Social Media Analytics with Big Customer Data. Some Marketing Decision Support Applications

Michel CALCIU Université Lille 1, RIME, France michel.calciu@iae.univ-lille1.fr	Jean-Louis MOULINS Université de la Méditerranée, CRETLOG, France jean-louis.moulins@univ-amu.fr	Francis SALERNO Université Lille 1, LEM, France francis.salerno@iae.univ-lille1.fr

Abstract

Cloud-computing provides easy and relatively inexpensive access to big data calculations. With clouds there is potentially no limit for the size of data and computing capacity. They can offer solutions to the « Data, data everywhere and not a byte to use » problem marketing managers are facing. Most of these data come from social media based marketing activities and need to be dealt with specific analytics. This paper presents several social media analytics models and applications and discusses their potential big data calculation complexities

Introduction

Social media is at the core of the so called “social commerce”, that is expected to generate tremendous business value in terms of operational efficiency (Wamba & al. 2016). Social media and related marketing activities generate vast amounts of data that need to be dealt with specific analytics. Social Media Analytics (SMA) has been identified by Fan and Gordon (2014) as an interdisciplinary modeling and analytical paradigm consisting of three steps: 1) capturing data from various sources; 2) understanding data using various analytics and models; and 3) summarizing and presenting the findings for decision making. SMA shares similarity with Big Data Analytics (BDA) in that both SMA and BDA can involve analysis, management and visualization of the similar types of datasets - accumulated traces of consumers’ online activities (Kiron, Perguson, & Prentice, 2013).

Big data in a narrow sense can be defined as data too large to be dealt with by one computer. In a larger sense it has been defined through the 3V model, which has been coined by Laney (2001) as : “high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making”. More recently, Gartner (Beyer & Laney, 2012) updated the definition of the 3V information assets as requiring new forms of processing to enable enhanced decision making, insight discovery and process optimization. The 3Vs have been extended in practice to 5V, adding data value and veracity as defining elements.

Social media data sources and big data

Multiple data sources contribute to the size, variability, and complexity of big data. Examples include weblogs, radio frequency identification, sensor networks, social networks, social data, Internet text and documents, Internet search indexing, call detail records, and large-scale e-commerce. New data collection practices like asyncron observation, crowdsourcing also become common.

A possible categorization of social media data is given in table 1

Table 1 Social media data categories

Demographic Data	Consumers' demographic information in social media, which includes age, gender, education, geography etc. (Kaplan & Haenlein, 2010). They are available in individuals' social media profile
Product Data	Product data in social media, which are generated when a particular brand or product name is mentioned (Mangold & Faulds, 2009). They can appear either in a brand's social media page or in personal area of social media user.
Psychographic Data	They indicate consumers' personality, values, attitude, interests and lifestyle related to a product or a brand. Consumers share their problems and expectation with a product, and product value and features on social media (Heinonen, 2011).
Behavioral data	Consumers' past buying behavior, such as buying record, in social media platform (Kietzmann, Hermkens, McCarthy, & Silvestre, 2011) can be used to predict future action.
Intention Data	Intentions expressed in social media help predict consumers' expectation with a product or a brand and future activities related to them (Ballings & Van den Poel, 2015).
Referrals Data	Ratings, reviews and nonverbal attitudes generate referral data. This type of data comes from positive or negative word of mouth on social media ((Trusov, Bucklin, & Pauwels, 2009).
Location Data	Real time location data of consumer. (Wagner et al., 2010).

Adapted from Wamba & al. 2016

Big social media customer data – Volume and Variety

Most of the social media customer data are potentially "big". Volume and variety, the first two V's from the above mentioned big data definition are best described by the datasets we use in our analysis. Velocity, the third V, will be discussed later in the paper where we present the big data computing techniques.

The first file we use is the Amazon *customer reviews dataset* (curtesy McAuley et al., 2015, file size 58,3G) that contains 82.68 million reviews after deduplication (142.8 million reviews originally) spanning May 1996 - July 2014.

We will refer to this dataset as the *web reviews dataset*. It will be used to predict customer ratings and helpfulness scores from verbatim feedback. We complete our text-mining exercise with a more extensive one from Liu et al. (2016), the only academic marketing research paper using big data cloud-analytics we have found in extant literature. It uses several unstructured web sources of data, many of them user-generated data (UGD), from various web platforms like Twitter[1], Google[2], Wikipedia[3] , IMDB[4] and Huffington Post[5] in order to produce a structured prediction model for TV show ratings. Three measures to extract information from the unstructured text data are applied in

[1] selecting relevant tweets demands the use of 4 identifiers (1) name of the show (e.g., Breaking Bad);14 (2) official Twitter account of the show (e.g., @TwoHalfMen_CBS); (3) a list of hashtags associated with the show (e.g., #AskGreys); and (4) the characters' names on the show (e.g., Sheldon Cooper)

[2] Google Trends provides total search volume for a particular search item. For the TV series data, one can use the name of the show (e.g., Two and a Half Men) and character names on the show (e.g., Walden Schmidt) as the keywords.

[3] Many of the Wikipedia editors are committed followers of TV and edit related articles earlier than the show's release date. Wikipedia edits or views may be good predictors of TV ratings.

[4] Consumers also post reviews on discussion forums such as the IMDB, chosen here because it has the highest Web traffic ranking (according to Alexa) among all TV-show-related sites.

[5] Consumers may also be driven to watch TV series by news articles. Huffington Post,is a site that offers news, blogs covering entertainment, politics etc. It ran 26th on Alexa as of January 29, 2015

order to produce three datasets that will be called in order *volume*, *sentiment* and *content* dataset. For Tweets they are described below.

The *volume* dataset records how many times a TV show is discussed. Users mentioning a show are likely to watch and their social network is likely to be influenced to watch.

