HAL open science

Business Intelligence from User Generated Content: Online Opinion Formation in Purchasing Decisions in High-Tech Markets

Karan Setiya, Jolien Ubacht, Scott Cunningham, Sertaç Oruç

To cite this version:

Karan Setiya, Jolien Ubacht, Scott Cunningham, Sertaç Oruç. Business Intelligence from User Generated Content: Online Opinion Formation in Purchasing Decisions in High-Tech Markets. 15th Conference on e-Business, e-Services and e-Society (I3E), Sep 2016, Swansea, United Kingdom. pp.505-521, 10.1007/978-3-319-45234-0_45 . hal-01702166

HAL Id: hal-01702166

https://inria.hal.science/hal-01702166v1

Submitted on 6 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Business Intelligence from User Generated Content: online opinion formation in purchasing decisions in high-tech markets

Karan Setiya[1], ✉ Jolien Ubacht[2], Scott Cunningham[2], Sertaç Oruç[3]
[1]Deloitte Consulting, Almere, The Netherlands
ksetiya@deloitte.nl;
[2] Delft University of Technology, Faculty of Technology, Policy & Management, Delft, the Netherlands
{j.ubacht, s.cunningham}@tudelft.nl;
[3]Sybo Games, Copenhagen, Denmark
sertac@sertacoruc.com

Abstract

User Generated Content (UGC) requires new business intelligence methods to understand the influence of online opinion formation on customer purchasing decisions. We developed a conceptual model for deriving business intelligence from tweets, based on the Classical Model of Consensus Formation and the Theory of Planned Behaviour. We applied the model to the dynamic high-tech smartphone market by means of three case studies on the launch of new smartphones. By using Poisson regression, data- and sentiment-analysis on tweets we show how opinion leadership and real-life events effect the volume of online chatter and sentiments about the launch of new smartphones. Application of the model reveals businesses parameters that can be influenced to enhance competitiveness in dynamic high tech markets. Our conceptual model is suitable to be turned into a predictive model that takes the richness of tweets in online opinion formation into account.

Keywords: business intelligence · User Generated Content · Twitter · sentiment analysis · opinion formation · smartphones · high-tech markets

1. Introduction

Nowadays having an online presence is the norm for enterprises. At the same time, the amount of data produced on the Web is increasing rapidly. Consequently, the ability of handling this data is becoming increasingly important for enterprises in order to gain competitive advantage [6] and retain market position. In our study, we specifically focus on how this phenomenon of the increased importance of User Generated Content (UGC) fits into high-tech industries. High-tech industries are characterized by high sunk costs, high risks and being dynamic with rapidly changing customer requirements and product characteristics/features [3]. Obtaining and sustaining competitive advantage requires different methods in such fast-paced industries compared to other (low-tech) industries where the operating environment is less prone to rapid change. This leads us to raise the question of whether the analysis of UGC can help generate better insights into a fast moving consumer market with short production cycles. In the smartphone industry, we recognize the aforementioned trends of increased data being generated from (mobile) devices; there is an increasing amount of mobile UGC and UGC is increasingly influencing public image and sales of products [16]. The smartphone industry itself is characterized by short product life cycles, rapid product and feature imitation, aggressive pricing strategies, highly price sensitive end-consumers and high barriers to entry [4]. In our study, we developed a conceptual model for deriving business intelligence from microblog posts. The objective of applying this model is to obtain new insights into the way that prevailing opinions are formed within UGC by analysing tweets. We applied the conceptual model on three real life cases of new smartphones. The main research question in this study is "*What kind of business intelligence on opinion formation of end users can be derived from tweets?*".

Business intelligence applications of UGC is a relatively new field of academic study, a literature review yielded a limited number of studies where UGC is used to relate to real world phenomena such as predicting movie box office returns based on tweets [2] or presidential election polls [15]. In both studies the UGC analysis produced more accurate predictions than traditional methods. While many studies have recognized the potential of UGC analysis for business intelligence, only very few provide a methodology and apply this methodology to real cases, whereas the importance of understanding the potential of UGC in business-settings is increasing. Therefore, by

means of developing a theoretical model on the role of opinion formation, we shed light on the types of business insights that can be derived from User Generated Content.

In section 2 we present the conceptual model that we developed for measuring online opinion formation, based on the *Classical Model of Consensus Formation* and the *Theory of Planned Behaviour.* Next, in section 3, we present the empirical data and the statistical analysis that we performed in order to further understand the process of opinion formation, which in turn can be influenced by companies to their own (competitive) advantage. The outcomes of the empirical twitter analysis are presented in section 4, in which we present a descriptive statistical model to describe the data, our Poisson regression analysis which shows the effect of our main independent variables (influencers, real world events and spam bots) on the tweets volume and sentiment and we show several aspects of User Generated Content that can be input for company (marketing) strategies. A reflection on our conceptual model is presented in section 5. Finally, in section 6 we answer our research question, reflect on limitations of the conceptual model and the empirical data analysis and formulate future research topics.

2. Conceptual Model

In order to develop our conceptual model for measuring online opinion formation, we examined two of the most important established models in the domain of online opinion formation and the drivers of (purchasing) behaviour (and in turn, competitive position): the *Classical Model of Consensus Formation* [5] and the *Theory of Planned Behaviour* (TPB)[1].

