ICMA Centre

Henley Business School
UNIVERSITY OF READING

Discussion Paper

Tweeting About Sustainability: Can Emotional Nowcasting Discourage Greenwashing?

February 2017

Andreas G F Hoepner

ICMA Centre, Henley Business School, University of Reading
Mistra Financial Systems, Stockholm School of Economics

Savio Dimatteo

SAP SE, Walldorf, Germany

Joé Schauld

University of St Andrews, Scotland

Pei-Shan Yu

ICMA Centre, Henley Business School, University of Reading
Mistra Financial Systems, Stockholm School of Economics

Mirco Musolesi

University College London, UK

Discussion Paper Number:
ICM-2017-02

The aim of this discussion paper series is to disseminate new research of academic distinction. Papers are preliminary drafts, circulated to stimulate discussion and critical comment. Henley Business School is triple accredited and home to over 100 academic faculty, who undertake research in a wide range of fields from ethics and finance to international business and marketing.

admin@icmacentre.ac.uk

www.icmacentre.ac.uk

© Hoepner et al, February 2017

Tweeting About Sustainability: Can Emotional Nowcasting Discourage Greenwashing?

Abstract

Fewer than 100 firms worldwide are recognised by Bloomberg to report accurate greenhouse gas emissions. Yet, tens of thousands of people are talking and tweeting about climate change every day. How can this attention be converted into accurate action? We propose that sustainable data science might help, specifically that '*emotional nowcasting*' of societal responses to sustainability related statements as expressed on Twitter. First, we differentiate between various types of corporate sustainability performance data and highlight the challenge that corporate greenwashing and a potential lack of financial independence of the assessor from the assessed poses for these data sets. Second, we introduce the concept of emotional nowcasting with two case studies of an emotionally non-ambivalent context, the football matches England vs. Germany and England vs. USA at the 2010 world cup. These case studies serve as a proof of concept for emotional nowcasting. Finally, we discuss the potential for emotional nowcasting to mitigate the pandemic of greenwashing currently experienced in sustainability communication. We conclude that emotional nowcasting can serve as one test of greenwashing which is on its own though not necessarily sufficient.

Keywords

Accounting Independence, Corporate Governance, Emotional Nowcasting, Sustainable Development, Twitter

Acknowledgements

We are very grateful to Alexander Bassen and Ralf Frank for the encouragement to work on such an innovative topic. Joé Schaul has left St Andrews since his work on this project and is now employed in the private sector. All remaining errors are our own.

Contact

Andreas G F Hoepner: a.hoepner@icmacentre.ac.uk

1 Introduction

When even Donald Trump tweets to postpone his party's planned assault on a key ethics committee,[1] it is certainly time for the responsible investment industry to take a clearer position on Greenwashing. Why? Greenwashing is a pandemic governance problem of the industry that professionally communicated corporate sustainability information, where well-meaning people get paid by companies to eventually report information that makes the company look as good as possible and may but also may not be accurate (Delmas and Burbano, 2011, Lane, 2013, Vos, 2009). Statistically, of the 1,000+ CO2e Scope 1 and Scope 2 disclosing companies, far less than 10% accurately disclose 100.0% of their emissions (Adamsson *et al.*, 2016, Liesen *et al.*, 2015). The rest green-washes their image, sometimes less, often more. So how can responsible investors tackle this greenwashing pandemic?

Two immediate proposals emerge. First, as demanded by Adamsson *et al.* (2016), a useful rule of thumb is to always follow the Precautionary Principle (Rio 1992) when interpreting sustainability information. Causally phrased, this can be stated as: "If in doubt, err on the side of the planet, not on the side of your company." Second, responsible investor can acquire commercial data sets on the sustainability performance of corporations. However, the failure of several agencies to commit to the recently launched Deep Data Delivery Standards[2] raises the question, how independent these rating agencies are from corporate payments?

