjan's *et al* work on the 2011 Dutch Senate elections they conclude that tweet counting is not a good predictor. Even though the performance of their method is below that of traditional polls and, in addition to that, the method relies on polling data to correct for demographic differences in the data, they provide strong proof that the methodology of the forenamed is not accurate.

Concerning all of the above analysis and opinions on Tumasjan's *et al* work, we ascertain that their methodology is neither replicable nor valid. However, we consider it as an interesting and significant part of the Twitter's predictive power literature, since it worked as a trigger for further investigation on the subject.

Twitter predicting Movies Box-Office revenues

Introduction to Related Work

The overwhelming development and use of social media has caused researchers to study one more aspect of one's personality, their online behavior. Easily accessible and estimated, ones online activity can provide predictions for future acts on a variety of fields, one of the most popular being a movies box-office.

In an attempt to penetrate on the predictions of movie box-office via Twitter, we came across different opinions on the matter. Some believe that counting the tweets referring to one particular movie gives enough data to be analyzed and evict to an accurate prediction of the movie's box office.Whereas others find the counting inefficient because there are so many different ways for referring to a movie which makes it humanly impossible to be all taken in consideration.

Furthermore, our references show that Twitter search history seems to be a source of predictive information for movie outcome. Though sometimes even less is needed as researches show

that just mentioning a movie makes one a possible audience member.

According to Sitaram Asur and Bernardo A. Huberman (2010), Twitter, one of the fastest growing social networks in the internet because of its large amount of interactive users, can easily observe the real world outcomes from box office. Users discuss about new outcomes constantly. Based on this belief, they state that social media can be effective indicators of real world performance. Also, twitter posts about movies can be a powerful machine concerning the box – office prediction. They also conclude that tweets can improve their prediction power only after movie is released.

Many other researchers came to state their opinions on this topic as well. Sheng Yu and Subhash Kak (2012) express a very similar opinion to Asur's one. They state that the large amount of information transferred every day and the different point of views that we can find in social media could be sufficient to predict movies box – office, one of the most studied areas. Movies are widely talked on social media, so they make box – office easily accessed and estimated. They also describe that because of the big quota of total sales on the opening weekend, we could get the approximate box – office by then. They conclude that there is a clear logical correlation between social media contents and movie box – office. At the end, they add that social media is also useful in predicting the winner of the Oscars, although the election process has little to do with the wisdom of the crowd.

On the other hand, Sharad Goel *et al* (2010) object to that belief, describing that what consumers are searching for online can also predict their collective future behavior days or even weeks in advance. They also believe that search counts are generally predictive of consumer activities such as attending to movies. That may be the case that movie searches are primarily intended to locate theaters or to purchase tickets online. Those are activities that are directly related to box – office prediction revenue. The usage of movie names makes it easier for the researchers to identify the movie that the user is searching for. Users also make a research on the movie they are interested on through other's posts on Twitter. In conclusion, the team opposes to the previous opinions stating that movies box – office can easily be predicted before the movie comes out on theaters, according to what users search online and the references that are being made on the specific movie.

Furthermore, Yafeng Lu *et al* (2014), based on their study and the "positive results" it brought, conclude that given the subjects of their self-reported lack of movie knowledge it is clear that the integration of social media and visual analytics for model building and prediction can quickly generate insight at a near professional prediction level. They also add that analysts can utilize the system to explore and combine information. Moreover, underlying mechanisms for similarity matching and data filtering can help a user quickly engage in exploratory data analysis as part of the model building process.

Used Methods

All of the above authors used a variety of methods to get to their points. Let's examine them one by one.

First, Asur and Huberman, started by searching Twitter with keywords and the movie title to make sure that the posts were related to the movie. With an average of 2 movie releases per week they collected data of a three-month-period. They defined the "critical period" of each movie from one week before it's released to 2 weeks after, where the popularity of the movie fades.

With that method they collected data for 24 movies. So with a variety of a 2.89 million tweets from 1.2 million users they figured out that 1) the busiest time for a movie is around the time it is released providing better the box – office revenue 2) the number of tweets per unique author changes overtime 3) the tweets display the distribution of tweets by different authors over the critical period. Afterwards, they examined how attention and popularity of the movies changes over time. The pre-release attention is earned by trailer videos, photos, blogs, news etc. On twitter this is characterized by sharing URLs as well as retweets which involves users re-sharing those data with their Twitter friends. Both these forms are pretty important to disseminate information regarding movies being released.