In the *sentiment* dataset Tweets are classified by polarity, here positive and negative. Four percent of the Tweets that are used are labeled manually by experts and the rest are labeled automatically using the LingPipe[6] linguistic analysis package.

The *content* dataset uses a measure that make inferences from the full content of the Tweets using the frequency of all n-grams of tweets in all analyzed TV shows. An n-gram is a continuous sequence of n words in the text. For example the Tweet "I love Pittsburgh Steelers" contains four 1-gram, three 2-grams, two 3-grams, and one 4-gram. Collecting the Tweets 24 hours before the show produced 6,894,624 selected Tweets related to the 30 TV series with their 2339 episodes. The 2339 episodes are described by the frequency of the 28,044,202 n-grams resulting from those Tweets.

The last series of datasets comes from a paper by Culotta & Cutler (2016) develops a fully automated method for inferring attribute-specific brand perception ratings by mining the brand's social connections on Twitter. We refer to this paper as the "social mining" paper. It matches followers of exemplar accounts representing a perceptual attribute with the followers of brands. GreenPeace for example is an exemplar for the eco-friendliness attribute of a brand.

The *brand-followers datase*t used Twitter's API to collect up to 500,000 followers for each brand. It consists of 239 brand lines containing their names and the Ids of their followers (30.6M followers, 14.6M unique, 314M file size). The distribution of these brands by sector is: Apparel 70, Cars 37, Food and Beverages 70 and Personal Care 62.

The *exemplar-followers datasets* collect for each of the exemplar accounts the IDs of up to 50,000 of their Twitter followers. They consist of 74 eco-friendly exemplars (2.0M followers, 1.0M unique, 25M file size), 110 luxury exemplars (4.4M followers, 2.3M unique, 46M file size), and 405 nutrition exemplars (4.7M followers, 2.7M unique, 48M file size)

Some Social Media Analytics Applications

Variety in data goes comes with variety in analytics. Some of the most frequent SMA analytics types or categories are shown in table 2 and illustrated in the rest of this paper.

Table 2 Most frequent SMA categories

Topic Modeling	Detecting dominant themes or topics by sifting through large body of captured text
Opinion Mining or Sentiment analysis	Opinion mining is similar to Sentiment analysis, but it more focuses on the views, believes and judgment rather considering positive or negative sentiment at first place. Sentiment analysis refers to more in-depth interpretation of data of public/consumer/user sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, Individuals, issues, events, topics, and their attributes
Social Network Analysis	Analysis of the social network that made up of individuals call nodes and connected with other nodes with similar interest, knowledge, opinion, etc.

[6] http://alias-i.com/lingpipe/demos/tutorial/sentiment/read-me.html

Topic modeling

The first two reviews[7] from the reviews dataset presented above separated by square brackets in json (JavaScript Object Notation) format are given in Listing 1.

Listing 1- The First two records in the Amazon Reviews Dataset

```
{"reviewerID": "A2LDF3FTTTS1JF", "asin": "147523144X", "reviewerName": "old
lady","helpful": [1, 1], "reviewText": "first half of book well written but last
chapters rather rushed probably so that a follow up book could be published or
so it seemed.", "overall": 3.0, "summary": "good first half", "unixReviewTime":
1363046400, "reviewTime": "03 12, 2013"}
{"reviewerID": "AK7A5Y6X483XZ", "asin": "0060827882", "reviewerName": "Mary
Boyd", "helpful": [1, 1], "reviewText": "A friend recommended this novel when I
commented that I would like to know more about why the Scots moved to Ireland.
This is a book about division and adversity, and how some people dealt with
it.The description of the Irish wake is something that I will long remember with
emotions of sadness as well as joy.If you are interested in Irish history, I
believe that you will enjoy this great book.", "overall": 5.0, "summary":
"Enjoyable Historical Novel", "unixReviewTime": 1282867200, "reviewTime": "08
27, 2010"}
```

As this popular format is not only machine but also human readable, it can be seen that the reviews are written by two different persons and concern two different books. The first review is not very favorable (overall rating 3) and has a rather short text with only one sentence. The second one is extremely favorable (rating 5) and four sentences long. Both reviews have been voted as helpful by 1 out of 1 voters. A summary of each review is given in table 3. It shows the number of sentences, words (tokens) and unique words (types).

Table 3 Reviews summary

Text	Types	Tokens	Sentences	lines
review1	24	26	1	1
review2	56	76	4	2

Following preliminary text mining practices each review text is then represented as a bag-of-words which contains the appearance frequency of each word per text. After common transformations, such as the removal of stop words, punctuation, and numbers the final representation used is the term-document matrix. It is a mathematical matrix that describes the frequency of terms that occur in a collection of documents here customer reviews.

Table 4 Term-document matrix

docs	first	half	book	well	written	last	chapters	rather ..
text1	1	1	2	1	1	1	1	1 ...
text2	0	0	2	1	0	0	0	0
...	...	...	...	...	...	...	...	...

As can be seen in table 4 the term book appears twice in both reviews. This is the main data structure that will be used in the topic modeling approaches that will be applied here to Amazon and Twitter datasets.