2 Corporate sustainability data

For responsible investors, the task of identifying an honestly reporting company is not as easy as it may sound, as companies have all incentives to show their sunny side and little reason to voluntarily display shadows. Supposedly, sustainability rating agencies advise investors on the quality of corporate sustainability disclosure but then several sustainability rating agencies also have business units consulting corporations on their sustainability performance. In other words, some agencies may be expected to bite the hand that feeds them.

In general, sustainability data can be classified into four types of information sets:

(i) information disclosed by issuers in their financial, sustainability or SEC reporting (e.g. provided by Bloomberg),

[1] http://edition.cnn.com/2017/01/02/politics/office-of-congressional-ethics-oversight-of-ethics-committee-amendment/
[2] www.DeepData.ai

(ii) responses of issuers to surveys of third parties (e.g. CDP),

(iii) due diligence assessments by third party rating agencies (e.g. MSCI)

(iv) data aggregations of independent external views on corporations (e.g. negative news turned into RepRisk's Reputation Risk Index).

With exception of type (iv) which relies on the availability of third party information, all of these types are at significant risk of being positively biased by corporate greenwashing. In fact, even type (iv) is at some risk of bias, if corporations manage to bury negative information in the deep web (i.e. out of the reach of search engines) while they repeatedly portray their positive side in the surface web.

To address this greenwashing challenge and make sustainability information as well as any third party data more useful, a group of dozens of academics and professionals launched the Deep Data Delivery Standards in September 2016. Apart from ensuring the machine readability of third party data, the standards, especially 6 and 7, directly address conflicts of interest in the production of third party sustainability data. According to these standards, "Deep data sets are expected to be delivered...

1. ... with a minimum of 5 years historical data on at least 30 independent indicators per data set (e.g. credit rating, ESG data) whereby any data point that is not delivered as reported at the respective point in time should be flagged as backfilled;

2. ... with 98% value weighted market coverage, where a market (e.g. equity index) is claimed to be covered;

3. ... with an assurance that ratings will be re-considered for at least 8.25% of the companies covered in the average month of the following year;

4. ... including considerate, accurate identifiers (eg ISINs) for 99% of the firms covered in every month of current and historical data coverage;

5. ... in machine readable format (eg CSV, XML) and with proper documentation of the data structure[3];

[3] Proper documentation may be understood as documenting each row and column or the data structure including relevant definitions, third party providers and methodological descriptions as well as any relevant adjustments to definitions, third party providers and methodologies that might have occurred over time.

6. ... with an assurance of individual rating independence meaning that none of the rated entities in the respective market (e.g. equity index) financially contributed to their rating or paid for access solely to their own rating;

7. ... with an assurance of organizational rating independence meaning that whenever rating agencies win entities as new clients which they also rate an independent analysis[4] is conducted if these new clients receive, statistically significant[5], higher ratings than in the year before and any biases found in this analysis will be addressed within 12 months;

8. ... with an assurance that all research or rating reports in the following year will indicate names and office locations of all analysts substantially involved in the analysis[6] as well as the extent to which their data sources exceed those self-reported by the rated entity[7];

9. ... with an assurance that all research or rating reports in the following year will include a logbook detailing any errata, where applicable, as well as the dates and roles of participants in communication with the rated companies;

10. ... accompanied by the ratio of the rating agency's research costs to total cost[8] or the ratio of research head count to total head count in the most recent financial year." [www.DeepData.ai, 2016:1]

If sustainability rating agencies are not willing to publicly or in private agreements with their clients adhere to these standards, especially standard 6 and 7, investors have to ask themselves, if some of the companies the rating agencies are assessing may have more or less directly paid for the information which the investors are intending to consume. In other words, without a clear commitment of a sustainability rating agency to financial independence of the assessed entity, the sustainability ratings are biased in favour of the specific companies.