Then, when the movie was released, they examined the first week box – office revenues prediction. They checked the "pre-release attention" data and compared them with the real-world outcomes. Their goal was to observe if the knowledge that can be extracted from the tweets can lead to reasonably accurate prediction of future outcomes in the real world.

Secondly, Yu and Kak used some several methods for their statement. One of them, as explained by the authors, is the regression method, which analyzes the relationship between the dependent variable, prediction results and one or more independent variables such as the social network characteristics. Also, there is the Bayes classifier, which is a probabilistic classifier using Bayes' theorem. Based upon the probability of the prediction event, Bayes classifier uses the Bayesian formula to calculate its posterior probability, that the object belongs to the result classes, and then selects the class with the posterior probability, as the event is most likely to have the result. If the prediction result is discrete, the Bayes classifier can be applied directly. Otherwise, the prediction result must be discredited first.

Then there is the K-nearest neighbor classifier, one of the simplest machine learning algorithms, which tries to cluster the objects according to their distance to others. Moreover, there is the Artificial Neural network. A.N.N is a computational model to simulate the human brain. According to that, the need of consistence of lots of artificial neurons is created. These neurons could belong to many interconnected groups, including input layer (responsible for receiving raw data and transmitting them to the next layer), hidden layer and output layer (gives us the final prediction result).

Afterwards the decision tree is a visual technique in data mining and machine learning. Travelling for "root node to leaf", one entity will get the

prediction result. The Classification tree (applied when the prediction output is discrete classes) and the regression tree (used when predicted outcome is continuous value) are two basic and major types of decision trees. Finally, the model based prediction method, possibly the hardest way to make predictions, requires the building of a mathematical model on the object before prediction, which also requires deep insight into the object. At this point the knowledge about social media is not enough for the development of effective models for them.

Goel *et al*, in order to identify user intent from queries, applied a simple and effective heuristic that leverages search engine technology. They categorized user queries as "movie-related" whenever an Internet Movie Database (IMDb) link appeared in the first page of the research. Then, by extracting the unique movie identifier in the corresponding IMDb link, they mapped the queries. When multiple IMDb links appeared they determined user intent from the top ranked result. This method brings easily results when the movie's name involves unique words. For example, for the movie *Transformers 2: Revenge of the Fallen* it's easy to conflate with other unrelated searches (e.g. for transformers of the electrical equipment variety). It is expected that other searches will be unrelated with the release of the movies. Then, after the movie is released they categorized the movies by the revenue rang based on the budget and number of opening screens obtained from IMDb.

For the last article, Lu *et al* for their conclusion used the VAST Box – Office challenge that the Visual Analytics Science and Technology conference ran, using social media to predict the opening weekend gross of movies. This challenge was about users' interaction with visualization tools to develop predictions.

Unlike most specialized data sources, movie data lends itself well to analyzing visual analytics modules as many casual users think of themselves as movie domain experts. To focus on the specific problem of predicting the opening weekend box – office gross of upcoming movies, Lu *et al* took advantage of data from Twitter and YouTube. From Twitter, they collected data for 112 movies released since 2013. The tweets that are collected are based on the hashtags, from the movie's official twitter account. Each of the 2.5 million tweets that are collected includes the posting time, retweet status, user profile information and Tweet text sentiment. From the YouTube, using a rule-based script to collect its data which contains the total view count and timestamps, he calculated a range of features such as comment volume and interpolated view counts prior to the opening weekend. Overall 7 million YouTube comments for 1104 movies were collected. Also, since IMDb has more than 2.8 million entries with each entry consisting of hundreds of features, to overcome the data sparseness, he calculated numeric values on a per-movie basis by aggregating gross incomes and ratings of previous movies that the cast of a new movie was involved in. Finally he obtained a high quality movie data set of approximately 2000 movies with up to 72 features per movie.

Outcomes

In our opinion, Asur and Huberman's methods are not effective, because they are not based on scientific systems or any statistical analysis. Of course, they may bring results but they are not trustworthy since they are only based on an amount of tweets.

Yu and Kak's method is quite trustworthy, since the results are examined thoroughly with many methods based on science, math and algorithm solution. Those methods can tell us accurately how box – office revenue will pass off. This way twitter and social media in general will be a precious tool on predicting how the future movie releases are about to evolve. And it's important that the final results of the research are published after the first weekend of the movie premiere since the results are detailed.