2.1 Classical Model of Consensus Formation

We used the adaptation of the Classical Model of Consensus Formation by Jackson [10] as a basis to explore the online opinion formation in the cases analysed. The original model by De Groot [5] describes that in a group of agents, no agent will simply share or strictly disregard the opinion of any other agent. Each agent in the network will take others' opinions into account when forming their own opinion. Repeatedly averaging this process of influencing leads to the newly formed opinions of the agents to either be brought closer to each other or be taken further apart. This can either lead to convergence into agreement or to divergence into disagreement. Jackson introduced opinion leaders to this Classical Model of Consensus Formation [10]. In the exploration and critical assessment of the model by De Groot [5] he states: *"Opinion leaders will arise naturally in the model, as individuals who are listened to by others and who have non-negligible influence on the opinions of at least some other agents."* (p. 296). He further states that when a network is strongly connected, such that there is a direct path from any agent to any other agent, and aperiodic, then it is convergent. Literature often presumes that this is the case; there are no agents who are not influenced by another agent's opinion. Translated to our research context this entails that if every person in a social network is, to some extent, influenced by the opinion of every other person in the same network, then a prevailing opinion will arise (convergence). The original model was adapted in our study to fit our more open context where the agents are in fact not strongly connected but linked via open social media. We used this model to describe the formation of the prevailing opinion from the collected tweets. Note that in our open context, a prevailing opinion which is carried by a majority of people in a network is not guaranteed since not every person in our data set is necessarily connected to every other person via a direct path. Therefore, multiple prevailing opinions can arise. This is important to acknowledge because multiple prevailing opinions adds complexity to understanding the formation of the prevailing opinion(s) and possible efforts to influence these.

2.2 Theory of Planned Behaviour (Subjective Norms)

The Theory of Planned Behaviour (TPB) by Azjen [1] is originally a theory from the psychology field. This theory describes the relation between beliefs and action and was proposed to build upon the Theory of Reasoned Action [9]. The TPB describes that human behaviour, specifically the intent to act, can be predicted by three types of considerations. The first consideration is attitude towards a certain behaviour: the way a person feels about the behaviour, personally. The second consideration is subjective norms: whether or not the specific behaviour is encouraged by friends and family. The third predictor of intention is *perceived behavioural control*: the extent to which the person feels able to overcome barriers to the behaviour. According to the TPB, these three considerations form the most important predictors of intention, subsequently intention is the most important predictor of behaviour. In our study, we consider prevailing online opinion as a proxy for *subjective norms*, which is one of the predictors of human behaviour.

2.3 Conceptualization

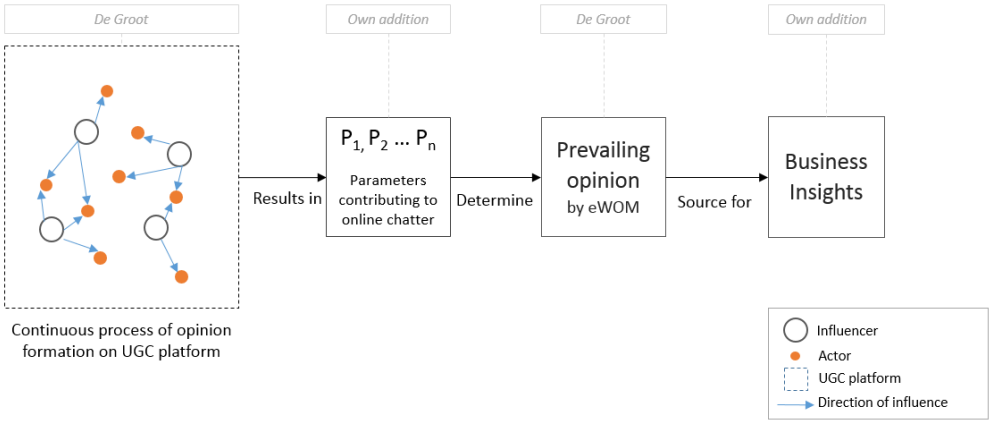

Fig. 1. Conceptual model for deriving competitive advantage from UGC analysis

In Fig. 1 we present our conceptual model based on elements from the Classical Model of Consensus Formation with our own additions. The model illustrates how business insights can be derived from online chatter that leads to opinion formation by electronic word of mouth (eWOM). We consider online chatter as a continuous process of opinion formation. In our context the opinions are formed about products, specifically smartphones. Next, from the online opinion formation process we identify a number of key parameters that are expected to be the main determinants of the prevailing online opinion (in De Groot's model this is the formed consensus). The key implications are that firstly this prevailing opinion, in a way, is already a business insight. When we are able to discover a certain opinion about a product emerge from online chatter, this is grounds for a company to formulate a strategy to respond to this online product sentiment in order to influence it for the better (or in order to damage competitors' online product opinions). Although not tested in our study, we also suggest that these parameters are valuable sources for predicting intention to buy according to the Theory of Planned Behaviour. In turn, this can potentially contribute to improve the competitive advantage of companies in the high-tech market, given that appropriate actions are taken to utilize these insights. Here, the prevailing online opinion would be a proxy for subjective norms. In other words, the online opinion about a product is highly valued when making a purchasing decision or even in creating the intention to buy. This intention to buy, in turn, is a predictor for actual buying behaviour [1]. If future studies provide evidence that there is indeed a relationship between the online opinion formation via eWOM and purchasing behaviour, businesses could make use of this in their strategy to improve their competitive position. In our study, we focused on the continuous process of online opinion formation. From this, we recognised a number of key parameters that shape the prevailing opinion that yield business insights in the role of opinion formation in high-tech markets.