This challenge of insufficient financial independence of the assessor from the assessed entity, however, does not only exist for corporate sustainability ratings or green bond ratings, it actually originates from mainstream credit ratings, for which significant evidence of bias exist (Poon, 2003, Baghai and Becker, 2016, Bolton *et al.*, 2012). The demise of Arthur Andersen showed very

[4] Independent analysis may be understood as an analysis by a third person or thirty party that was neither directly involved in the client acquisition process nor the rating process.
[5] The analysis may use 1%, 5% or 10% as common statistical significance levels.
[6] Whenever a rating process is fully automated, a rating agency may indicate the data scientist(s) substantially involved in designing the rating process.
[7] This extent may be communicated by classifying data that is self-reported by the rated entity as representing (i) all, (ii) most, (iii) about half, (iv) some or (v) none of the information underlying the assessment.
[8] Research costs or head count includes staff costs of researchers, staff providing or costs of supplies needed for research and IT related to research (i.e. data processing and data delivery).

practically how significant the conflict of interests are in the Anglo-American audit model, despite corporations being mandated to report the audited financial information. The only exception may be regulated indicators which give the company little discretion for massaging such as operating cash flows. But even with respect to these indicators, forensic accountant tend to flag the accruals based timing opportunities though have hope that the newly introduced extended audit reports may enhance matters. In summary, investors may need more technology beyond in-depth accurate analysis.

This need may indeed be addressed using sustainable data science or "green data science" as Van der Aalst (2016) calls it. Conceptually, the increase in popularity of online social networks such as Twitter1 leads to an ever-growing amount of data about human activity. Since computational analysis of this data can yield insights into human social activity, we can nowadays perform studies to measure the emotion of a target group of users (in our case defined by geographic region) on a certain issue.

The key technology element used here is automated sentiment analysis of text. This technique allows us to first train algorithms2 by giving them positive, negative and neutral sentences, after which they can then classify new, unseen sentences into a class of negative (-1), positive (1) and neutral (0). While natural language analysis is always hard for computers (due to language challenges such as misspellings, the lack of context, use of irony, etc), an accuracy of roughly 70% on one test set is nevertheless sufficient to yield some very interesting results. Since such automated sentiment analysis could be undertake in near real time, we call it emotional nowcasting. To provide a the proof that such automated sentiment analysis practically works, the next section display two case studies from the world of sports, where the real world emotion is very much predictable. Our sustainable data science methodology is illustrated in Figure 1 and explained in the subsequent section.

3 Sustainable data science methodology

3.1 Acquiring initial data

Twitter allows its users to specify whether their tweets are public or can be seen only by their friends. Public tweets can be collected by anyone, but there is a limit imposed on the amount of

tweets any user can collect. We therefore operated on data sets collected by other people, such as the service offered by TwapperKeeper[9].

3.2 Adding geographic information about users

Once we have the user names or user IDs of the relevant twitter users, we can query the Twitter API (2016) to find the registered location of these users. These locations are 90% of the time in plain text and may contain any string (e.g. "London", "St. Andrews, Scotland, UK", "Reading, Berkshire" or "Somewhere over the rainbow..."), but some of them have precise latitude/longitude coordinates (e.g. "52.346, 13.398"). From the locations which are useful and unambiguous, we can then group users by country, state, or city.

Grouping users by geographic region is done primarily through keyword matching, where we compare keywords in the location string with the set of all city and town names of a specific state or country. This set of place names is currently obtained using the Geonames (2016) database, and pre-processed to avoid most ambiguities such as London, which could be "London, UK" or "London, Ontario, Canada".

Additionally, the location of further users can be determined by checking if their coordinates are within the desired region.

3.3 Determinging the average emotion over time through sentiment analysis

Sentiment analysis algorithms such as the Naive Bayes are able to analyse the text of a tweet and classify them as either positive (1), negative (-1) or neutral (0). This requires the algorithm to first be trained on a training set, which consists of tweets and a manual ranking of whether the tweet carries a positive, neutral or negative emotion. We chose a Naive Bayes classification algorithm, as it has been found that it generally provides some of the best results (Go *et al.*, 2009, Choi and Cardie, 2008, Pak and Paroubek, 2010, Parikh and Movassate, 2009). An accuracy of about 70% was achieved on a test set of tweets.