Goel's method is similar to Asur and Huberman's one, despite the fact that this one is only based on tweets of the official movie's account. This may not be accurate about the box – office prediction since box – office is caused by the audience. Thereafter the Asur and Huberman's method is more trustworthy since is based on the users of twitter who are the audience and the people who are responsible for a movie's box – office.

Lu, uses maybe the most reliable method, since it's based on the audience. Collecting so much information is the best way to actually be able to tell the box – office revenue. It's the easiest way to observe the discussions between the audience and also how they react to any news or new marketing moves of the movie. This is what makes this method the best.

Twitter is a powerful machine concerning the predictive analytics. The every-day rising number of active users who interact with each other, share opinions and influence each other can give us many predictions about a large amount of issues such as box- office revenue, elections etc. even issues that have nothing to do with the audience such as the Oscar awards.

Extracting intelligence from Social media

In recent years, we have witnessed the ubiquitous influence of social media in our lives, while at the same time the collected content of this social media interactions is as much as needed so as to predict outcomes and to shape the future [12, 31]. Additionally, there has been a great development in the field of artificial intelligence (AI) and on the way AI processes the information. We propose a theoretical foundation on how sentiment analysis and other predictive tools can extract intelligence from Twitter and Facebook, in some cases, shaping in that way multidimensional AI behaviors. We further demonstrate on how social media can interfere constructively with those AI behaviors so as to formulate judgments about the right course of actions to be taken in a given situation.

Background

Artificial Intelligence is the science and engineering of making intelligent computer programs or machines through pattern

recognition, predictive modeling, text mining and data analytics. Artificial intelligence is commonly referred to as machine learning and was originally developed to enable computers learn [32]. Today, the technology is based on a number of advanced mathematical methods for optimization, regression and classification. Applications has been in a wide variety of fields including the airline industry, medicine, insurance, mechanical engineering, manufacturing, software engineering and finance [33]. On the other hand, because of its ease of use, speed and reach, social media is fast changing the public discourse in society and setting trends and agendas in topics that range from the environment and politics to technology and the entertainment industry [12]. As a form of collective interactions, we presume that social media and especially Twitter and Facebook have the power at predicting real-world outcomes, even in specific occasions. We focus on examining the case of teaching AI systems through social media because it could mean a low-cost solution of data collection, data analysis but also data storage. Through social media, information can be categorized and filtered so as to generate a plethora of patterns representing personalities. Our goal in this case study is to present the real influence of predictive analytics on social media for AI behavioral development. Supporting our goal we further analyze the recommended procedure, by giving examples and instructions.

Related Work

Social media has been considerably studied as a source of information for many fields. Although, there are not many implementations concerned specifically in what we analyze. Sitaram and Bernardo [12] use Twitter database to predict outcomes and improve forecasting skills, Valerio *et al*. [17] prove how tie strength of offline networks and online ones can be similar in predicting ties between people. Java *et al* [34] investigated community structure and isolated different types of user intentions on Twitter. While, Johan, Huina and Xiaojun [14] are making a deep psychometric analysis for public mood on neural networks.

Data set description

A new level of connectedness among peers, creates a huge database by providing new ways for the dissemination and consumption of data [35]. Evolving data mining technologies and the increasing processing power of today's computers, support the desire to appropriately analyze parts of today's growing public deluge in real time so as to conclude with patterns of real reactions and habits. Data are being extracted from API's of every social network itself. In our paradigm we are concerned mostly with Twitter and Facebook. Both serve there data free, although Facebook enters few limitations making only the data from groups and fan pages open to the public. As Valerio *et al*. [17] believe, the widespread use of online social networks such as Facebook and Twitter, is generating a growing

