# External Validity in Social Media Predictions

Author: Dennis Paschek
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

**ABSTRACT:** There is a growing number of papers which attempt to predict the future based on Social Media. Although these papers were able to present rather successfully results per se, there is a gap in the literature. Current approaches lack of external validity, i. e. to what extent a prediction model can be applied to a different setting than applied to by its author. This paper seeks out to close this gap. First I will provide a literature review concerning Social Media predictions on the one hand and external validity on the other hand. Following will be a conceptualization and application towards current papers. I conclude that external validity in Social Media predictions still remains a challenge and requires in the future more attention.

**Supervisors:** DR. A.B.J.M. (FONS) WIJNHOVEN

**Keywords**
Socail Media, Prediction, External Validity, Segment Mining, Machine Learning

*5th IBA Bachelor Thesis Conference*, July 2nd, 2015, Enschede, The Netherlands.

# 1. INTRODUCTION

The Web 2.0 and especially Social Media are sources of sheer infinite amount of data, which has the potential to deliver useful information for researcher and organizations alike. Kietzmann et al. (2013) defines Social Media as a highly interactive platforms via which individuals and communities share, co-create, discuss, and modify user-generated content. There is already a large body of established literature about Social Media in general. In the last couple of years many articles got released concerning its predictive power as social media messages can have an important application for organizations and individuals that want to track the impact of products, services, events and people on social media (Tjong Kim Sang, 2014).

However there are still gaps in the literature present which need further research. One of these gaps is external validity of Social Media Predictions. With this paper I attempt to close this gap and bring the attention of future papers towards this challange. Shadish et al. (2001) defines external validity as “inferences about whether the cause-effect relationship holds over variations, settings, treatments, and outcomes.” In terms of Social Media prediction, models should be capable of delivering equally satisfying results in other settings than the original one it got introduced. In order to analyze this topic, the paper begins with a literature review which gives an overview of the current state of literature related to this field. First I group different prediction model based on the applied field. The next part relates to the aspect of external validity and its application for Social Media Prediction.

# 2. LITERATURE REVIEW

Predicting the future on basis of Social Media is a rather new and therefore still developing topic, which has definitely potential, but requires a lot of further development. (Gayo-Avello, 2012; Shoen et al. 2013). Nonetheless there is a growing interest from many different fields, like politics, sales prediction, unemployment rate financial markets or the entertainment industry. As Social Media Prediction is not limited to any industry or field, in consensus with Shoen et al. (2013) it is possible to divide the literature in different sections based on popularity in the literature and success. A brief summary of these field will be in the next paragraph.

## 2.1 Approaches of Social Media Prediction

In terms of Social Media prediction there has to be a distinction made between models based on expression volume and models based on a sentiment analysis. Prediction models solely based on the expression volume only take into account the amount and/or frequency a specific set of key words gets mentioned in Social Media. Examples for this would be the amount of polical parties mentioned in Twitter (Tumasjan et al. 2010) or the frequency of influenza-related posts on Twitter (Lampos and Christianni (2010). Sentiment based predictions go one step further. These approaches use a form of machine learning to analyze the language of posts in Social Media and tries to determine the polarity of each in categories like “negative” “neutral” or “positive”. Examples for sentiment based predictions are Tjong Kim Sang & Bos (2012) or who used Twitter mentions to predict the Dutch Senate Elections and Asur and Huberman (2010) who also used Twitter to predict upcoming movie sales. Sentiment Analysis in general is a well-studied field .A very elaborative and highly respected summary of this field is made by Pang and Lee in 2008.

