Twitter Analytics for the Horticulture Industry

Marijn ten Thij, Sandjai Bhulai
Vrije Universiteit Amsterdam,
Faculty of Sciences,
Amsterdam, The Netherlands
Email: {m.c.ten.thij, s.bhulai}@vu.nl

Wilco van den Berg
GroentenFruitHuis,
Zoetermeer, The Netherlands
Email: vandenberg@groentenfruithuis.nl

Henk Zwinkels
Floricode,
Roelofarendsveen, The Netherlands
Email: h.zwinkels@floricode.com

Abstract—**In our current society, data has gone from scarce to superabundant: huge volumes of data are being generated every second. A big part of this flow is due to social media platforms, which provide a very volatile flow of information. However, leveraging this information, which is burried in this fast stream of messages, poses a serious challenge. A vast amount of work is devoted to tackle this challenge in different business areas. In our work, we address this challenge for the horticulture sector, which has not received a lot of attention in the literature. Our aim is to extract information from the social data flow that can empower the horticulture sector. In this paper, we present our first steps towards this goal and demonstrate key examples of this empowerment.**

Keywords–Twitter; horticulture; social media analytics.

I. Introduction

In recent years, there have been a lot of overwhelming changes in how people communicate and interact with each other, mostly due to social media. It has revolutionized the Internet into a more personal and participatory medium. Consequently, social networking is now the top online activity on the Internet. With this much subscriptions to social media, massive amounts of information, accumulating as a result of interactions, discussions, social signals, and other engagements, form a valuable source of information. Social media analytics is able to leverage this information.

Social media analytics is the process of tracking conversations around specific phrases, words or brands. Through tracking, one can leverage these conversations to discover opportunities or to create content for those audiences. It is more than only monitoring mentions and comments through social profiles, mobile apps or blogs. It requires advanced analytics that can detect patterns, track sentiment, and draw conclusions based on where and when conversations happen. Doing this is important for many business areas since actively listening to customers avoids missing out on the opportunity to collect valuable feedback to understand, react, and provide value to customers.

The retail sector is probably the business area that utilizes social media analytics the most. More than 60% of marketeers use social media tools for campaign tracking, brand analysis, and for competitive intelligence [1]. Moreover, they also use tools for customer care, product launches, and influencer ranking. Social media analytics is also heavily used in news and journalism for building and engaging a news audience, and measuring those efforts through data collection and analysis. A similar use is also adopted in sports to actively engage with fans. In many business areas one also uses analytics for event detection and user profiling. A vast amount of work is devoted to tackle the challenges in the mentioned business areas. In our work, we address this challenge for the horticulture sector, which has not received a lot of attention in the literature.

The horticulture industry is a traditional sector in which growers are focused on production, and in which many traders use their own transactions as the main source of information. This leads to reactive management with very little anticipation to events in the future. Growers and traders lack data about consumer trends and how the products are used and appreciated. This setting provides opportunities to enhance the market orientation of the horticulture industry, e.g., through the use of social media. Data on consumer's appreciation and applications of products are abundant on social media. Furthermore, grower's communication on social media might indicate future supply. This creates a need for analytic methods to analyze social media data and to interpret them.

In this paper, we present our first steps towards this goal and demonstrate key examples of this empowerment. We start with discussing related research in Section II. In Section III, we describe our dataset that we used in the analysis. Next, we present the results of our data analysis in Section IV. Finally, we conclude the paper in Section V with some discussion and future work.

II. Related Research

Many studies have focused on detecting trends and/or events using data obtained from Twitter, examples are building a prediction system for the number of hit-and-runs [2], or detecting locations of earthquakes [3], or detecting flu spreads and activity [4], [5]. Currently, researchers have extended their scope to a wide range of fields where these methods are being applied.