These classifications of tweets can then be averaged over a time window, resulting in a decimal score. This score can subsequently be plotted over time, showing the evolution of the average emotion in a population of twitter users from a specific location. A sliding window smoothing

[9] As of January 9th, 2012 TwapperKeeper is now fully integrated with the HootSuite Twapperkeeper. (2016). HootSuite (Formerly Twapperkeeper) [Online]. Available: https://hootsuite.com/ [Accessed 2010]. dashboard.
http://twapperkeeper.com/index.html

can be used to more clearly see the important trends and reduce the noise of short-time fluctuations.

4 Case studies: England vs. Germany and England vs. USA at the world cup 2010

We first try to see how well current sentiment analysis methods can perform for events with predictable emotion, whereby we use the national pride of football supporters as a clear cut predictor of the expected emotion. For the experimental case study, we use about 34,000 tweets3 containing the hashtag4 #world cup during the world cup game Germany-England. For the second game studied, USA-England, we use a different data set and only had about 1600 tweets from the USA and 600 from England, which means results are less representative.

The sentiment score on the y axis is obtained by taking an average over the positive (+1) and negative (–1) tweets in a given time frame. If all tweets were positive, the sentiment score would reach its maximum value or 1. If all tweets were negative, it could fall to –1. The more neutral tweets exist, the more the sentiment score is drawn towards the value of zero. The results of our first case analysis are plotted in Figure 2. Please note that while German goals have clear correlations on the English emotion (in the 4 minutes and even 1 minute interval), the German emotion is a bit less clear, which is most likely due to the classifier not being so good at measuring emotion in tweets written in German, as it was trained only on English tweets. Also, this data is based on an uneven amount of tweets from the nations, with roughly 17000 tweets from England but only 7000 from Germany (as people tweeting in German were less likely to include the hashtag #worldcup, but rather included another hashtag in their language).

Our initial goal was to study all 32 games to have measurements of statistical relevance. However this proved impossible for 2010 data, as we working from inconsistent data sets which were more complete in certain time periods than others, so we simply did not have the data to do this. Further, analysis of emotion is made harder when several languages are involved, as (A) different nations use different hashtags, (B) it is hard to train algorithms on multiple languages, as fewer people have done previous sentiment analysis work on languages other than English. Nevertheless, the results of our second case study, the England vs. USA game plotted in Figure 3 are still shows reasonable results, in particular with respect to the goal scored by USA in the 40th minute. If 2010 data allows to already emotionally nowcast a football game of mediocre relevance such as England vs. USA, most events that command reasonable attention should be nowcast going forward as the social media space appears to be ever increasing.

Specifically for the world cup games, our data set is only considering tweets written in English with the hashtag #worldcup, which is not representative for users from non-English speaking countries, as they use different hashtags. As grouping users by geographic location and automatic text analysis are not precise processes results have to be considered with a margin of error.

Further analysis on different issues with sufficient attention promises to hold some potential, especially considering that the measured emotional data can then be related to various other types of data which are affected by human emotion. As shown by Asur and Huberman (2010), the sheer volume of tweets already holds strong predictive power in the movie business. It has been shown by O'Connor *et al.* (2010) that there is a strong correlation between sentiment of tweets and polls. And perhaps most interestingly, Bollen *et al.* (2011)show a strong correlation of public mood (measured via tweets) with the Dow Jones Industrial Average. Climate change certainly demands sufficient attention, as shown by Barkemeyer *et al.* (2017) or simply by the Climate Change Attention Index.[10]

5 Concluding discussion

This initial study of a nation's emotion in the context of a world cup game is just the beginning of what emotional nowcasting may be able to do in our view. It is meant to demonstrate the power that large amounts of public data combined with automated analysis cab have. Despite the many limitations of automated sentiment analysis on public twitter messages and the small scope of this study, we have nonetheless two important results: (i) Emotion of one subset of the population can be measured and compared to that of another subset within a given context, and (i) with enough data, this can be done over very short time periods of just minutes, which can be important as emotions tend to fluctuate rapidly.