## 2.2 Epidemiology

Epidemiology is the scientific study of disease patterns among populations in time and space. One of the first attempts to use Big Data in order predict the incidence of a disease was made by Cooper et al. (2005) in which Yahoo search queries were used to predict estimated the cancer incidence, estimated cancer mortality, and the volume of cancer news coverage. In the year 2008 two particular papers received a lot of attention. On the one hand published Polgreen et al. (2008) a paper in which also Yahoo got used to find a correlation between influenza-related web searches and reported infections. On the other hand there is the work of Ginsberg et al. (2008) which came to similar results like Polgreen et al. though using Google as a source of data for their prediction. Furthermore became the latter basis for the Google Flu Trends website[1]

Inspired by the relatively high success of web search prediction, did Lampos and Christanini (2012) build a model which used geo-tagged (i.e. posts of user which state their current location) Twitter updates to measure regional influenza statistics. Similar applications are found in the work of Dredze (2012). Also did Collier (2012) and Signorini et al. (2011) use Twitter for general public health and infection rates.

Based on the high amount of papers about predicting epidemiological factors via web searches and more recently Social Media, this area has some potential, however not everybody is as optimistic about it. So far predicting the future in the field of epidemiology could provide the best results compared to other fields and also received the most attention from the literature.

## 2.3 Sales Prediction

Sales predictions were mostly related to the sales of video games and book sales music charts and the box office for moves. As a reason for these specific queries, there are several reasons; Firstly, do they all have in common that the actual sales figures were relatively easy to find, for example on pages like www.imdb.com or www.billboard.com . A second benefit of these choices is the fact that they are very popular among Social Media user and that they therefore generate a bigger data set (Goel et al. 2012). Finally it is reasonable to assume that the signaled intention to watch a movie or buy a book on a Social Media platform eventually is followed by the actual purchase (Asur & Huberman, 2010). An early paper got published by Gruhl et al. in the year 2005 in which they presented a positive correlation between mentions in personal blogs and books sales. They also went a step further and were able to predict spikes sales days and sometimes weeks in advance. Asurs and Hubermans model to predict the box office of movies is arguably the most famous one with over 820 recorded cites on Google Scholar. The authors used Twitter data to predict success of movies two weeks in advance. They managed to outperform the Hollywood Stock Exchange prediction market which is known as the “golden standard of the industry”. However in 2012 Wang et al. doubted the validity of Asur & Hubermans work and constructed a similar approach which lead to a less successful, but methodologically more valid, performance. In line with Wang et al., mentioned Goel et al. a lack of literature, in particular related to the predictive power of search among different domains. Although the reasons are ultimately unclear, they propose three possible factors for the variance. First, the varying sample size within different domains. Secondly the varying difficulty to relate search queries to a specific product and as the last point , contrary to Asur and Hubermans point of view, that in some domains a post in Social Media or a web search about a certain product is more likely to lead to a purchase than others.

## 2.4 Stock Market Movements

This field in is very promising, because even slightly improvements in prediction the stock market can lead to significant benefits and thus the high amount of paper concerning this particular area. One of the first attempts was made by Wüthrich et al. (1998) who used articles from online financial newspapers to predict the stock markets of several countries in Europe, Asia and America. Although the system was not always correct, it was overall successful. Later on a similar research was made by Tumarking and Whitelaw (2001). But instead of using the content of online newspaper, they extracted with data from a business-forum. And despite the similarity in the design of the research, but articles lead to contradicting results. While Wüthrich et al. see a causal relation of newspaper and stock market movements, Tumarking and Whitelaw see it the other way around, rather is the stock market dictating news articles and forum post respectively. Despite the early complicates in the field, are more recent articles offering a more positive outlook.More recent research provides a more optimistic outlook on the topic. As an example can be the work of Choudhury et al. (2008) be mentioned in which the authors used content from weblogs. Another famous work is made by Bollen et al. (2011). Here, the source of data are mentions received from Twitter.