An example of such a field is journalism. In one of the first descriptive works analyzing the network and content of Twitter, Kwak et al. [6] found that it is mainly used for News (85% of the content). This lead to further analysis of the spread of news in Twitter. For instance, [7] studies the diffusion of news items in *Twitter* for several well-known news media and finds that these cascades follow a star-like structure. Furthermore, [8] studies the life cycle of news articles posted online and describes the interplay between website visit patterns and social media reactions to news content. Through this, the authors show that this hybrid observation method can be used to characterize distinct classes of articles and find that the overall traffic that the articles will receive can be modeled accurately. Also, [9] focuses on Twitter as a medium to help journalists and news editors rapidly detect follow-up stories to the articles they publish and they propose to do so by leveraging transient news crowds, which are loosely-coupled groups that appear in Twitter around a particular news item, and where transient here reflects the fleeting nature of news.

Another active field of study where Twitter is used is in the sports industry. For instance, [10] examines the effectiveness of using a filtered stream of tweets from Twitter to automatically identify events of interest within the video of live sports transmissions. They show that using just the volume of tweets generated at any moment of a game actually provides a very accurate means of event detection, as well as an automatic method for tagging events with representative words from the twitter stream. Also, [11] investigates to what extent we can accurately extract sports data from tweets talking about soccer matches and show that the aggregation of tweets is a promising resource for extracting game summaries. Building on the knowledge of these previous works, [12] describes an algorithm that generates a journalistic summary of an event using only status updates from Twitter as a source. Finally, [13] uses sentiment analysis on tweets of players and other data to model performance of NBA players.

A last example of a sector where Twitter data can be used is the retail industry. For instance, [14] analyzes the perceived benefits of social media monitoring (SMM) and finds that SMM enables industrial companies to improve their marketing communication measurement ability. Also, [15] develops a framework for leveraging social media information for businesses, which focuses on sentiment benchmarks. Furthermore, [16] analyzes the use of social media by destination marketing organizations and finds that they are exploring several ways to leverage social media. Finally, [17] discusses the setup and the key techniques in social media analytics and gives an overview of the possibilities for the industry.

In this work, we use Twitter to empower the horticulture industry by analyzing topic-relevant tweets.

III. Dataset

In this section, we describe how we obtained the tweets that we use in our analysis. These tweets are scraped using the filter stream of the Twitter Application Programming Interface (API) [18]. We use two streams (called "Netherlands (NL) general" and "NL specific"), which are set up with the goal to scrape as many Dutch tweets as possible.

In the "NL general" stream, we use filter stream with the option **track**, where a list of words must be defined. All tweets containing one of these words are caught. We define a list of general Dutch words (e.g., *'een, het, ik, niet, maar, heb, jij, nog, bij'*, which translates to *'a, the, I, not, but, have, you, still, with'*). In total, this list consists of 130 words. Then, "NL specific" uses the filter stream combining **track** and **follow**. For this option, we add a list of user IDs for which all tweets are caught. For this list, we include local news outlets and other news-heavy Twitter accounts, such as neighborhood police accounts. In total, we define a list of 1,303 users. Examples of accounts are *@NUnl,@TilburginBeeld*. The list of terms consists of 395 entries (e.g., *'brandweer, politie, gewond, ambulance'* which translates to *'fire brigade, police, injured, ambulance'*). Note that any caught tweet may be contained in multiple streams, we have not distinguished duplicates in our dataset.

Since we do not have access to the Twitter Firehose, we do not receive all tweets that we request due to rate limitations by Twitter [19]. To give an insight in the volume of tweets that we analyze, we display the total number of tweets that we received on a monthly scale in Table I, which shows that the number of tweets we receive per month is steady throughout the year. All months contain at least 30 million tweets that contains the keywords we used. Using such a large tweet dataset allows us to do studies into less popular topics, without a large loss of accuracy.

Table I. Number of received tweets from August 1st 2014 to August 1st 2015.