3. Data collection and analysis method

In this section we explain the methodology and technical architecture to make the operationalization and the practical application of our conceptual model explicit. First, we present our key design choices for the empirical data (analysis) such as the selection of tweets as a proxy for UGC and the rationale behind the selection of the specific cases. Next, we explain the high-level technical architecture which was used for data-collection and, in part, for data analysis. Finally, we explain how we interpreted the data by means of our conceptual model.

3.1 Data collection

In order to test our conceptual model, we made a number of important design choices. First, we chose tweets as our UGC source. We consider tweets to be representative for a wider range of UGC (however not all UGC), similar to, amongst numerous other studies, [2], [11], [15]. Public tweets are relatively simple to collect through the Twitter APIs. Over 90% of tweets are public, and can thus be captured programmatically, without the need to obtain special permissions to collect this (public) data. Also, it is important for businesses to be able to respond quickly to rapidly

changing customer requirements in a dynamic market environment. Tweets lend themselves for such an objective as tweets can be collected almost in real-time, unlike most other forms of UGC. We collected tweets pertaining to the launch of three new smartphones in the market. We then used volume and sentiment of the tweets as metrics to find out the prevailing opinion about each product. According to Liu et al. [14], other studies of online word of mouth (eWOM), and research on traditional word of mouth (WOM), have identified volume (measured as the number of messages) and valence (classification of positive or negative) as two important measures of WOM activity.

The three smartphones selected for our case study are: Samsung Galaxy S5, Sony Xperia Z2 and HTC One (M8). Our case selection was based on both pragmatic considerations and an attempt to maximize the reliability versus construct validity trade-off. The three key considerations for our case selection were: (1) *the timing of the product launch*: we expected most online chatter about the products around the launch date, due to anticipation and buildup of online and offline hype (due to e.g. marketing efforts from manufacturers and brand loyalty from customers); (2) *similar product features*: each smartphone caters to a similar target market. With similar target markets, we can compare the opinion formation for the three cases and recognize specific capabilities that the companies show (such as employing a certain parameter to influence the prevailing opinion); (3) *Expected volume of online chatter*: for our statistical analyses, we needed to maximize the collected tweets in order to have a representative sample of the population of tweets. For collecting the tweets, we developed a web application which automatically queries the Twitter API with relevant key words for each case, processes the response by parsing the returned data and saving it in a database.

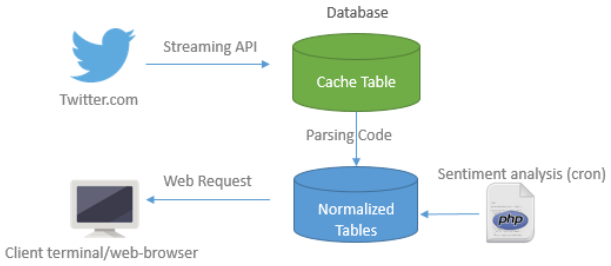

Fig. 2. High level technical architecture

Fig. 2 illustrates the high level technical architecture. Our web application queries Twitter's streaming API for tweets about our search terms such as "Samsung Galaxy S5". The response is dumped into our cache table of the database without parsing the data. Next, the raw data is parsed and saved to the respective tables in a MySQL database. This process of translating the raw data into our own data structure is known as *database normalization.* Our normalized data in the populated database is now ready for analysis. In our analysis, the key parameters that shape the prevailing opinion of each product are those factors that influence the volume and sentiment of our collected tweets. Before finding these parameters, we first needed to do a sentiment classification of our data set. To this end, we implemented a PHP class which classified our collected tweets into positive, neutral and negative sentiments based on natural language processing techniques. This sentiment classifier would run automatically each day to tag each tweet with its corresponding sentiment class. Once our populated, normalized database was classified, we used this dataset to analyse and interpret the data. Sentiment and volume of tweets are our dependent variables, and the key parameters that shape these variables become the independent variables.

3.2 Data analysis

In total, we collected 1,448,799 tweets, which is a 369 days worth of tweets (with three simultaneous data streams for approx. 120 days). The analysis of the tweets was subdivided into two major parts. First, we built a descriptive statistical model, in order to help explain the data trajectory and its distribution. An important trade-off in building this model is: the more parameters we add, the more data-specific the model becomes, but this also sacrifices the level of generalizability of the model. With this trade-off in mind, based on the theoretical underpinning described in section 2, and a first inductive exploration of the data, we considered the following parameters to be the most important contributors to the expected rate of tweets per day (volume):

$$\boldsymbol{\lambda} = N + E(t) + C + O + B$$

Where:

- **λ :** Poisson expected rate of daily tweets
- **N**: Base rate of daily tweets. This is the number of "noise" tweets that is more or less constant throughout the data collection. This number of tweets was picked up for the given query, regardless of any other (external) factors
- **E(t)**: An event that triggers tweets as a function of time. For example a large-scale mobile conference, a feature announcement or the launch of the smartphone in question. From our data, it becomes apparent that the tweets triggered from such a real world event are a function of time because the tweet rate often shows anticipatory and spillover increase in tweet rate surrounding the event
- **C**: The manufacturing company-specific capabilities. Brand equity and the ability of a manufacturer to directly or indirectly effect the online and offline chatter about their product is the third building block of the expected daily rate of tweets. Note that measuring this brand equity and quantifying its contribution to λ is out of scope of our study. Brand equity is a composite variable, built up from among others: performance, value and social image [12]
- **O**: Presence of opinion leaders or key influencers. We use two metrics to identify the key influencers in our data set: the number of mentions per user account in our collected tweets, and the number of retweets for each tweet. This gives us the most popular user accounts and the most popular tweets and with these two metrics, we were able to indicate the most influential user accounts and show the most popular tweets in each case
- **B**: The presence of spam bots. Spam bots can dilute the opinion formation and the classification of sentiment. If a negatively classified tweet is repeated many times by a spam bot, this could unjustly skew the prevailing opinion toward negative opinions. Therefore, we attempted to identify such tweets by calculating the likelihood of every user to tweet more than the overall average number of tweets per day (Poisson likelihood distribution). If a user's tweeting frequency has a likelihood of less than 1%, we consider this a spam bot (account)

The second major part of the statistical analysis consists of a number of Poisson regression analyses and modeling the interaction effects of the main independent variables to answer the main research question. The Poisson regressions model the volume of tweets based on the three parameters *tweets from influencers, occurrence of events* and *tweets from spam bots* (these are the independent variables). We also modeled the effect of these parameters on the sentiment of tweets and volume of tweets (the dependent variables). By manipulating the aforementioned parameters, we derived company-specific capabilities which were identified in terms of inducing online chatter.

4. Statistical Analyses and interpretation

A summary of the main findings from our descriptive statistical analysis is presented in table 1. The base rate of tweets (N) is clearly different in each case, from 154 tweets per day for the Sony Xperia Z2, to 1729 tweets per day for the HTC One (M8). Further, in all three cases, the kinds of opinion leaders were similar. Each case showed that the highest number of mentions and retweets came from three categories of opinion leaders: the smartphone manufacturer, a number of mobile network operators (such as T-Mobile, Verizon etc.) and a number of well-known technology bloggers/vloggers.

The HTC case yielded the most tweets per day, however the Samsung case yielded similar numbers in terms of volume of tweets per day and the fraction of (spam) bots. In both cases, the fraction of spam was well over half (53.5% and 65.4% for Samsung and HTC, respectively). The Sony case yielded far less tweets, however also relatively less spam with 46.4% coming from spam accounts. This indicates that products from smartphone manufacturers with higher market share and online presence will also induce more spam.

The case with the most positive online chatter is the Sony Xperia Z2, which is surprising since this is the least popular in terms of volume of tweets. Not only does the Sony case produce the highest percentage of positive tweets, it also has the least negative tweets. This indicates that Sony is able to create a more positive buzz online than larger players on the smartphone market such as HTC and Samsung.

Table 1 Overview of cases essential statistics

	Total # tweets	**Λ**	**N**	**# Event days**	**O**	**B**	**% Positive**	**% Negative**	**% Neutral**
Samsung Galaxy S5	679,704	5,189	268	48	4,634	357,400	32.72	8.29	58.99
HTC One (M8)	677,383	6,048	1,729	31	2,896	424,933	29.17	7.45	63.39
Sony Xperia Z2	91,712	728	154	42	830	41,397	34.48	4.08	61.43

The second part of our statistical analysis is the Poisson regression analysis, where we tested a number of hypotheses in order to model the effect of the identified parameters on the daily rate of tweets and sentiment.

Table 2 Poisson regression: daily tweet rate

	Prediction (# tweets)	**Influencers**	**Event**	**Bots**
Daily tweet rate	809.97 ***	0.993 ***	2.028 ***	1.0002 ***
Daily positive tweet rate	243.47 ***	0.994 ***	2.131 ***	1.0002 ***
Daily negative tweet rate	68.92 ***	0.995 ***	1.85 ***	1.0002 ***
Daily neutral tweet rate	512.35 ***	0.992 ***	1.999 ***	1.0002 ***

*Signif. codes: 0: "***" 0.001: "**" 0.01: "*" 0.05: "." 0.1: " " not significant: "n.s." (as given by R)*

In Table 2, we present the results of the Poisson regression for the overall tweet rate and tweet rates per sentiment class. The model predicts nearly 810 tweets per day *without* the effect of influencers, bots or events (when independent variables have no effect). The values of the independent variables (Poisson parameters) are a factor increase or decrease on the predicted rate of daily (positive, negative or neutral) tweets if that independent variable increases by 1 unit, *all else being unchanged.* From the table it becomes evident that the occurrence of a real-world event triggers the strongest effect on overall daily tweet volumes (factor 2.028) as well as for each of the sentiment classes. This is in line with expectations; real world events such as the Mobile World Conference or the announcement of a specific functionality of the new smartphone are, to a great extent, marketing and promotional methods for manufacturers, meant to create awareness and buzz around their products. These events are the strongest determinant of daily tweet rate.