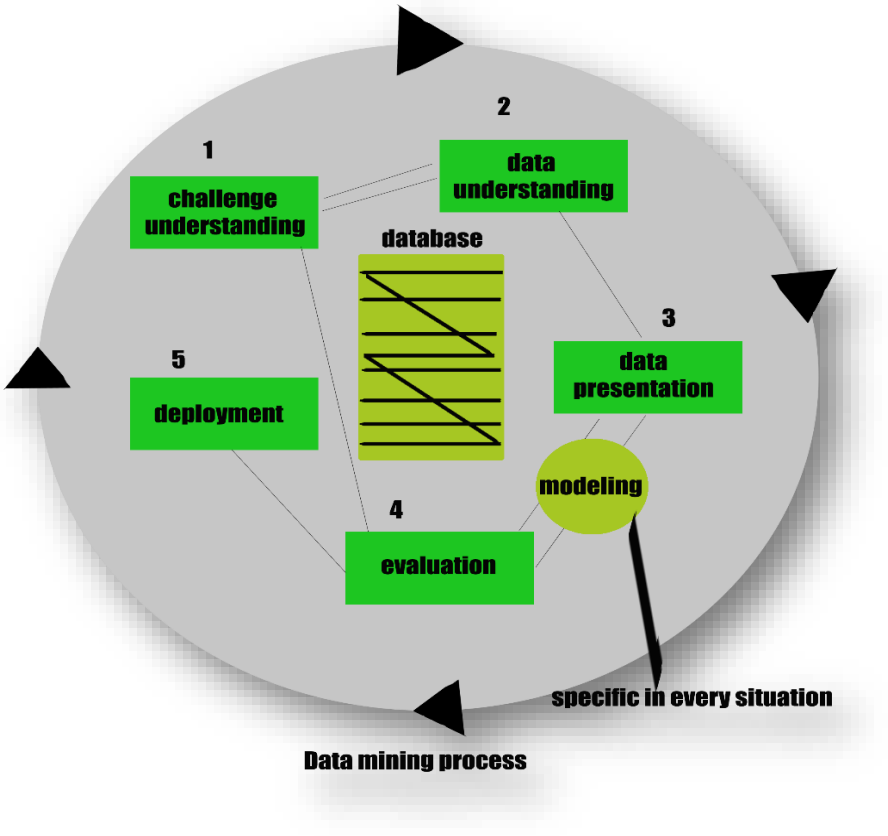
amount of accessible data, concerning social relationships.

Extracting intelligence from Social media using analytics

Analytics are the methods of decomposing concepts or substances into smaller pieces, to understand their workings. Predictive analytics encompasses a variety of techniques from statistics, data mining and game theory that analyze current facts to make predictions about future events.

Nann *et al*. [35] are stating that after analyzing data, the more context-specific the algorithms are designed the higher the quality of sentiment recognition will be. So, In order to analyze our data as we want we should specify what kind of algorithms we will use.

We are about to separate the process in 6 steps as shown below in the Figure.



Collecting Twitter and Facebook data begins with identifying the topic of interest, as Figure suggests, to understand the challenge. Here, our challenge is about machine learning from Social media. This is the step of measuring our interest using keywords or hashtags executed on given APIs. We acquire those APIs through tools such as GNIP and DataSift for Twitter and Graph API for Facebook [15]. Those data sets in the first step are unstructured and difficult to access. So as to

produce outputs from our data we move on step 2. Step 2, is similar to what is generally known as content analysis, referring to a broad set of natural language processing (NLP) and text mining methods, for extracting intelligence from social media data [15]. A careful consideration of text cleaning and processing is prerequisite for intelligence gathering. In this step we are including the functionality of descriptive analytics, which focuses on descriptive statistics, such as the number of tweets, distribution of different types of tweets and the number of hashtags. In our case, in order to move on the 3rd step we should also run a sentiment analysis, which is primarily interested in extracting subjective information (e.g. emotions, opinions) in tweets. Sentiment analysis, is a well-studied problem where a given text is being labeled as Positive, Negative or Neutral. A tool for this reason can be the DynamicLMClassifier which is a language model that accepts training events of categorized character sequences [12]. After the analysis and the removal of all the "noise" from our data, we are monitoring our results and arranging our next step which is the modeling. Before the modeling process we make a better approach on what we want to achieve. In our case study, we want to examine if there is a chance to transmit some of our habits and actions into machines, so as to create uniquely reactions based on already examined patterns. In our modeling we suggest the development of an algorithm which will give the base to filter every possible income we get. This algorithm should contain some psychological measures. We want our machine to separate what is vice and what is virtue. The algorithm should be structured to measure more than one aspects, such us politics, cultures, societies, perspectives etc.

Sociologists, anthropologists and psychologists have largely studied social relationships in humans from two different points of view. On the one hand, the analysis of personal networks, called ego, study the relationship an individual has with others. On the other hand, social network analysis on public networks studies the existing relationships between people inside a bounded population or community. Results from ego network analysis, already highlight a number of key properties characterizing the social behavior of the users [17]. We aim to make a comparison between the results derived from ego networks' analysis and from the characteristics asked from our algorithms.