So does this potential of emotional nowcasting have the ability to mitigate greenwashing? We believe that there are two reasons in favour of this potential but also two reasons for caution. First, sentiment analysis on social media data can certainly reveal with type of sustainability information the general public believes and which one it distrusts. Similar to the analysis of negative news published in the media, such a big data analysis of societal distrust is likely to identify a significant proportion of the more obvious corporate attempts to greenwash. Second, the simply awareness of corporations that their most simplistic greenwashing practices can be identified ex-post via sentiment analysis or even in near real time via emotional nowcasting

[10] http://ccai.sociovestix.com/en/

should prevent various obvious forms of greenwashing, as corporations retreat from the practices since the reputational risks have become bigger than the reputational benefits.

However, the year 2016 has taught the world a Churchillian lesson about the intellectual scrutiny of the average social media user. This implies that sentiment analysis will either have to weight individual users by their historical ability to identify deceptions, as the less obvious cases of greenwashing are unlikely to be identified through an analytical set up relying on the intellect of the average voter. Consequently, we do not see emotional nowcasting in itself as sufficient to address the greenwashing pandemic. However, it is likely to be a useful tool alongside industry initiatives such as the Deep Data Delivery Standards and simply a sharp and somewhat cynical mind set of the data scientist tasked with cleaning a sustainability data set from greenwashing bias.

References

Adamsson, H, Hoepner, A G F & Yu, P-S (2016). Towards a carbon data science. *Henley Business School Discussion Paper* ICM-2016-01.

Asur, S & Huberman, B A (2010). Predicting the future with social media. In: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, 2010. IEEE, 492-499.

Baghai, R & Becker, B (2016). Non-rating revenue and conflicts of interest, *CEPR Discussion Paper* No. DP11508,

Barkemeyer, R, Figge, F, Hoepner, A, Holt, D, Kraak, J M & Yu, P-S (2017). Media coverage of climate change: An international comparison. *Environment and Planning C: Politics and Space*, 0263774X16680818.

Bollen, J, Mao, H & Zeng, X (2011). Twitter mood predicts the stock market. *Journal of Computational Science,* 2 (1): 1–8.

Bolton, P, Freixas, X & Shapiro, J (2012). The credit ratings game. *Journal of Finance,* 67 (1): 85–111.

Choi, Y & Cardie, C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2008*. Association for Computational Linguistics, 793–801.

Delmas, M A & Burbano, V C (2011). The drivers of greenwashing. *California Management Review*, 54 (1): 64–87.

Geonames.Org (2016). GeoNames Data [Online]. Available: http://www.geonames.org/export [Accessed 2016].

Go, A, Bhayani, R & Huang, L (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report*, Stanford, 1 (12).

Lane, E L (2013). Greenwashing 2.0. *Columbia Journal of Environmental Law*, 38 (2): 280–331.

Liesen, A, Hoepner, A G, Patten, D M & Figge, F (2015). Does stakeholder pressure influence corporate GHG emissions reporting? Empirical evidence from Europe. *Accounting, Auditing & Accountability Journal*, 28 (7): 1047–74.

O'Connor, B, Balasubramanyan, R, Routledge, B R & Smith, N A (2010). From tweets to polls: Linking text sentiment to public opinion time series. In: *Fourth International AAAI Conference on Weblogs and Social Media, 2010*. AAAI Publications, 1–2.

Pak, A & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: LREc, 2010.

Parikh, R & Movassate, M (2009). Sentiment analysis of user-generated twitter updates using various classification techniques. CS224N Final Report, 1–18.

Poon, W P (2003). Are unsolicited credit ratings biased downward? *Journal of Banking & Finance*, 27 (4): 593–614.