In addition, influencers and spam bots also significantly affect the daily tweet rate. Here, it is noticeable that influencers have a dampening effect on the daily tweets volume from 'real users' (non-influencers and non-spam

bots). This could indicate that an increase of tweets from influencers shift 'real users' into a more passive role of publishing tweets about the product.

Furthermore, in terms of sentiment, we see that the occurrence of events are a stronger determinant of an increase in positive tweets (increase by 213%), than on negative tweets (increase of 85%). Events also trigger a stronger increase in positive tweets than the overall increase of 103%. The data analysis shows that events trigger a relatively strong positive sentiment as compared to negative and neutral sentiment online chatter (tweets)
Diving deeper into our statistical analysis, we modeled the autonomous effect of our independent variables on the daily tweet volume and sentiment per case. The results of this analysis can be found in Table 3.
Consistent with the base rate of tweets (N), the Poisson regression predicts the most tweets for the HTC case and the least tweets in the Sony case, when none of the independent variables (influencers, events, bots) are at play. Furthermore, spam bots and events are the strongest determinants of tweet volume and sentiment in the Sony case and influencers have the strongest effect in the Samsung case. Surprisingly, influencers have a slightly decreasing effect on the real users' tweets in the HTC case. This is not the case in the Samsung and Sony case. However, since most of these results are not more than 0.1 significant in the HTC and Sony cases, further analysis would be needed in order to determine the validity of this specific effect.

Table 3 Poisson regression: total and sentiment tweet rate per case

		Prediction	**Influencers**	**Event**	**Bots**	**Pseudo R^2**
Samsung Galaxy S5	*All*	2,411.49 ***	1.002106 ***	1.774112 ***	1.00012 ***	0.644
	Positive	206.43 ***	1.00491 ***	1.776597 ***	1.000118 ***	
	Negative	87.97 ***	1.006669 ***	1.839144 ***	1.000096 ***	
	Neutral	714.0836 ***	1.002693.	1.693674 ***	1.00012 ***	
HTC One (M8)	*All*	3,238.935 ***	0.9964912 **	1.773402 ***	1.000117 ***	0.723
	Positive	324.4074***	0.9984911	1.951503***	1.000095***	
	Negative	96.93106***	0.996837	1.510892**	1.000115**	
	Neutral	655.239***	0.998553	2.443912***	1.000089.	
Sony Xperia Z2	*All*	358.2151 ***	1.000198	1.951005 ***	1.001031 ***	0.642
	Positive	63.68824***	1.019967	3.553738***	1.000089	
	Negative	9.18002***	1.007648	1.125418	1.001048.	
	Neutral	135.9369***	0.9905421	2.245549***	1.000704*	

4.1 Company specific capabilities

By comparing the case-specific regression results, we identified company specific-capabilities which can help the smartphone companies recognize their strengths, weaknesses, opportunities and threats in the process of online opinion formation. If we assume that each company's objective is to increase the number of tweets from real users about their product, then according to the results of our analysis, Sony has been most effective in using real world events to trigger additional tweets from real users. At the same time, spam bots (which may also have been employed by Sony), have the strongest effect out of all three cases.

Samsung was most effective in using influencers to drive chatter from real users. Specifically, when looking at the most retweeted tweet in the Samsung data, we found that a tweet by football star Cristiano Ronaldo, promoting the Samsung Galaxy S5 as part of Samsung's marketing campaign, was most influential.

In terms of the effect on sentiment of real users, a few things stand out from our results. Firstly, the strongest increasing effect on positive tweets by influencers is found in the Samsung case, at the same time, they also have the strongest increasing effect on negative tweets. Comparing the effects of the parameters on sentiment of real users, we found that in the Sony case, the effect of real world events have the strongest increasing effect on positive tweets. Another notable result is that the effect of events on negative tweets is stronger than the effect on positive tweets in the Samsung case. Although the absolute number of predicted negative tweets is arguably negligible, this

is an important finding which Samsung could use to strategically engage with their (potential) customers in order to improve the prevailing opinion(s) about their product online and offline. Each company in our analysis has their own specific capabilities and with our data-driven analysis, we can shed light on these capabilities, allowing these companies to act upon them.

4.2 Interaction effects

In the previous section, we modeled the isolated effect of the independent variables influencers, events and spam bots on the daily tweeting volume and sentiments. In order to understand the dynamics and interaction between the independent variables and the effect of those interactions, we also modeled the effects of a combination of independent variables on the dependent variable: the real user tweet rate. Since we are dealing with an open multi-actor and multi-variable environment where numerous factors can have an effect on both sentiment and volume of tweets, it is important to consider these interaction effects.