Constructing our model, we suggest OpinionFinder, a tool which provides us with a more detailed view of changes in public mood aligned on a daily time series, by analyzing the content collected. This may however ignore the rich, multi-dimensional structure of human mood. So as to capture additional dimensions of public mood we apply a second mood analysis tool, labeled GPOMS, measuring human mood states in terms of different mood dimensions namely calm, alert, sure or vital to name but a few [14].

We run an analysis of LIWC2007, a text analysis software developed to assess emotional, cognitive and structural components of text samples using a psychometrically validated internal dictionary. This software determines the rate at which certain cognitions and emotions are presented in the text. Then, use predictive model so as to compare the

results with our multidimensional profiles. Description models, run again the comparison, but this time focusing on as many variables as possible. Decision modeling uses optimization techniques to predict results. Strengthening our model, we equip the procedure with a tool called Oracle Real Time Decisions which learns from every single interaction derived from previous steps. Each decision is made using a statistical model predicting the outcome of the interactions between human and machine.

In order to make our decisions stronger, we force our results to pass through "scenarios" and what-if patterns, included in Crystal Ball tool, in the form of questions, based on likelihood and realism in order to determine and implement the best outcome. Crystal Ball can handle both structure and creativity so as to create the data for the future [31].

Evaluation and deployment

In this particular step we examine our model. We figure out our accuracy and we make new assumptions based on our results, from its time we run the sample. This is also the point, where we compare the results that we get from our training method with the results of the existing training methods, such as the artificial neural network A.N.N. As previously described A.N.N. are computing systems made up of a number of highly interconnected processing elements, which process information by their dynamic state response to external inputs.

Case study findings

In this case study, we theoretically examine how we can extract intelligence from social media networks based on previously conducted actions. Using Predictive tools and few algorithms filtering our behavior results we believe that there can be progress in how we reach intelligence behind this massive data. We propose to teach machines by a structured pattern counting its time the best possible reaction, in the way humans learn from other humans' behaviors. The nimbleness of the human brain through its complex structure of neurons and interactions for now is something that simply cannot be programmed. That is why, it is proposed that predictive analytics in social media will unlock the benefits, and narrow AI in particular that will bring together big data and other sources of information to create large and informative data pictures of personalities.

Future research could transfer all the theoretical part into technical testing. Also, there would be a lot of development if more analysis could be applied. Furthermore, there could be a combination between the method introduced and the prospect of getting also feeling patterns through social media.

Conclusions

In this paper, we are generally trying to prove the existence of the prediction power that we acquire through the social media. We especially concentrate in Twitter, because it is considered a legitimate communication channel. In sum, our results demonstrate that Twitter can be considered a valid indicator of political opinion and box office through appropriate methods. We further assign a theoretical approach on the way social media can interfere with the development of Artificial Intelligence.

The fact that even the simplest methodology used in our study, because of cost consideration, was able to generate justifiable results is encouraging and arguments to additional prospects can be derived.

References

1) Daniel Gayo-Avello (2012). *A Balanced Survey on Election Prediction using Twitter Data*. arXiv.org
2) Daniel Gayo-Avello, Panagiotis T. Metaxas, Eni Mustafaraj (2011). *Limits of Electoral Predictions Using Twitter.* Proceedings of the 5th International AAAL Conference on Weblogs and Social Media.
3) Joseph DiGrazia, Karissa McKelvey, Johan Bollen, Fabio Rojas (2013*). More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior*. PLOS ONE.
4) Sharad Goel, Jake M. Hofman, Sebastien Lahaie, David M. Pennock, Duncan J.Watts (2010). *Predicting consumer behavior with Web search.* PNAS.
5) Aravindan Mahendiran(2014) *Automated Vocabulary Building for Characterizing and Forecasting Elections using Social Media Analytics*
6) Brendan O'Connor,Ramnath Balasubramanyan,Bryan R. Routledge,Noah A. Smith, (2010) *From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series* Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media
7) Pete Burnap , Rachel Gibson , Luke Sloan, Rosalynd Southern and Matthew Williams (2015) *140 Characters to Victory?: Using Twitter to Predict the UK 2015 General Election* http://ssrn.com/abstract=2603433
8) Yu, Sheng, and Subhash Kak. (2012). "*A Survey of Prediction Using Social Media."* arXiv preprint arXiv:1203.1647
9) Stefan Stieglitz, Linh Dang-Xuan (2012). *Social media and political communication: a social media analytics framework*. Social Network Analysis and Mining, 4(3), 1277-1291.
10) Erik Tjong Kim Sang and Johan Bos(2012) *Predicting the 2011 Dutch Senate Election Results with Twitter*. Proceedings of SASN 2012, the EACL 2012Workshop on Semantic Analysis in Social Networks
11) Adam Bermingham and Alan F. Smeaton(2011) *On Using Twitter to Monitor Political Sentiment and Predict Election Results* .Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011, pages 2–10,
12) Sitaram Asur, Bernardo A. Huberman (2010). *Predicting the Future with Social Media*. arXiv preprint arXiv:1003.5699
13) Andranik Tumanjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe (2010). *Predicting Elections with Twitter: What 140 characters reveal about political sentiment.* Four more Years, 2012-2013.
14) Johan Bollen, Huina Mao, Xiajun Zeng (2010). *Twitter Mood Predicts the Stock Market.*