	Received
August 2014	36,364,128
September 2014	29,418,425
October 2014	32,298,112
November 2014	36,597,687
December 2014	37,065,838
January 2015	38,377,668
February 2015	34,454,250
March 2015	38,722,934
April 2015	33,931,939
May 2015	34,492,494
June 2015	33,274,631
July 2015	31,078,206
Total	416,076,312

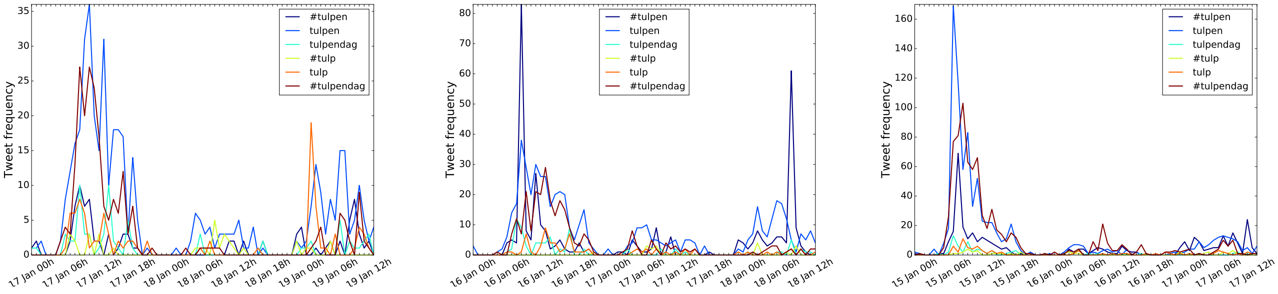
For the real-time analysis of Twitter-information, we set up our own Twitter scraper using the Twitter API. Here, we used a list of product names, provided by our partners from GroentenFruitHuis and Floricode, to locate possibly interesting tweets. Again, we tailored these lists to acquire Dutch tweets and used to language filter provided by the Twitter API to ensure that all received tweets are indeed Dutch. The terms are split up into two lists, one containing fruits and vegetables, e.g., apple, orange, and mango, and the other containing flowers and plants, e.g., tulip, rose, and lily.

IV. Data studies

After retrieving the tweets, the first step towards empowering the sector is knowing what kind of information can be retrieved from the social feed. To investigate this, we performed several studies, for which we present the results below. These studies range from an analysis of what is discussed at the current time in Twitter, to a study into the frequency of postings with regards to particular products, to an analysis of the online footprint caused by real-life events.

A. GfK time series

For this analysis, we have used the year long data of Twitter messages from August 2014 to August 2015 and counted the number of occurrences of fruit products mentioned in the text of the tweets. From these mentions, we construct a weekly time series reflecting the number of mentions. Then, we compared these numbers to sales numbers of the most occurring product



(a) Tulpendag in 2014 (b) Tulpendag in 2015 (c) Tulpendag in 2016

FIGURE 1. TULIP DAY TWEET FREQUENCIES DURING THE DAY ITSELF AND THE TWO DAYS AFTER THE EVENT.

type of these products. Thus, we compared the number of Dutch tweets mentioning 'pears' or 'pear' to the number of Conference pears that are sold in the same time-frame. Similarly, we compared the number of tweets mentioning 'apple' or 'apples' to the number of Elstar apples that are sold. In Figure 2, we present these time series, normalized on the maximum value in the series. In this figure, we see that in both cases the time series for the tweets and the sales are comparable. Using an eight-hour shift for the sales time series, we find a Pearson correlation coefficient of 0.44 for the apples series and a coefficient of 0.46 for the pears series.

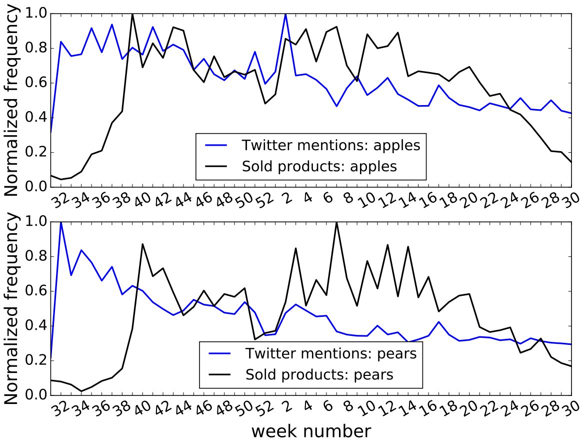


FIGURE 2. WEEKLY NORMALIZED TWITTER MENTIONS AND SALES.