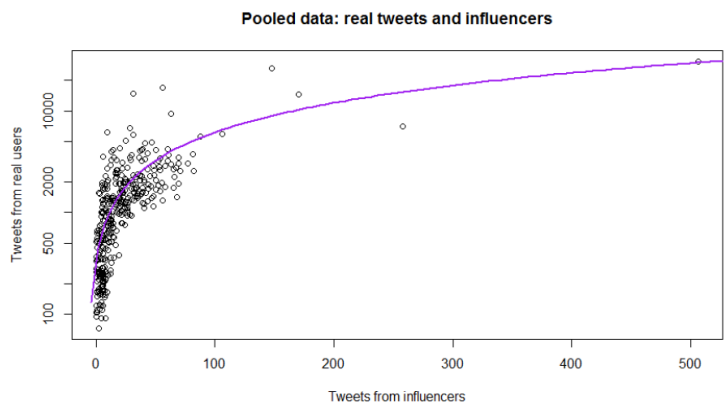

Fig. 3. Pooled tweets from influencers and daily tweets from real users

In Fig. 3 we can see that, for the pooled tweets of all cases, there is a positive trend line for the daily tweets volume of real users as a function of the tweets coming from influencers. This is different from what we found in our isolated effect of influencers in the Poisson regression in the previous section. This difference can be explained by the fact that in Fig. 3, the effect of the other parameters (bots and events) are not unchanged. Therefore, it seems that there is an interaction between the events, spam bots and influencers in their effect on the daily tweet volume from real users. Again, we use Poisson regression to model this interaction. An interaction between independent variables A and B implies that the effect of A depends on the value of B and that the effect of B depends on the value of A [7]. In Table 4 the results of the modeled interaction effects are presented. The interaction for events and influencers has the strongest effect on the predicted daily rates of tweets from real users. Also, the three-way interaction of events, influencers and bots results in a slight increase of predicted daily tweets. This tells us that when influencers tweet and there is also an event, the tweets from real users (non-influencers and spam bots) decrease. This same effect is apparent for the modeled two-way interactions of events, influencers and spam bots.

Table 4 Pooled data: Interaction coefficients

Interaction	Coefficient
Events and Influencers	0.9590 ***
Influencers and Bots	0.9999 ***
Bots and Events	0.9995 ***
Events and Influencers and Bots	1.00002 ***

However, when the interaction of all three parameters are modeled, the predicted number of tweets from real users increases slightly. This slightly unintuitive effect of events and influencers' interaction leading to a decrease in tweets from real users might be explained by the nature of the interactions between influencers and real users. Similar to traditional media (TV, radio etc.), when an event occurs, the media is the first to publicize this to a passive audience. Similarly, key influencers such as smartphone manufacturers, technology bloggers and mobile network operators are expected to tweet more on days of events in order to report on these events and real users are then adopting a more passive role as they are consuming the newly published information before expressing their own opinions.

4.3 Synthesis of findings

When we consider both the isolated effects and interaction effects on volume and sentiment of tweets from real users, it seems that in our (pooled) data, interaction effects do occur (as seen in Fig. 3) as the coefficients of the effect in our individual parameters show a decrease in tweets volume when influencers increase their tweets volume. Whereas the interaction effect of all three parameters shows an increasing coefficient (1.00002) as can be seen in Table 4. Our analyses thus show that when there is an interaction of all three (core) parameters then the tweets from real users are predicted to increase. In short, in the HTC case, influencers have the strongest decreasing effect on real users' tweets. Spam bots have the strongest increasing effect on real users tweet volume in the Sony case and events have the strongest effect in the Sony case. Finally, the effect of events on real users' tweets volume are very similar in the Samsung and HTC case (coefficients of 1.774 and 1.773 respectively).