Computational Science. Journal of Computational Science.

15) Bongsug (Kevin) Chae (2014). *Insights from hashtag #Supplychain and Twitter Analytics: Considering Twitter and Twitter Data for supply chain practice and research.* Productional Economics.
16) Michael Ballings, Dirl Van Den Poel (2013). *CRM in Social Media: Predicting increases in Facebook Usage Frequency.* European Journal of Operational Research.
17) Valerio Arnaboldi, Andrea Guazzini, Andrea Passarella (2011). *Egocentric online Social Networks: Analysis of Ket Features and Prediction of tie Strength in Facebook*. Computer Communications.
18) Yafeng Lu, Robert Krüger, Dennis Thom, Feng Wang,Steffen Koch, Thomas Ertl, and Ross Maciejewski, (2014). *Intergrating Predictive Analytics and Social Media*. Vast, IEEE.
19) Neil Savage (2011). *Twitter as medium and message*. Communications of the ACM
20) Amrapali Mhaisgamali, Dr. Nupur Giri (2014). *Detailed Descriptive and Predictive Analytics with Twitter based TV ratings*. IICAT.org.
21) Jennifer Golbeck, Cristina Robles, Karen Turner (2011).*Predictive Personality with Social Media. alt.chi: Playing Well with Others.* CHI'11 Extended Abstracts on Human Factors in Computing Systems. ACM, 2011.
22) Michael D. Conover, Bruno Goncalves, Jacob Ratkiewicz, Alessandro Flammini and Filippo Menczer(2011) *Predicting the Political Alignment of Twitter Users.* 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing
23) Daniel Gruhl , R. Guha , Ravi Kumar , Jasmine Novak , Andrew Tomkins (2005).*The Predictive Power of Online Chatter.* KDD'05.
24) Bollen, J., Mao, H., Zeng, X.(2011)*Twitter mood predicts the stock market.* J. Comput. Sci. 2, 1–8.
25) Park, H., Rogers, S., Stemmie, J.(2013*) Analyzing health organizations' use of Twitter for promoting health literacy*. J. Health Commun. 18, 410–425.
26) Lasorsa, D.L., Lewis, S.C., Holton, A.E.(2012). *Normalizing Twitter*. Journal. Stud. 13, 19–36
27) Aral, S., Dellarocas, C., Godes, D.(2013). *Social media and business transformation: a framework for research.* Inf. Syst. Res. 24, 3–13.
28) Gayo-Avello, D.(2012). *A Meta-analysis of State-of-the-art Electoral Prediction from Twitter Data*. Social Science Computer Review (2013)
29) Jansen, B., Zhang, M.(2009). *Twitter power: tweets as electronic word of mouth*. J. Am. Soc. Inf. Sci. Technol. 60, 2169–2188.
30) Dodds, P., Harris, K., Kloumann, I., Bliss, C., Danforth, C.(2011). *Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter*. PLoS One 3, 1–26.
31) Oracle(2010). *Predictive analytics: Bringing the tools to the data.*
32) Stuart Jonathan Russel, Peter Norvig(2013). *Artificial Intelligence.*
33) Philip Klahr, Carlisle Scott(1992). *Innovative Applications on Artificial Intelligence 4.*
34) Akshey Java, Xiaodan Song, Tim Finin and Bell Tseng (2007) *Why we Twitter: understanding microblogging usage and communities* p.56-65. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, 2007.
35) Nann Stefan, Krauss Jonas, Schoder Detlef(2011) *Predictive analytics on public data – the case of stock market*. 21st European Conference of information systems.
36) Andreas Jungherr, Pascal Jürgens, and Harald Schoen(2012)*Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions.* Social Science Computer Review