These results indicate that it could be possible to predict the sales of a product type eight weeks in advance, however, this will need to be confirmed using other product types.

B. *Tulip day*

As a kick-off of the tulip season, which indicates the period in which a large variety of tulips is available at the vendors, the Dutch tulip growers organize the so-called Tulip day in Amsterdam. During this day, growers place a field of a large range of different tulips on the Dam in Amsterdam. During a few hours in the afternoon, visitors are allowed to pick the tulips in this field. For this event, we analyzed the presence on Twitter over a few years. For 2014, 2015, and 2016, we counted the occurrences of the tags 'tulip', 'tulips', and 'tulip day', both during the event and the two days after the event. The results of this study are presented in Figure 1. Here, we clearly see an increase in the number of mentions of the Tulip day on Twitter, this increase in mentions coincides with an intensified campaign by the growers and the governing body to broaden the attention to the Tulip day event. This analysis shows that the impact of the effort by marketing can be measured through Twitter analysis.

C. *Top 10 rankings*

The studies we discussed so far are based on a static collection of tweets. Another approach to extract value from Twitter is to see what a continuous feed of messages contains. To better understand how the products we are interested in are discussed on Twitter on a real-time basis, we set up two lexicons as described at the end of Section III. Using these lexicons, we scan Twitter in real-time for tweets that match these products. Using the tweets we receive this way, we can analyze what is currently discussed about these products. As a first step, we visualize this information in a top-10 application. The main page of this application shows the top 10 most discussed products on Twitter in the last day for both fruits/vegetables and plants/flowers, of which an example is shown in Figure 3. By clicking on one of the products in these top lists, we are redirected to a page, shown in Figure 4 for bananas, which shows us both the current messages mentioning the product and a detailed analysis of these messages, e.g., in terms of the most occurring tokens and a time series in which the terms are mentioned.

D. *Story detection*

The real-time data is also used for the detection of stories and discussions that suddenly pop-up. We do this by clustering incoming tweets by their tokens, using the Jaccard index. For sets A and B, this index equals

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

If two tweets are more similar than a predefined threshold, which we set at 0.4, then these two tweets will be represented by the same cluster. Therefore, if a topic is actively discussed on Twitter, it will be represented as a cluster in our story detection. Since these clusters are renewed every hour we add

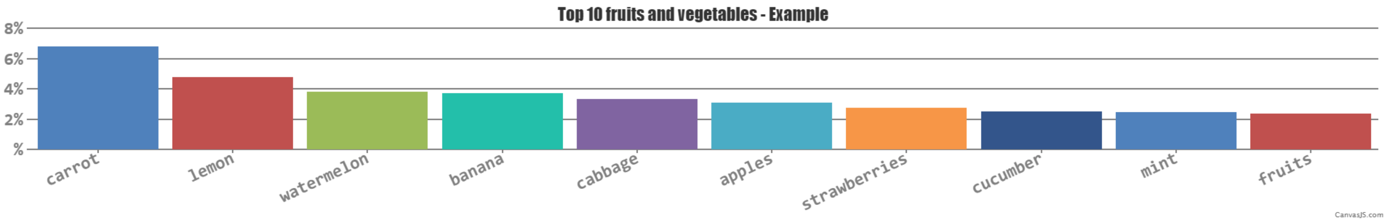
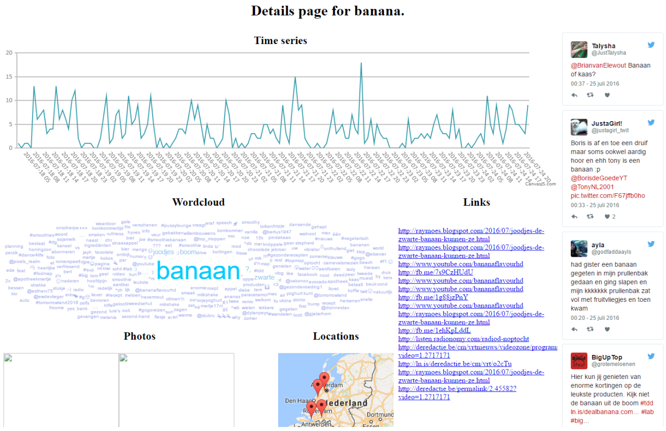