## 2.5 Electoral Predictions

Electoral predictions had a peek between the years 2010 until 2011. For example stated Tumasjan et al. (2010) that the search volume towards a politician and party could potentially be an indicator for an upcoming election. Similar results got published by O´Conner et al. (2010) who were able to use the frequency of mentions of Obama and his rival McCain to predict the election However, in contrast to previous work, did Chung and Mustafaraj (2011) find out that search volume predictions on Twitter are not a competitive alternative to sentiment-based predictions. An in-depth analysis of electoral prediction from Twitter data is made by Gayo-Avello (2012). Although he acknowledges the potential this field has, he criticizes previous work due to weak methodology or poorly execution of the machine learning process. As the main weaknesses he states possible demographic- and self-selection bias, varying performance compared to baseline models and the fact that all of the researches were post-doc.

# 3. EXTERNAL VALIDITY

The first concept of validity was published by Campbell and Stanley in 1966 and extended by Cook and Campbell (1976, 1979). Validity is based on four pillars, one of them being external validity. Cook and Campbell define as following ‘‘External validity asks the question of generalizability: To what populations, settings, treatment variables, and measurement variables can this effect be generalized?” Since then a number of varying definitions occurred: Shadesh et al. (2001) defines external validity as “inferences about whether the cause-effect relationship holds over variations, settings, treatments, and outcomes.” McTavish and Loether (2002) refers to External validity as “whether the results of a study can be legitimately generalized to some specified broader population.” Another definition is given by Monette et al. (2002): “External validity concerns the extent to which causal inferences . . . can be generalized to other times, settings, or groups of people.”

Lukas (2003) divides, based on the distinct definitions existing, external validity into two aspects. The first one is about generalizing from a sample to a larger population, while the second treats external validity as generalizing to populations or settings other than those studied. And indeed both aspects apply in relation to Social Media prediction, too. Researcher try to predict a varying events (i.e. elections or sales) occurring within a bigger population, using on a smaller sample (i.e. Twitter).

According to Calder et al. (1982) are there two dominant point of view about external validity. On the one hand there is the opinion that external validity has a high priority as it is a good indicator for a sound construct under any circumstances. On the other hand there is a different view provided by Cook and Campbell (1979), which says that external validity is not a crucial aspect when the application is mostly theoretical.

However, as Prediction models attempt to provide a practical tool for different fields, external validity has to be taken into account while designing and applying said prediction models. Winer (1999) proposes to do so by better understanding how external variables may interact with the theory which seem to be irrelevant during the beginning phase of the research. In alignment with that, state Alm et al. (2015) that external validity can only be improved by investigating possible, external factors.

## 3.1 Threats to External Validity

In a practical environment it is crucial to consider the factors that threaten the external validity of their studies (Rich & Oh, 2000). Campbell and Stanley (1963) identified four variables which could harm external validity, namely interactional effects of testing, interactional effects of selection biases with experimental variables, reactive effects of experimental arrangements, and multiple treatment interferences. Later on, Shadish added another threat which leads to following list:

1.) Interaction of causal relations with units
2.) Interaction of causal relations with treatments
3.) Interaction of causal relations with outcomes
4.) Interaction of causal relations with settings
5.) Context-dependent mediations

The first threat is concerning the properties of a sample. Babbie (2009) states in his book the most common ones as gender, age, location, and education.

The second threat indicates that a different treatment of the sample might lead to different results.

The third threat is regarding extrapolating the results. For example, if the aim of a study is to analyze the general sentiment towards several universities, it would be an error to equalize these results with the sentiment of each faculty within every university. Although both outcomes seem to be related to each other, are they not similar and thus will lead to varying results.

The fifth and last threat is concerning defining the causal pattern. The explanation for a specific causal relationship may vary from setting to setting, e.g. might a declining number of young professionals in Spain due the weak economy, while in Germany it is due the decreasing number of graduating students.

## 3.2 Application to Social Media Prediction

Inspired by Wijnhoven and Bloems (2014) paper on external validity of sentiment mining reports, the above mentioned threats can be applied to Social Media Predictions.Table1 at the end gives an overview of external validity threats within Social Media Predictions.