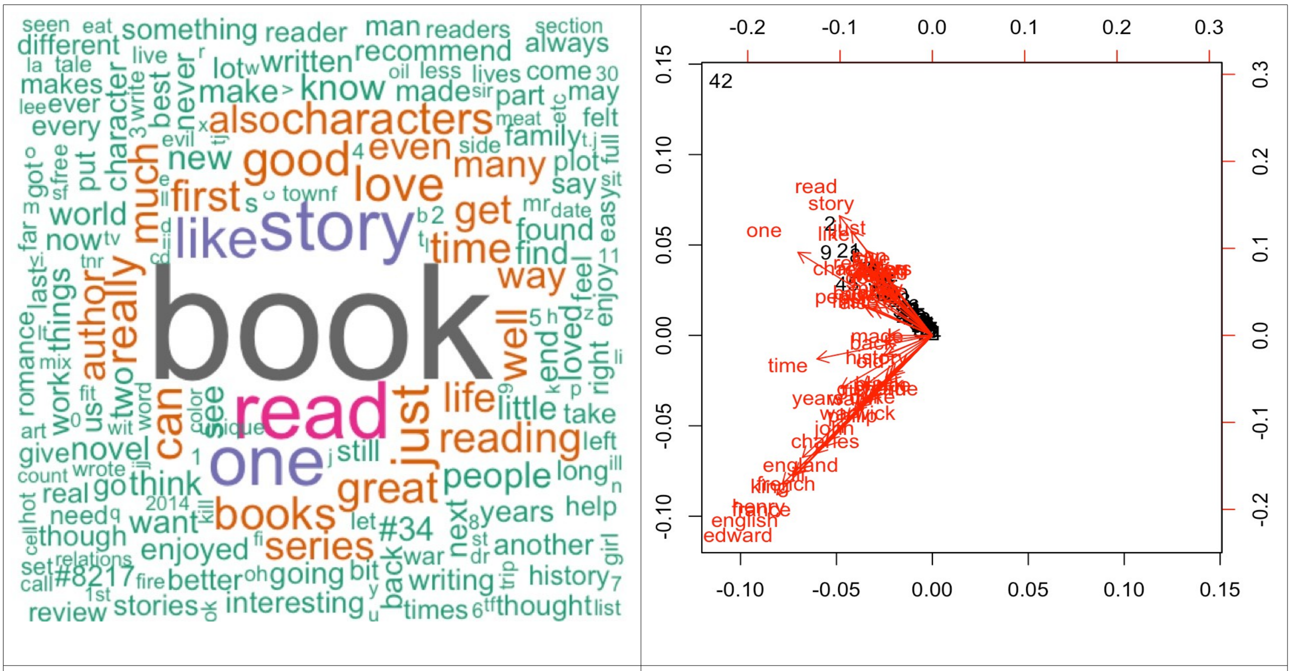
We will essentially introduce models capable of reducing the feature space generated by text mining approaches in order to predict ratings. The Amazon Reviews dataset has widely been used in machine learning research (McAuley et al., 2015, Martin & Pu, 2014) and various models have been tested as to their ability to predict consumer ratings from features extracted from the reviews and from the other non-text elements that are included. The Lasso regression, applied here, uses a

[7] Each review consists of the following labels: (1) reviewerID – the ID of the reviewer; (2) asin – the product ID of the item being reviewed;(3) reviewerName – the name of the reviewer;(4) Helpful – the first number is the amount of people who voted the review as being helpful and the second number is amount of people who voted on the review; (5) reviewText – the entire review in text form; (6) overall – the rating out of 5 that the reviewer gave the product; (7) summary – a shortened version of the review; (8) unixReviewTime – time of the review (9) reviewTime – time of the review in dd/mm/yyyy.

form of Regularized Least Squares that like Ridge regression is suited when the number of independent variables is big, and has the advantage over the latter to automatically select more relevant features and discard the others. For the calculations we applied the latest big data cluster computing technologies Apache-Spark and the Scala language as show in the Appendix.

A classical data reduction technique is Principal Components Analysis. It has been applied in order to retain a small number of dimensions (components) regrouping the most important features (words) to represent the content of the reviews. An interactive web application using a small set of randomly chosen book reviews can be found at http://marketing.iae.univ-lille1.fr/shiny/sample-apps/textminreviews/ . The main outputs are shown in figure 1

Figure 1 - Feature reduction using Principal Components Analysis applied to book reviews



(a) word cloud of book reviews

(b) word loadings on PCA components

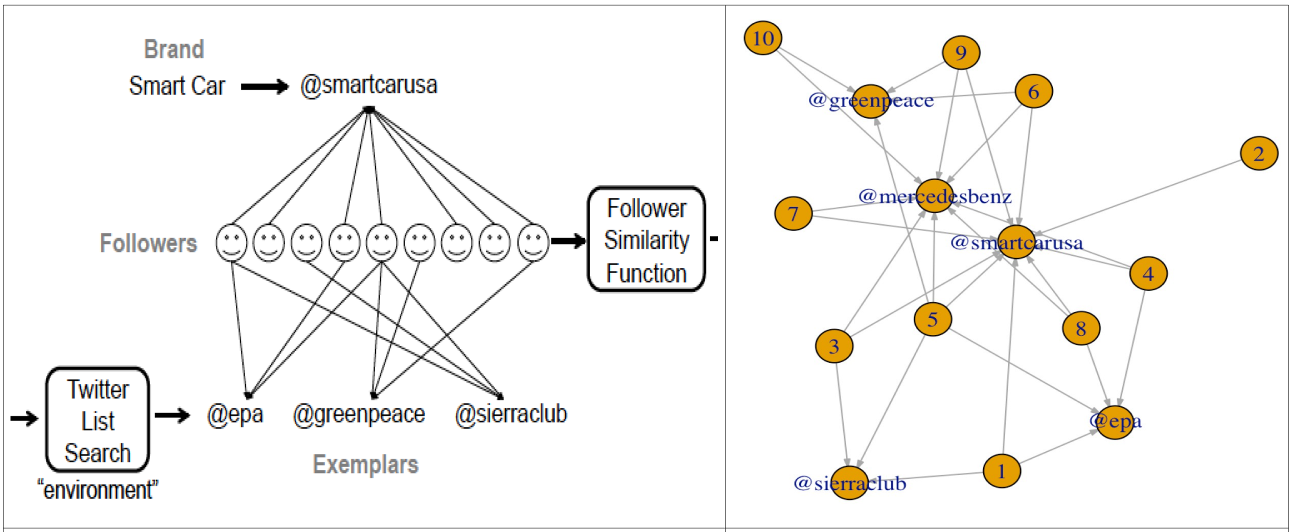
A similar approach has been applied to the *content* dataset using the frequency of all n-grams of tweets (not only words or 1-grams) in all analyzed TV shows. An n-gram is a continuous sequence of n words in the text. For example the Tweet "I love Pittsburgh Steelers" contains four 1-gram, three 2-grams, two 3-grams, and one 4-gram. Collecting the Tweets 24 hours before the show produced 6,894,624 selected Tweets related to the 30 TV series with their 2339 episodes. This generates a significant feature space with 28,044,202 n-grams resulting from Tweets describing the TV episodes mentioned above. It is represented as a "short-and-fat" term-document matrix that is too large to be stored in memory. Applying the PCA dimension reduction technique on such a "fat" matrix required the use of SSVD (Stochastic SVD) method developed by Halko (2012). Both SSVD and the related SVD methods are available in the open-source Mahout machine learning library that applies the well known big data computing approach MapReduce on Hadoop. SVD adapted for MapReduce breaks down into two basic operations, which are matrix multiplication and orthogonalization. As a result four principal components (PC) from the 28,044,202 n-gram features could be selected using the "elbow" rule. Phrases such as "tonight," "can't wait," and "watch" have the largest projection on the first PC. Overall, the first four PCs cover consumers' intention to watch the shows and are the most important independent variables in the final regression model explaining TV show ratings (R-squared 0.756) which is comparable with the R-squared of the model with only the lagged rating included.