FIGURE 3. EXAMPLE OF THE TOP 10 DISCUSSED FRUITS AND VEGETABLES ON TWITTER.

the notion of stories, which clusters the clusters over time. By doing this, we can also track which clusters are prevalent for a longer period of time and therefore will be very likely to be of value for the industry. A current implementation of this method has been successfully implemented for the news reporting industry. This system can be found at [20].



http://raymoes.blogspot.com/2016/07/joodjes-de-zwarte-banaan-kunnen-ze.html
http://raymoes.blogspot.com/2016/07/joodjes-de-zwarte-banaan-kunnen-ze.html
http://www.youtube.com/bananaflavourhd
http://fb.me/7s9CzHUdU
http://www.youtube.com/bananaflavourhd
http://www.youtube.com/bananaflavourhd
http://fb.me/1g88jzPnY
http://www.youtube.com/bananaflavourhd
http://raymoes.blogspot.com/2016/07/joodjes-de-zwarte-banaan-kunnen-ze.html
http://fb.me/1ehKpLddL
http://listen.radionomy.com/radiod-noptocht
http://deredactie.be/cm/vrtnieuws/videozone/program video=1.2717171
http://ln.is/deredactie.be/cm/vrt/o2cTu
http://raymoes.blogspot.com/2016/07/joodjes-de-zwarte-banaan-kunnen-ze.html
http://deredactie.be/permalink/2.45582?video=1.2717171

Talysha
@JustTalysha
@BrianvanElewout Banaan of kaas?
00:37 - 25 juli 2016

JustaGirl!
@justagirl_twit
Boris is af en toe een druif maar soms ookwel aardig hoor en ehh tony is een banaan :p @BorisdeGoedeYT @TonyNL2001 pic.twitter.com/F67jffb0ho
00:33 - 25 juli 2016
♥ 2

ayla
@godfaddaayls
had gister een banaan gegeten in mijn prullenbak gedaan en ging slapen en mijn kkkkkkk prullenbak zat vol met fruitvliegjes en toen kwam
00:20 - 25 juli 2016

BigUpTop
@grotemeloenen
Hier kun jij genieten van enorme kortingen op de leukste producten. Kijk niet de banaan uit de boom #tdd ln.is/dealbanana.com… #lab #big…

FIGURE 4. EXAMPLE OF DETAILS PAGE FOR TWEETS MENTIONING BANANAS.

V. DISCUSSION AND FUTURE WORK

In this paper, we describe our first steps towards empowering the horticulture industry by analyzing topic-relevant tweets in Twitter. During our first exploration of the Twitter data, we encountered some interesting results. For instance, we found that there could be predictive power in the number of times a specific product is mentioned on Twitter for the future sales numbers of that particular product. Furthermore, we developed methods to visualize the current industry specific content in real-time and filter out interesting information in the process.

These ideas can be fruitfully adopted in marketing analytics to directly measure the impact of marketing activities. Furhermore more, Section IV-A yields promising result in the production planning. These first results provide a good basis for further study.

ACKNOWLEDGMENTS

We thank RTreporter for supplying access to their tweet collections for our analyses.

REFERENCES

[1] D. Metric. Social media bechmark report. Retrieved: August 29, 2016. [Online]. Available: http://www.netbase.com/blog/why-use-social-analytics (2013)

[2] X. Wang, M. Gerber, and D. Brown, “Automatic Crime Prediction using Events Extracted from Twitter Posts,” in Social Computing, Behavioral-Cultural Modeling and Prediction. Springer, 2012, vol. 7227, pp. 231–238.