The first threat is concerning the sample used as a basis for the prediction model. Twitter, for example, is a very popular Social Media Site and therefore often a prediction tool alike. Yet, only a handful researcher took into consideration that the Twitter population might not be representative. And indeed, Mislove et al. (2011) came to the conclusion that Twitter users significantly over represent the densely population regions of the U.S., are predominantly male, and represent a highly nonrandom sample of the overall race/ethnicity distribution. Similar to these findings states Link (2013) that only 13% of the US Population is active on Twitter. In 2013 Couper writes about a possible selection bias in organic data. He points out that researcher working with Social Media focus too much on the "haves" and not enough on the "not haves". Even if this problem does not eliminate the validity of Social Media prediction, researcher have to be aware of this threat and be careful with generalizations towards the whole population.

The second threat relates to possible, unpredictable events which might have a negative influence on a prediction model. This threat is existing, because the phenomenon Social Media a still developing sometimes rapidly changing. For example, only few years ago was MySpace, without a doubt, the leading platform for Social Media. However due crucial mismanagement and the rise of alternatives, in 2015 the rank of MySpace in the Alexa Ranking (A website which rates and displays the accumulated traffic for many internet sites) for only 1576 compared to rank 2 for Facebook, rank 10 for Twitter and rank 14 for LinkedIn respectively. This threat is not par se due a possible rise or decline of a specific Social Media network in particular, rather a lack of stability. Nobody knows how the industry will develop in the future and which site will be dominant then. But there is also a threat for external validity on the macro level. For example may the legal framework concerning Social Media hinder or even make it impossible to collect and use data of user for third parties. Another related issue is privacy. Wilson et al. (2012) document the privacy options for Facebook since its release in 2004. Although this threat does not influence current models which are designed to predict outcomes only a few weeks in advance, it can become an issue if researcher try move a step further and attempt to predict outcomes more fare away in time.

The third threat indicates that a result of a specific prediction cannot be enlarged to different settings or products. An example would be the Lego- videogames. If a model is predicting the success for, e.g. the latest Lego: Star Wars game on the computer, the results from this specific research cannot be simply applied to another Lego game or to another gaming console like the Xbox from Microsoft. Beforehand of the research it has to be clear which aspects of a prediction she or he has to involve and which to avoid.

Threat number is four concerning the setting of the research. In case of Social Media Prediction it is obviously related to the different forms of Social Media platforms. Kietzmann et al. (2010) created a framework with the most prevailing aspect of each Social Media site and the differences among them. Furthermore there are certain distinctions about the function or purpose about Social Media. While there are, for example, sites like Facebook or Twitter which are more about representing yourself and sharing opinions with friends or followers, social networks like LinkedIn are more business-orientated and target probably a more sophisticated group of people. Therefore it would not be too surprising if the general sentiment about, e.g. a complicated economics book, extracted from LinkedIn would be different than the sentiment based on data extracted from Twitter. And indeed, Frické (2015) sees inductivism as a problematic issue for sentiment mining. He suggests a need for a theoretical framework which should make it possible to apply results from one setting into another in order to increase the external validity of these.

The last threat is related to the explanation of causal relationships within a setting between the data used in the model and the predicting based on these. This problem is a fundamental problem in sentiment mining in general (Pang & Lee, 2009; Missen et al., 2010; Wijnhoven & Bloemen, 2014) thus many articles, which are more empathizing the challenges of sentiment mining, are tackling this problem. Therefore I will not go into detail and refer to the above mentioned sources.