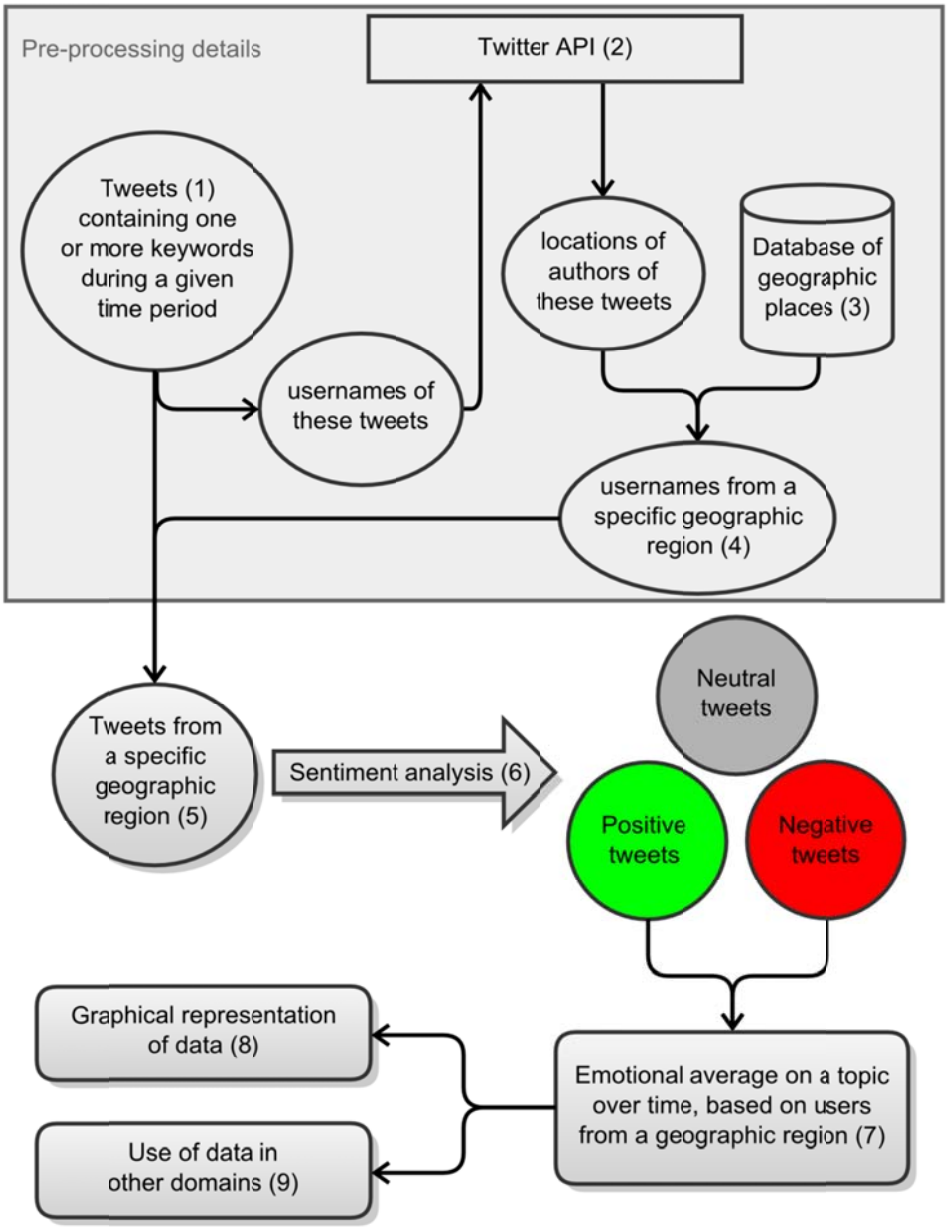
Twapperkeeper (2016). HootSuite (formerly Twapperkeeper) [Online]. Available: https://hootsuite.com/ [Accessed 2010].

Twitter. (2016). Twitter api [Online]. Available: http://apiwiki.twitter.com/ [Accessed 2016].