This is also the main contribution of this *text mining paper*, showing that easily accessible online content such as Twitter Tweets carefully extracted, sorted and reduced using big data techniques can provide timely representations of consumer intentions. The research company Nielsen considers real-time twitter data to build TV Audience Rating using its 140 million members in order to give more accurate TV rating information to its clients(Kiron, Palmer, Phillips, & Berkman, 2013).

Social mining: Brand Social Perception Score

Figure 2a illustrates the core methodology used for matching brand and exemplar followers using the brand and exemplar followers datasets in order to adapt a similarity function between the brand and the attribute represented by exemplars[8]. For an interactive web application one could visit http://marketing.iae.univ-lille1.fr/shiny/sample-apps/socmintwit.

Figure 2 – Matching eco-friendly Exemplars followers with Brand followers



(a) Followers of one brand

(b) Followers of two brands

Adapted from Cullota & Cutler (2016)

Observing follower's adjacencies with brands and exemplars in table 5a, the privileged measure was the Jaccard index that defines the similarity of two sets as the cardinality of their intersection divided by the cardinality of their union (table 5b) . In order to keep brands with different numbers of followers comparable, in the Social Perception Score (SPS) the Jaccard scores are normalized by weighting each exemplar inversely proportional to its number of followers[9] (table 5c).
The higher this affinity score, the more strongly consumers associate the brand *B* with the attribute represented by the set of exemplars (***E***).

Table 5 - Building a Brand Social Perceptions Score using Followers of Exemplars

a) Follower adjacencies with Brands (B_i) and Exemplars (E_i)

	B1	B2	E1	E2	E3
1	1	0	1	0	1
2	1	0	0	0	0
3	1	1	0	0	1
4	1	1	1	0	0
5	1	1	1	1	1

b) Brand/Exemplar follower similarity

$$J(B_j, E_i) = \frac{|B_j \cap E_i|}{|B_j \cup E_i|}$$

	@smartcarusa	@mercedesbenz
@epa	0.44	0.33
@greenpeace	0.3	0.5
@sierraclub	0.33	0.22

[8] For the example illustrated in figure 1, the brand smartcar had 11052 followers out of which 953 (8,6%) were also followers of environmental friendliness exemplars.

[9] this is analogous to the "inverse document frequency" adjustment used in information retrieval to encourage documents containing rare query terms to be ranked higher than documents containing common query terms (Manning et al. 2008).

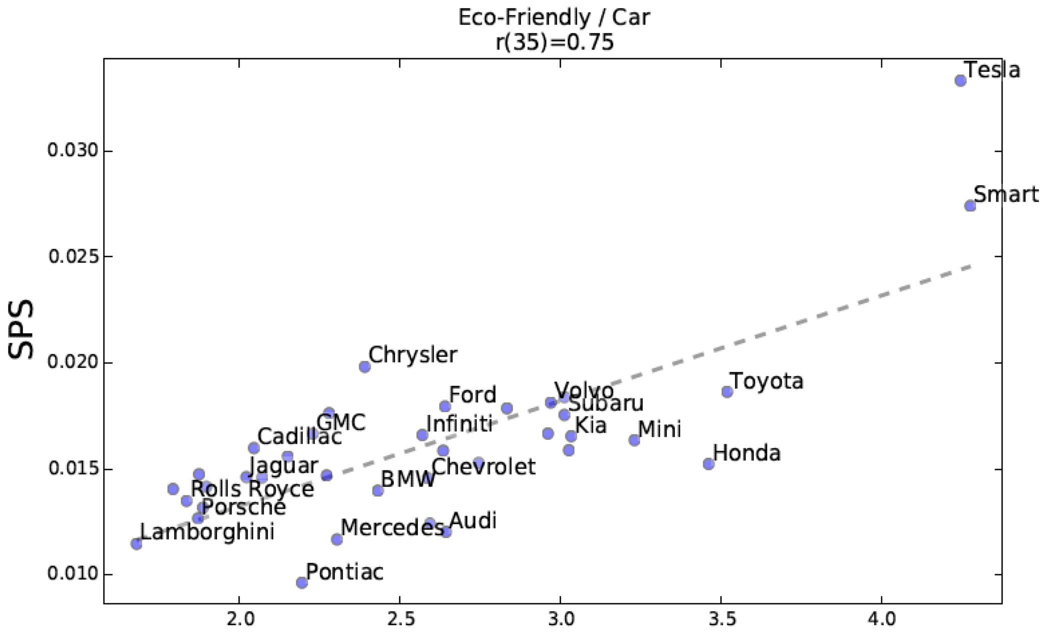
6	1	1	0	1	0
7	1	1	0	0	0
8	1	1	1	0	0
9	1	1	0	1	0
10	0	1	0	1	0

c) Brand SPS $$SPS(B, \boldsymbol{E}) = \frac{\sum_{E_i \in \boldsymbol{E}} \frac{1}{|F_{E_i}|} J(F_B, F_{E_i})}{\sum_{E_i \in \boldsymbol{E}} \frac{1}{|F_{E_i}|}}$$

@smartcarusa	@mercedesbenz
1.13	1.12

This method has the advantage to automatically produce social perception scores for brand attributes that have been shown to be similar to manual survey based ones as is shown in figure 3.