Table1: External Validity Threats for Social Media Predictions

| External Validity Threat | Applied to Social Media Prediction |
|---|---|
| Interaction of causal relations with units | Demographical Bias |
| Interaction of causal relations with treatments | Inconsistency of Social Media |
| Interaction of causal relations with outcomes | Biased Application of Results |
| Interaction of causal relations with settings | Settings Bias |
| Context-dependent mediations | Algorithm Bias |

## 4. LIMITATIONS AND DISCUSION

This work solely focuses on the conceptualization of external validity in terms of Social Media prediction, more precisely for prediction based on sentiment mining. However, further research would be necessary to validate my findings and in terms of theory and, more importantly. Within a practical application. As mentioned before in the introduction, was the aim of this paper to fill a gap of literature. Most likely there is a more elaborated approach than a literature review required to do so, but nonetheless does this paper underline the need for this task and can after all function as a ground for further research.

## 5. CONCLUSION

Opinion mining (or sentiment analysis) arouses great interest in recent years both in academia and industry (Brun, 2011). Therefore it is no surprise that literature made a lot of process in the last few year which lead to continuous improvements of the predictions based on Social Media. However, there seems to be a collective neglect of the external validity of prediction reports. This paper tries to conceptualize external validity in order to be applicable to the field. With this paper I do not want to, by any means, try to accuse previous work for lacking a decent methodology. Rather it attempts to provide basic guidance for future papers which might take a more precise look on external validity issues when designing prediction models and also present their results. Another important aspect is the so-called "file drawer effect" (Fanelli, 2010; Gayo-Avello, 2012). The tendency of researcher to only publish positive results, can have a very damaging effect for the research as it leads to an overly successful and too promising view. Although it is still a promising and interesting field, it is most likely not as progressive as many authors desire to. In order to be methodically sound and therefore being able to deliver useful predictions, a model does not only have to be able to provide satisfying result, but it also has to be externally valid.

## References

1. Asur, S., & Huberman, B. (2010). Predicting the Future with Social Media. *International Conference on Web Intelligence and Intelligent Agent Technology*, *01*, 492-499.
2. Babbie, E. (2009). *The practice of social research*. Belmont, Calif.: Wadsworth.
3. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal Of Computational Science*, *2*(1)
4. Brun, C. (2011). Detecting Opinions Using Deep Syntactic Analysis.
5. Calder, B., Phillips, L., & Tybout, A. (1982). The Concept of External Validity. *Journal Of Consumer Research*, *9*(3), 240. doi:10.1086/208920
6. Campbell, D., Stanley, J., & Gage, N. (1966). *Experimental and quasi-experimental designs for research*. Chicago, Ill.: R. McNally.
7. Choudhury, M., Sundaram, H., John, A., and Seligmann, D. 2008, “Can blog communication dynamics be correlated with stock market activity?”
8. Chung, J., & Mustafaraj, E. (2011). Can Collective Sentiment Expressed on Twitter Predict Political Elections?
9. Collier, N. (2012). Uncovering text mining: A survey of current work on web-based epidemic intelligence. *Global Public Health*, *7*(7), 731-749. doi:10.1080/17441692.2012.699975
10. Cook, T., & Campbell, D. (1979). *Quasi-experimentation*. Boston: Houghton Mifflin.
11. Cooper, C., Mallon, K., Leadbetter, S., Pollack, L., & Peipins, L. (2005). Cancer Internet Search Activity on a Major Search Engine, United States 2001-2003. *J Med Internet Res*
12. Couper, Mick (2015): „Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys“. In: Survey Research Methods. 7 (3), pp 145-156.
13. Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries.*Scientometrics*, *90*(3), 891-904. doi:10.1007/s11192-011-0494-7
14. Frické, Martin (2014): „Big data and its epistemology“. In: *Journal of the Association for Information Science and Technology*. 66 (4), pp. 651-661
15. Gayo-Avello, D. (2013): „A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data“. In: *Social Science Computer Review*. 31 (6), pp 649-679
16. Ginsberg, Jeremy; Mohebbi, Matthew H.; Patel, Rajan S. u. a. (2008): „Detecting influenza epidemics using search engine query data“. In: *Nature*. 457 (7232), pp 1012-1014

17. Gruhl D, Guha R, Kumar R, Novak J, Tomkins A (2005) The predictive power of online chatter. Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (ACM, New York), pp 78–87.