Van Der Aalst, W M P (2016). Green data science using big data in an "environmentally friendly" manner. In: *18th International Conference on Enterprise Information Systems*, 2016 Rome, Italy.

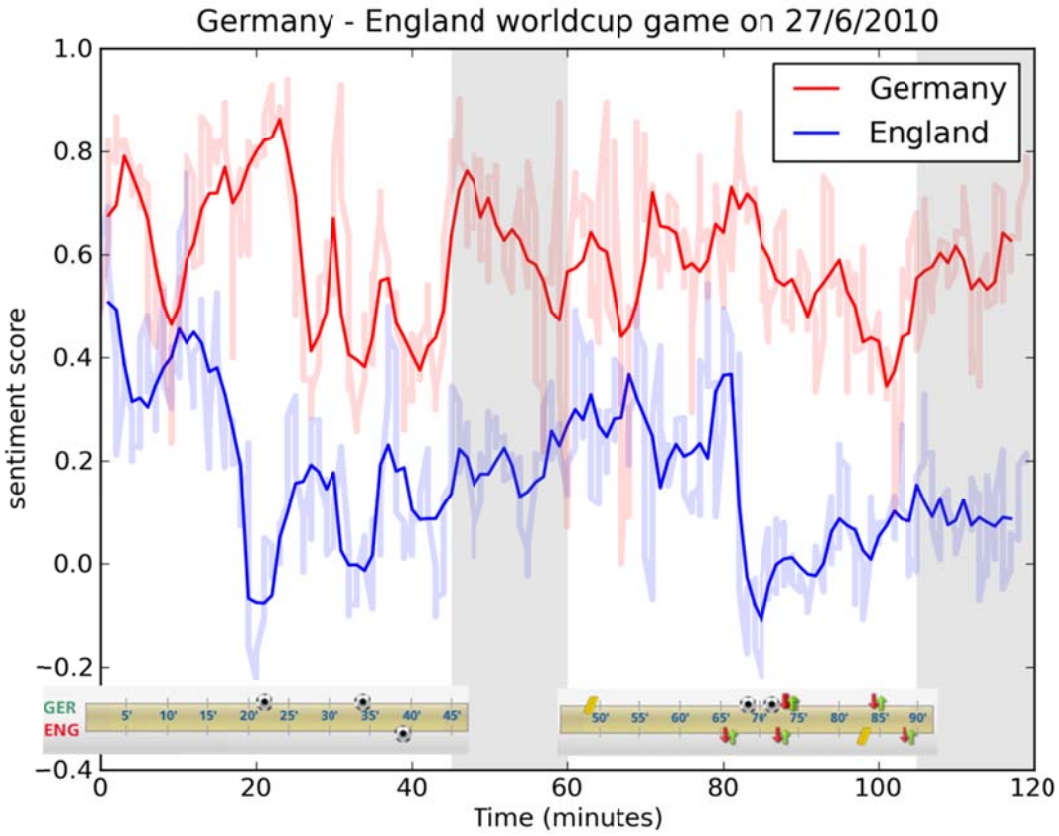
Vos, J (2009). Actions speak louder than words: Greenwashing in corporate America. *Notre Dame JL Ethics & Public Policy*, 23, 673–97.

Figure 1: Details of how emotion data is obtained.