5. Conceptual model applied

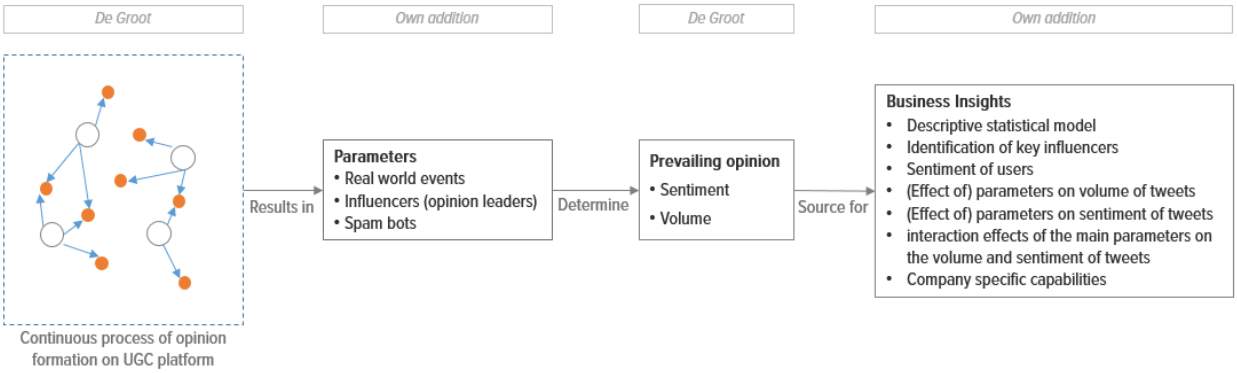

Fig. 4. Application of conceptual model

In Fig. 4 we present the application of our conceptual model, as illustrated in Fig. 1, by starting with analyzing the process of opinion formation. To this end, we took a snapshot of this continuous process of opinion formation with our dataset of 369 days worth of tweets. We then recognize a number of opinion leaders (or influencers). Next, we classified the tweets in positive, negative or neutral tweets. With this classification we were able to model the effect of our identified opinion leaders and events on the sentiment as presented in the previous sections. By aggregating the tweets with these classifications, we could indicate the prevailing opinions surrounding the respective topics (smartphones). This prevailing opinion is seen as the subjective norms as described in the TPB[1]. We suggest that these subjective norms can be strategically shaped by incentivizing actors in the network and in this way steering the (sentiment of) online chatter about a product. By analyzing tweets, we recognized three key parameters that shape the prevailing opinion: the first two are real-world events and a small number of opinion leaders who determine the prevailing opinion amongst these agents. The third parameter is spam-bots. It is important to recognize that spam-bots can potentially influence this opinion as well since popular products and brands tend to elicit more spam-messages which misuse the brand recognition in order to attract people to their websites. The main 'real' influencers however, are the handful of opinion leaders and real-world events. In turn, the resulting prevailing online opinion, has a number of important implications in terms of a product's potential performance. Some of the most important business insights derived from our applied model are the identification of key influencers and events, and their effect

on the tweets volume and sentiment surrounding a specific topic. With such insights, companies can actively steer the online (and consequently offline) conversation about their products by strategically engaging with (potential) customers, influencers and organizing events. Our conceptual model can be further developed into an 'always-on' solution where, based on the parameters, simulations can be run which provide predictions into the consequences of managerial decisions. For example, such a decision support system could simulate real world events and its repercussions on sentiment based on historical data to support management decisions with regard to e.g. marketing strategy or product release timing, etc.

6. Conclusion

In this section we present our overall conclusions while answering our main research question "*What kind of business intelligence on opinion formation of end users can be derived from tweets?*". The purchasing behavior of customers is increasingly being influenced by the online chatter about products and services. Therefore, methods to steer this online opinion formation are becoming increasingly valuable for companies. Understanding the online opinion formation and strategically acting upon the insights gained from the analysis of user generated content underlie the ability to steer this prevailing opinion about specific topics. With the application of our conceptual model, we presented a number of methods to find such valuable insights in tweets.

Firstly, in order to understand the formation of prevailing opinions online, we can identify opinion leaders. These key influencers contribute to the subjective norms surrounding a product, online. *Top-down (proactive):* Proactively understanding the drivers of online opinion formation will enable companies to establish strategies to influence the prevailing opinions that are bound to arise. More concretely, one of the main drivers of prevailing (online) opinions are opinion leaders. By measuring the influence of opinion leaders upon the sentiment and volume of tweets from (potential) customers, companies can employ such opinion leaders in order to drive the online opinion. However, the objectivity of influencers such as technology vloggers is imperative to their credibility and, by extension, their core business. These influencers will less likely promote a product in exchange for compensation from the manufacturer as the risk of being exposed and losing credibility as an objective technology reviewer would be detrimental. *Bottom-up (reactive):* At the same time, with an online and social media presence, companies can get much closer to their customers than ever before. By actively measuring the online sentiment surrounding a product or service, companies can leverage online channels to engage with their customers on a personal level. The offline world and the online realm cannot be considered independently of each other. We have seen in our twitter analyses, that real world events strongly influence the online chatter and sentiment about smartphones. Finally, we are dealing with a multi-variable and open context where influencing factors upon tweets volume and online sentiment do not operate in isolation. Therefore, it is important to understand the interaction effects of the influencing factors. Here, the objective should be to find the right combination of influencing factors (events, opinion leaders and (spam) bots) and effectively formulating a strategy to employ these factors to influence the online opinions about a product or service. Our conceptual model, in combination with the concrete application of this model in our study, can serve as a basic framework for operationalizing such a strategy.

The main business insights that we were able to derive from our method of UGC analysis are:

- **Description of the buildup and trajectory of the data with a descriptive statistical model**: $\lambda = N + E(t) + C + O + B$. Each of these parameters can be influenced to shape online opinion formation.
- **Identification of key influencers (opinion leaders)**: Smartphone manufacturers, technology bloggers/vloggers, MNOs. Each of these influencers can be employed to influence the prevailing online opinion.
- **Sentiment of real users (over time)**: The trajectory of sentiment about the respective products can be analysed so that triggers for negativity and positivity can be identified and acted upon.
- **Main parameters that determine volume of tweets**: Events, influencers and (spam) bots. These main parameters will have a relatively strong impact on the prevailing opinion.
- **The effect of the main parameters on volume of tweets**: Events trigger the strongest uplift in tweets volume. Therefore, events are an important tool for manufacturers to consider when attempting to steer the online opinion formation.
- **The effect of the main parameters on sentiment of tweets**: Similar to the volume of tweets, events trigger the strongest uplift in positive tweets.

- **The interaction effects of the main parameters on the volume and sentiment of tweets**: When influencers tweet and there is also an event, the tweets from real users decrease.
- **Company specific capabilities**: Sony has been most effective in using real world events to trigger additional tweets from real users. Samsung was most effective in using influencers to drive chatter from real users. In the Sony case, the effect of real world events have the strongest increasing effect on positive tweets. Each company has specific capabilities which, we can recognize with our data-driven analysis, allowing these companies to act upon them in order to shape the online opinion regarding their product.