18. Goel, S., Hofman, J., Lahaie, S., Pennock, D., & Watts, D. (2010). Predicting consumer behavior with Web search. *Proceedings Of The National Academy Of Sciences*, *107*(41)

19. Kietzmann, J.H.; Hermkens, K.; Silvestre, B.S. (2011): „Get serious! Understanding the functional building blocks of social media". In: *Business Horizons*. 54 , S. 241-251

20. Lampos, V., & Cristianini, N. (2012). Nowcasting Events from the Social Web with Statistical Learning. *ACM Transactions On Intelligent Systems And Technology*, *3*(4), 1-22. doi:10.1145/2337542.2337557

21. McTavish, D., & Loether, H. (2002). *Social research*. Boston: Allyn and Bacon.

22. Mislove, A., Lehmann, S., Ahn, Y. Y., Onnela, J. P., & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. *ICWSM*, *11*, 5th.

23. Missen, M., & Cabanac, G. (2010). Opinion Detection in Blogs: What Is Still Missing?.

24. Monette, D., DeJong, C., & Sullivan, T. (2002). *Applied social research*. [Belmont, Calif. u.a.]: Wadsworth.

25. O'Connor, B., Balasubramanyan,, R., Routledge, B., & Smith, N. (2010). Linking text sentiment to public opinion time series.

26. Paul, M., & Dredze, M. (2012). A model for mining public health topics from Twitter.

27. Pang, Bo; Lee, Lillian (2008): „Opinion Mining and Sentiment Analysis". In: *Foundations and Trends® in Information Retrieval*. 2 (1–2)

28. Polgreen, P., Chen, Y., Pennock, D., & Nelson, F. (2008). Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases*, *47*(11)

29. Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013). The power of prediction with social media. *Internet Research*, *23*(5)

30. Rich, R., & OH, C. (2000). Rationality and Use of Information in Policy Decisions: A Search for Alternatives. *Science Communication*, *22*(2), 173-211. doi:10.1177/1075547000022002004

31. Shadesh W.R., William R., Cook, T. D., Campbell, D. (2001): *Experimentaland quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

32. Signorini, A., Segre, A., & Polgreen, P. (2011). The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *Plos ONE*, *6*(5), e19467. doi:10.1371/journal.pone.0019467

33. Tjong Kim Sang, E. (2014). Using Tweets for Assigning Sentiments to Regions.

34. Tjong Kim Sang, E., & Bos, J. (2012). Predicting the 2011 dutch senate election results with Twitter.

35. Tumasjan, A., Springer, T., Sander, P., & Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.

36. Tumarkin, R., and Whitelaw, R.F. 2001, “News or noise? Internet postings and stock prices”, Financial Analysts Journal, vol. 57, no. 3, p. 41.

37. Wang, H., Can, D., Kazemzadeh, A., Abe, F., & Narayanan, S. (2012). System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle.

38. Wijnhoven, F., & Bloemen, O. (2014). External validity of sentiment mining reports: Can current methods identify demographic biases, event biases, and manipulation of reviews?. *Decision Support Systems*, *59*, 262-273. doi:10.1016/j.dss.2013.12.005

39. Wilson, R., Gosling, S., & Graham, L. (2012). A Review of Facebook Research in the Social Sciences.*Perspectives On Psychological Science*, *7*(3), 203-220. doi:10.1177/1745691612442904

40. Winer, R. (1999). Experimentation in the 21st Century: The Importance of External Validity. *Journal Of The Academy Of Marketing Science*, *27*(3), 349-358. doi:10.1177/0092070399273005

41. Wüthrich, B., Permunetilleke, D., Leung, S., Cho, Zhang, J., and Lam, W. 1998, “Daily prediction of major stock indices from textual WWW data”, in Proceedings of KDD-98.