Figure 3 - Correlation between manual survey based and automatic social eco-friendliness scores



Opinion mining and sentiment Analytics

In the *sentiment dataset* mentioned before Tweets are classified by polarity, here positive and negative. A mixed manual vs. automatic classification method is used. Four percent of the Tweets are labeled manually by experts and the rest are labeled automatically using the LingPipe linguistic analysis package (see above). The frequencies of positive and negative Tweets are then used as independent variables in a prediction model for TV show ratings. This and other mainly content and topic modeling approaches, that have more explanatory power, presented by Liu et al. (2016) can be substituted to Nielsen's TV Audience Rating.

Next we present an approach, that we have adapted from Felbermayr & Nanopoulosto (2016), for extracting emotions and negative vs. positive sentiments from online reviews in order to measure their importance based on the review quality perceived and evaluated by other customers through their helpfulness ratings. For an interactive web application one could visit http://marketing.iae.univ-lille1.fr/shiny/sample-apps/sentimentreviews/).

We concentrate our analysis on comparing the rather different reviews emotions and sentiments distribution two product categories that are bought at the same kind of stores, between Video_Games and Toys_Games [10].

[10] We use Toys and Games 5-core (167,597 reviews) and Video Games 5-core (231,780 reviews). These data have been reduced to extract the k-core, such that each of the remaining users and items have k reviews each.

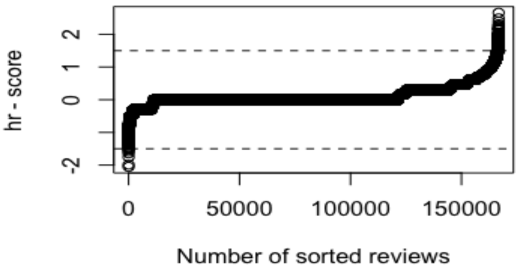
Here are the main steps to follow: 1) derive helpfulness scores ; 2) select indisputably helpful or not helpful reviews; 3) extract emotion features from review texts; 4) building a classifier 5) fitting the model and extracting importance measures of emotions and sentiments for the given product category.

First the helpfulness score is derived from the review field *"helpful": [x, y]* where *x* is the number of voters finding the review helpful and *y* the total number of voters. If *x* stands for positive helpfulness then *y-x* indicates negative helpfulness. A logarithmic expression of positive vs. negative helpfulness ratio intuitively rescales the values without changing the ranking (monotonicity) as can be seen from figure 4a. It allows to discriminate between helpful and not helpful reviews.

Figure 4- Helpfulness score calculations

helpful(x)	Out of (y)	$\left(\frac{x+1}{y-x+1}\right)$	$\log_{10}\left(\frac{x+1}{y-x+1}\right)$
4	4	5	0.7
3	4	2	0.3
2	4	1	0
1	4	0.5	-0.3
0	4	0.2	-0.7
...	...	...	...
1	1	2	0.3
0	1	0.5	-0.3
0	0	1	0

(a) sensitivity analysis of the helpfulness score



(b) application to Toys and Games reviews

Second we observe that all 0 rated reviews have helpfulness ratios x/y that are either 0/0 meaning no votes and do not contribute any helpfulness rating or 1/2 and more generally 0.5y/y rated reviews meaning half of the voter favor the review and half are against it. In the Toys and Games reviews dataset for example (see figure 4b) 66% were zero rated reviews with 62.4% no votes and 3.6% have half favorable/half not votes. As can be seen from figure 4b in the middle region, the majority of reviews show an helpfulness score equal to, or close to, zero. These reviews are hard to be characterized according to their helpfulness, since they have not received any votes, or the amount of positive votes is not substantially distinctive from the amount of negative votes.

Therefore in order to select indisputable helpful and not helpful reviews, in the Toys and Games dataset for example we picked all reviews below hr = -1 and the same amount of reviews with the highest *hr* helpfulness scores.

Third, after common transformations, such as the removal of stop words, punctuation, and numbers, we extract all emotion words per selected reviews checking for their appearance in the NRC emotion lexicon (Mohammad and Turney 2013).

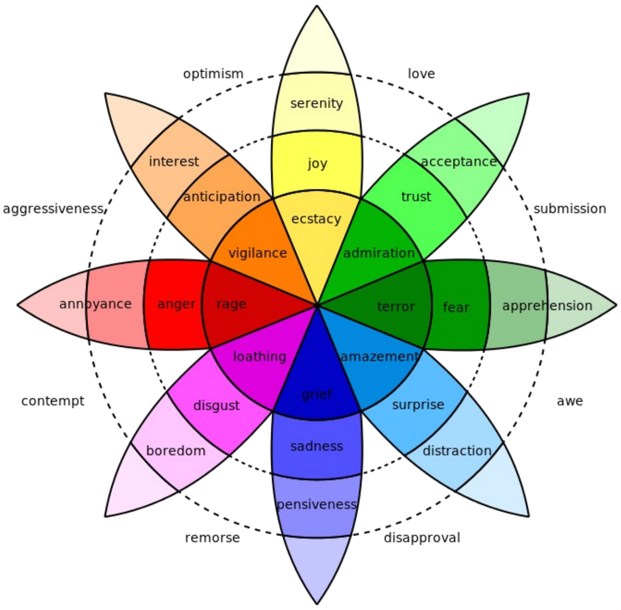
Figure 5b shows Plutchik's model with a bottom (the inner circle), which represents the most intense feelings of a person (e.g., ecstasy, admiration).These feelings might be visibly expressed by the first layer (e.g., joy, trust) and lose their intensity vertically when considering the outer layers (e.g., serenity, acceptance). Mixing the first layer of emotion dimensions would lead to a combined emotion dimension, i.e., joy and trust, combines to love. The eight emotion dimensions from the first layer and two sentiments (negative and positive) are associated to 14,182 words (unigrams) in the NRC Emotions Lexicon we used.