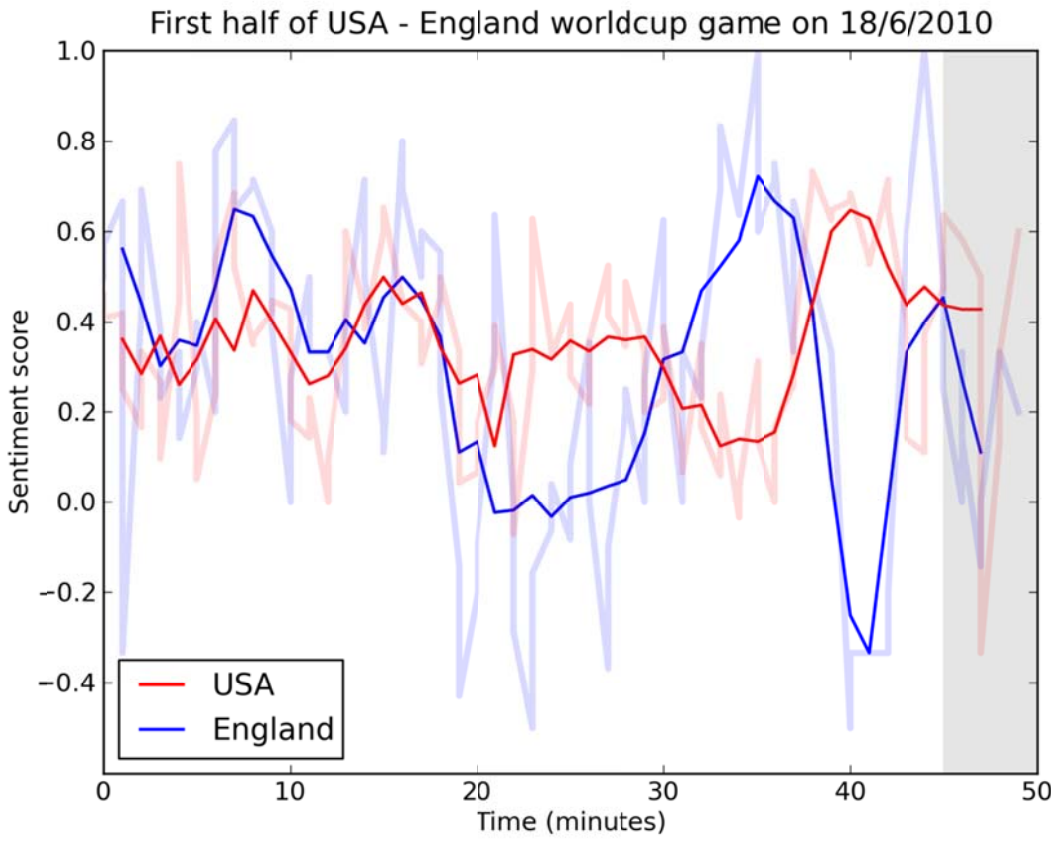
Tweets (1) contain not only messages, but also a time stamp, and a username. These usernames can then be used to query (2) the Twitter Application Programming Interface [1], a service offered by Twitter to allow computer programs to search for certain tweets or users. This allows us to obtain the geographic location of users, which can then be filtered by country or region using a geographic database (3) such as Geonames [2]. From the filtered usernames (4), we can then only take a subset of tweets (5) which have been written by these users. On those we perform sentiment analysis (6) (see section 3.3) to obtain tweets with positive, negative, or neutral emotions. We then take a simple average of positive and negative ones (7) over time, from which we can either plot a graph (8) or use the data further (9).

Figure 2: Emotions measured of people living in Germany and England (excluding the rest of the UK) during their nations' world cup game in 2010.



Data coming from a total of ~34000 tweets. Moving average MA_t of sentiment score smoothed over different time windows: 1 minute (light lines) and 4 minutes (dark lines). White background shows actual playing time, and FIFA match cast was added for illustration. This shows clearly the expected effects of English mood dropping after a German goal in minutes 20, 32, 67 and 72.

Figure 3: Emotions measured of people living in the United States and England (excluding the rest of the UK) during their nations' world cup game in 2010.



Data coming from a total of only ~2200 tweets. Moving average MA_t of sentiment score smoothed over different time windows: 1 minute (light lines) and 4 minutes (dark lines). White background shows actual playing time. Only first half of the game is shown due to lack of data for the second half, also there was overall less data available for us during this time period, which explains the more erratic lines. While the English goal at the 4th minute is less clear, the impact of the US goal in minute 40 can be seen clearly with a vast drop of English sentiment and an increase in US emotion.