[3] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes Twitter users: real-time event detection by social sensors,” in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 851–860.

[4] E. Aramaki, S. Maskawa, and M. Morita, “Twitter catches the flu: detecting influenza epidemics using Twitter,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011, pp. 1568–1576.

[5] A. Signorini, A. M. Segre, and P. M. Polgreen, “The use of Twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic,” PLoS One, vol. 6, no. 5, 2011, p. e19467.

[6] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a social network or a news media?” in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 591–600.

[7] D. Bhattacharya and S. Ram, “Sharing news articles using 140 characters: A diffusion analysis on Twitter,” in Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on. IEEE, 2012, pp. 966–971.

[8] C. Castillo, M. El-Haddad, J. Pfeffer, and M. Stempeck, “Characterizing the life cycle of online news stories using social media reactions,” in Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing, ser. CSCW ’14. New York, NY, USA: ACM, 2014, pp. 211–223. [Online]. Available: http://doi.acm.org/10.1145/2531602.2531623

[9] J. Lehmann, C. Castillo, M. Lalmas, and E. Zuckerman, “Transient news crowds in social media,” in Proceedings of the Conference on Weblogs and Social Media, ser. ICWSM. AAAI, 2013, pp. 351–360.

[10] J. Lanagan and A. F. Smeaton, “Using Twitter to Detect and Tag Important Events in Live Sports,” Artificial Intelligence, vol. 29, no. 2, 2011, pp. 542–545. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2821/3236

[11] G. Van Oorschot, M. Van Erp, and C. Dijkshoorn, “Automatic extraction of soccer game events from Twitter,” in CEUR Workshop Proceedings, vol. 902, 2012, pp. 21–30. [Online]. Available: http://ceur-ws.org/Vol-902/paper_3.pdf

[12] J. Nichols, J. Mahmud, and C. Drews, “Summarizing sporting events using twitter,” in IUI ’12: Proceedings of the 2012 ACM international conference on Intelligent User Interfaces. ACM, 2012, pp. 189–198. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2166966.2166999$\backslash$npapers3://publication/doi/10.1145/2166966.2166999

[13] C. Xu, Y. Yu, and C.-K. Hoi, “Hidden in-game intelligence in nba players’ tweets,” Commun. ACM, vol. 58, no. 11, Oct. 2015, pp. 80–89. [Online]. Available: http://doi.acm.org/10.1145/2735625

[14] J. ”Järvinen, A. Töllmen, and H. Karjaluoto, ”Marketing Dynamism & Sustainability: Things Change, Things Stay the Same…”. Springer International Publishing, 2015, ch. ”Web Analytics and Social Media Monitoring in Industrial Marketing: Tools for Improving Marketing

Communication Measurement", pp. 477–486. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10912-1_157

[15] W. He, H. Wu, G. Yan, V. Akula, and J. Shen, "A novel social media competitive analytics framework with sentiment benchmarks," Information & Management, vol. 52, no. 7, 2015, pp. 801–812, novel applications of social media analytics. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378720615000397

[16] S. Hays, S. J. Page, and D. Buhalis, "Social media as a destination marketing tool: its use by national tourism organisations," Current Issues in Tourism, vol. 16, no. 3, 2013, pp. 211–239. [Online]. Available: http://dx.doi.org/10.1080/13683500.2012.662215

[17] W. Fan and M. D. Gordon, "The power of social media analytics," Commun. ACM, vol. 57, no. 6, Jun. 2014, pp. 74–81. [Online]. Available: http://doi.acm.org/10.1145/2602574

[18] Retrieved: August 29, 2016. [Online]. Available: https://dev.twitter.com/streaming/reference/post/statuses/filter

[19] Retrieved: August 29, 2016. [Online]. Available: https://dev.twitter.com/rest/public/rate-limits

[20] Retrieved: August 29, 2016. [Online]. Available: http://www.rtreporter.com