Figure 5 - Emotions and sentiments associated to review text words

Review text: I prefer board games to be fun and simple enough to enjoy with the young ones. While this seemed fun and interesting, we couldn't get into it. We tried a few times but my gamers lost interest.

word	**n**	**sentiment**
board	1	anticipation
enjoy	1	anticipation
enjoy	1	joy
enjoy	1	positive
enjoy	1	trust
fun	2	anticipation
fun	2	joy
fun	2	positive
lost	1	negative
lost	1	sadness
prefer	1	positive
prefer	1	trust

a) words & emotions association in a review



b) the wheel of emotions - based on Plutchik (1980).

Fourth emotion features extracted from the reviews are used to build a classifier. Only the words that appear in both the reviewers' texts and the emotion dictionary are kept (see figure 5a). Each review (r) is represented as a bag-of-words which contains the appearance frequency (f_t^r) of each word or term (t) per review text (t*a* T_r) to which emotions ($\boldsymbol{e_t}$) are associated. Once a review contains at least one term of the dictionary, the corresponding feature vector (v_r) is created:

$$v_r = \sum_{t\,a\,T_r} \left(\sqrt{f_r^t}\boldsymbol{e_t} \right) \qquad (1)$$

The more frequently a term of the emotional lexicon occurs in a review, the more it contributes to the emotional loading of the review. To reduce the impact of the repetitive use of a term, the frequency of terms within one review (f_t^r) is regularized by the square root.

This produces a matrix of reviews (see table 6) expressed as features of emotional and sentiment intensities explaining the helpfulness scores (*hr*)

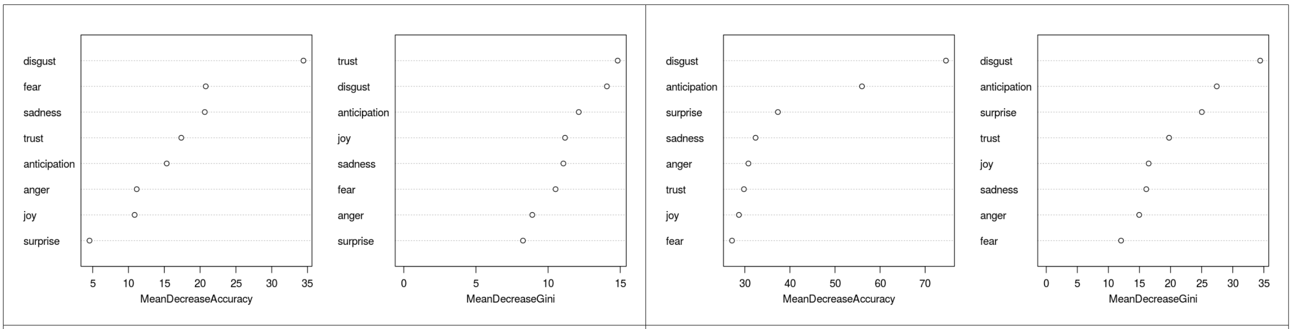
Table 6 - Reviews as features of emotional and sentiment intensities

review	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	positive	negative	hr
1	1	0	1	1	0	2	0	0	0	3	-2.1
2	0	0	0	0	0	1	0	0	0	2	-2
3	2	1	1	1	1	2	0	1	3	3	-1.8
...	...	...	...	...	...	...	...	...	...	...	...
15	0	3.4	0	0	2.4	1	0	2	3.4	1	-1.4
...	...	...	...	...	...	...	...	...	...	...	...
187	2	6.2	1	2	7.7	1	2	4	19.1	4	2.4
188	2	3	2	3.4	5.4	4.4	2	4.4	11.8	4	2.5
189	2	6.2	1	2	7.1	1	2	4	18.5	4	2.7

The 15th review, that has already served as an example in figure 5a appears in this matrix as a row features vector. Following formula 1, it is calculated as the sum over the square root of word frequencies grouped by the emotion and sentiment dimensions they are associated with. For example the anticipation dimension is associated with three terms in this review, one of which , the term "fun", appears twice. Its adjusted frequency score becomes $1+1+\sqrt{2}=3.4$. Please note that in contrast to Felbermayr & Nanopoulosto (2016) we have extended formula 1 to sentiment dimensions.

Fifth our classifier has been trained on the basis of positively and negatively helpful reviews in order to identify the importance of emotion and sentiment dimensions in customer reviews. Random forests (Breiman 2001), the classification methodology used , enables a more intuitive interpretation of the importance of emotion dimensions. Applied to the two datasets it shows a different influence of emotion dimensions for distinct product categories see figure 6.

Figure 6 - Random Forests Importance of emotion dimensions



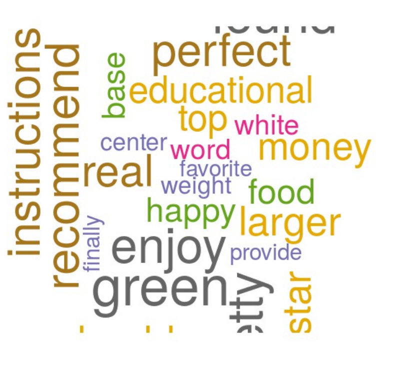
a) Toys and Games Reviews

b) Video Games Reviews

For Toys and Games Reviews the first emotion dimension is “trust” represented word frequencies shown in figure 7a, while for Video Games the first emotion dimension is “disgust” represented word frequencies shown in figure 7b.