Our main academic contributions lie firstly in our way of analyzing UGC for business insights based on a theoretical model. Second, by using a modern application of the Classical Model of Consensus Formation we found concrete – new- parameters i.e. real world events and spam bots (in addition to opinion leaders) that shape online opinion formation. We discovered the roles of the diverse sources of tweets and their influence on volume and sentiment of tweets. This differs from the offline world as there are more sources and types of agents forming the prevailing opinion. The application of this model in the online world enabled these nuances as we could work with large volumes of data, as opposed to an offline setting with much more limited and more indirect data. Third, we found categories of opinion leaders who each showed significant effect on the formation of online opinion, namely: manufacturers, MNOs and technology bloggers/vloggers.

6.1 Limitations

In order to increase the reliability such that the results of our study become generalizable for a wider range of products or industries, the application of our conceptual model should be done on more cases, while maximizing representativeness of the respective product-group or industry. Nonetheless, we have made a first step toward delivering more insight in this field of study, where academic research is sparse. Our technical implementation is modular and scalable such that it can be applied to other products and topics where online chatter is expected. Furthermore, we collected data from one source: Twitter. Although tweets have proven to be a relevant, effective and rich source of data for our purposes, gathering data from various social media sources to cover a wider demographic and build a richer data set could improve the application of our conceptual model.

A number of improvements could be made on the text analysis techniques. Natural language processing is a complex and strongly-evolving discipline. Basic techniques were combined in this study in order to classify the tweets into sentiments and finding spam tweets. Some ways to improve the classification accuracy are: (1) expanding the matching dictionary used to classify each tweet's sentiment, (2) expanding the classifications to more nuanced sentiment based on severity of language and (3) considering word order, grammar and punctuation. In general, similar to sentiment detection, spam detection is considered a text classification problem [8]. By using e.g. a Naïve Bayes approach, similar to our sentiment classifier, spam classification in the application of our model can be improved isn future research.

6.2 Future research

In our research, we revealed insights which answer important questions in the domain of business intelligence from UGC as well as identify knowledge and theory gaps. First, we have laid the ground work for obtaining business insights from tweets and in future research, we suggest to measure the competitive advantage that can be derived from these insights. Moreover, explicit measurement of the relation between online opinion (as subjective norms) and buying behaviour (expressed as business performance metrics such as sales or profit) can be measured as part of this competitive advantage from UGC analysis. Second, our descriptive statistical model identifies the five main parameters that contribute to the expected tweets volume per day. This descriptive model can be combined with our Poisson regression analysis and be further developed into a predictive model. An implementation of our conceptual model can automatically draw from historical data and compute predictions based on influencing coefficients similar to our analysis (influence of events on positive tweets from users etc.). In closing, our study has laid ground work for understanding online opinion formation, UGC analysis and the implications on competitive advantage. Understanding the interactions between the online and offline world, in a business context and also in social and cultural contexts is paramount in today's rapid technological advancements, ubiquity of technology and online media in our lives.

References

1. Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes, 50*(2), 179-211.
2. Asur, S., & Huberman, B. A. (2010). *Predicting the future with social media.* Paper presented at the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT).
3. Baer, W. J., & Balto, D. A. (1998). Antitrust Enfocement and High-Technology Markets. *Mich. Telecomm. & Tech. L. Rev., 5*, 73.
4. Cromar, Scott (2010). Smartphones in the US: Market analysis. Last accessed April 18th 2016: https://www.ideals.illinois.edu/bitstream/handle/2142/18484/Cromar,%20Scott%20-%20U.S.%20Smartphone%20Market%20Report.pdf
5. De Groot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association, 69* (345), 118-121.
6. El Sawy, O., & Pereira, F. (2013). Anticipating Game Changers for "Enterprise 2020" in a Digitally-Intensive World *Business Modelling in the Dynamic Digital Space* (pp. 1-12): Springer Berlin Heidelberg.
7. Evans, S. M., Janson, A. M., & Nyengaard, J. R. (2004). *Quantitative methods in neuroscience: A neuroanatomical approach*: Oxford University Press, USA.
8. Fawcett, T. (2003). In vivo spam filtering: a challenge problem for KDD. *ACM SIGKDD Explorations Newsletter, 5*(2), 140-148.
9. Fishbein, M. (1979). A theory of reasoned action: some applications and implications. Nebraska Symposium on Motivation, Vol 27, 1979, 65-116.
10. Jackson, M. O. (2010). *Social and economic networks*: Princeton University Press.
11. Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology, 60*(11), 2169-2188.
12. Keller, K. L. (1993). Conceptualizing, measuring, and managing customer-based brand equity. *the Journal of Marketing*, Vol. 57, No. 1 (Jan., 1993), pp. 1-22
13. Kim, S.-M., & Hovy, E. (2004). *Determining the sentiment of opinions.* In: Proceedings of the 20th international conference on Computational Linguistics.
14. Liu, Y., Chen, Y., Lusch, R. F., Chen, H., Zimbra, D., & Zeng, S. (2010). User-Generated content on social media: Predicting market success with online word-of-mouth. *IEEE Intelligent Systems*, Volume 25, Number 1, Pages 75-78, January/February, 2010.
15. O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *ICWSM, 11*, 122-129.
16. Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science, 31*(2), 198-215.