Figure 7 - Frequency of words associated to the most important sentiment

a) Toys and Games Reviews - 1st emotion - trust

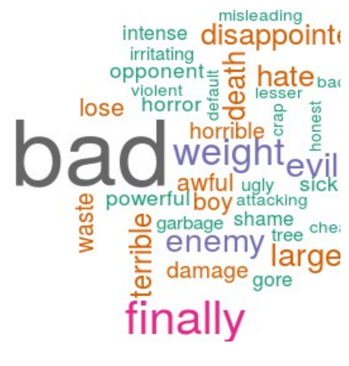


word cloud

word	n
green	27
found	25
enjoy	24
pretty	24
recommend	23
instructions	22
perfect	22
real	21
durable	19
larger	19

frequency

b) Video Games Reviews 1st emotion - disgust



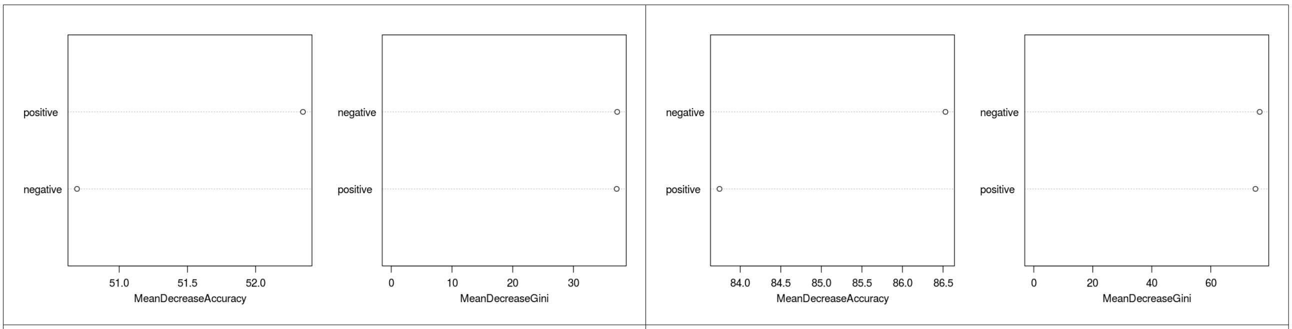
word cloud

word	n
bad	94
finally	41
weight	30
evil	27
enemy	24
larger	23
hate	21
death	20
disappointed	20
terrible	20

frequency

Although less significative, we observe also a different influence of sentiment dimensions for the two product categories see figure 8.

Figure 8 - Random Forests Importance of sentiment dimensions

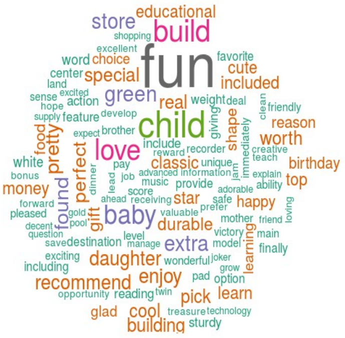
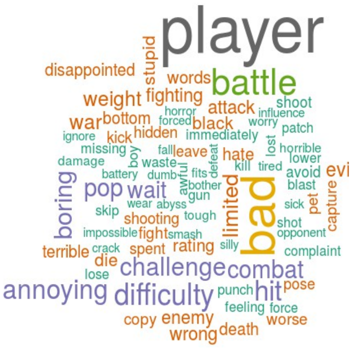


a) Toys and Games Reviews

b) Video Games Reviews

Similarly, for Toys and Games Reviews the first sentiment dimension is “positive” represented word frequencies shown in figure 9a, while for Video Games the first emotion dimension is “negative” represented word frequencies shown in figure 9b.

Figure 9 - Frequency of words associated to the most important sentiment

<table>
<tr><td colspan="2">Toys and Games Reviews – sentiment positive</td><td colspan="2">Video Games Reviews sentiment negative</td></tr>
<tr>
<td>
</td>
<td><table>
<tr><td>word</td><td>n</td></tr>
<tr><td>fun</td><td>96</td></tr>
<tr><td>child</td><td>53</td></tr>
<tr><td>build</td><td>42</td></tr>
<tr><td>love</td><td>39</td></tr>
<tr><td>baby</td><td>35</td></tr>
<tr><td>extra</td><td>27</td></tr>
<tr><td>green</td><td>27</td></tr>
<tr><td>store</td><td>26</td></tr>
<tr><td>found</td><td>25</td></tr>
<tr><td>enjoy</td><td>24</td></tr>
</table></td>
<td>
</td>
<td><table>
<tr><td>word</td><td>n</td></tr>
<tr><td>player</td><td>131</td></tr>
<tr><td>bad</td><td>94</td></tr>
<tr><td>battle</td><td>67</td></tr>
<tr><td>difficulty</td><td>49</td></tr>
<tr><td>hit</td><td>47</td></tr>
<tr><td>annoying</td><td>44</td></tr>
<tr><td>challenge</td><td>41</td></tr>
<tr><td>boring</td><td>39</td></tr>
<tr><td>pop</td><td>39</td></tr>
<tr><td>combat</td><td>38</td></tr>
</table></td>
</tr>
<tr><td>word cloud</td><td>frequency</td><td>word cloud</td><td>frequency</td></tr>
</table>

Conclusion and discussion

This paper presents a hands on approach Social Media Analytics in marketing and applies BigData calculation techniques to potentially vast social media data. To our knowledge it is the first attempt to apply the newer enhanced MapReduce technologies like Spark to marketing science problems.
It contributes to the still very reduced marketing literature that deals with big consumer behavior data using cloud analytics by summarizing some of the main extant academic research and by introducing new applications, datasets and technologies in order to complete the picture.
Explaining their importance, relative simplicity and applicability, based on a variety of marketing datasets, can contribute to the adoption of BigData computational techniques among marketing scientists.

References

Ballings, M., & Van den Poel, Dirk. (2015), CRM in social media: Predicting increases in Facebook usage frequency. *European Journal of Operational Research,* 244,1, 248- 260.

Beyer, M.A., Laney D. (2012) *The Importance of ‘Big Data’: A Definition*, Gartner, Stamford, CT

Breiman, L., (2001), Random Forests, *Machine Learning*, 45, 1, 5–32.

Culotta, A. & Cutler J. (2016), Mining Brand Perceptions from Twitter Social Networks. *Marketing Science*. 35, 3 (May/Jun), 343-362

Fan and Gordon (2014)

Fan, W., & Gordon, M. D. (2014), The Power of Social Media Analytics. *Association for Computing Machinery. Communications of the ACM*, 57(6), 74.

Felbermayr, A., & Nanopoulosto, A., (2016), The Role of Emotions for the Perceived Usefulness in Online Customer Reviews, *Journal of Interactive Marketing,* 36, 60–76

Halko, N.P (2012) Randomized methods for computing low-rank approximations of matrices. Unpublished doctoral dissertation, University of Colorado, Boulder.

Heinonen, 2011).

Heinonen, K., (2011), Consumer activity in social media: Managerial approaches to consumers' social media behavior. *Journal of Consumer Behaviour,* 10, 6, 356- 364.

Kaplan & Haenlein, 2010)

Kaplan, Andreas M, & Haenlein, Michael. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons,* 53, 1, 59- 68.

Kiron, D., Palmer, D., Phillips, A.N., & Berkman, R. (2013), Social business: Shifting out of first gear. *MIT Sloan Management Review*, 55, 1, 1

Kiron, D., Perguson, R.B. & Prentice, P.K. (2013), From Value to Vision: Reimagining the Possible with Data Analytics. *MIT Sloan Management Review*, March.

Kietzmann, J.H, Hermkens, K., McCarthy, I.P. & Silvestre, B.S. (2011), Social media? Get serious! Understanding the functional building blocks of social media. *Business horizons,* 54, 3, 241- 251.

Laney, D. (2001), *3D Data Management: Controlling Data Volume, Velocity, and Variety*, Technical Report, URL https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed October 2017).

Liu, X. Singh, P.V. & Srinivasan, K. (2016), A Structured Analysis of Unstructured Big Data by Leveraging Cloud Computing, *Marketing Science*, 35, 3 (May/Jun), 363-388.

Mangold, G.W., & Faulds, D.J. (2009), Social media: The new hybrid element of the promotion mix, *Business horizons*, 52(4), 357- 365.

Martin, L. & Pu, P. (2014), Prediction of Helpful Reviews Using Emotions Extraction, *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

McAuley, J., Pandey, R. & Leskovec J (2015) Inferring networks of substitutable and complementary products, *KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Mohammad, S.M. & Turney P.D. (2013), Crowdsourcing a Word-Emotion Association Lexicon, *Computational Intelligence*, 29, 3, 436–65.

Plutchik, R. (1980), A General Psychoevolutionary Theory of Emotion, *Theories of Emotion*, 1.

Trusov, M., Bucklin, R.E., & Pauwels, K. (2009), Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site. *Journal of marketing*, 73, 5, 90- 102.

Wagner et al., 2010

Zeng & al. (2010).

Wagner, D., Lopez, M., Doria, A., Pavlyshak, I., Kostakos, V., Oakley, I. & Spiliotopoulos, T. (2010). Hide and seek: location sharing practices with social media. *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*

Zeng, D., Chen, H., Lusch, R. & Li, S-H. (2010), Social media analytics and intelligence. *Intelligent Systems, IEEE*, 25, 6, 13- 16.

Appendix

Listing 2 - Measuring customer sentiment on the Amazon Reviews Dataset*

```
import org.apache.spark.ml.feature.{HashingTF, IDF, Tokenizer}
import org.apache.spark.ml.regression._
import org.apache.spark.ml.{Pipeline, PipelineModel}
import org.apache.spark.ml.tuning.{ParamGridBuilder, CrossValidator}
import org.apache.spark.ml.evaluation.RegressionEvaluator

# Load dataset and cache it
val data = spark.read.json(/media/storage1/reviews-train.json).cache()

# Define a pipeline combining text feature extractors + linear regression
val tokenizer = new Tokenizer().setInputCol("reviewText").setOutputCol("words")
val hashingTF = new HashingTF().setInputCol("words").setOutputCol("features")
val lasso = new LinearRegression().setLabelCol("overall").setElasticNetParam(1.0).setMaxIter(100)
val pipeline = new Pipeline().setStages(Array(tokenizer, hashingTF, lasso))
val paramGrid = new ParamGridBuilder().addGrid(lasso.regParam, Array(0.005, 0.01, 0.05)).build()

# Define evaluation metric
val evaluator = new RegressionEvaluator().setLabelCol("overall").setMetricName("r2")
val cv = new CrossValidator().setEstimator(pipeline).setEvaluator(evaluator).setEstimatorParamMaps(paramGrid)

# Run everything!
val cvModel = cv.fit(data)

#Evaluate on test data:
val test = spark.read.json("/media/storage1/reviews-test.json")
var r2 = evaluator.evaluate(cvModel.transform(test))
println("Test data R^2 score:", r2)

val sparkPredictions = cvModel.transform(test)
sparkPredictions.write.format("json").mode("overwrite").save(/media/storage1/predictions.json)
```

** The listing is adapted to scala from a python version (Bradley, 2016)*