L. Molteni & J. Ponce de Leon, Int. J. of Design & Nature and Ecodynamics. Vol. 11, No. 3 (2016) 220–229

FORECASTING WITH TWITTER DATA: AN APPLICATION TO USA TV SERIES AUDIENCE

L. MOLTENI[1] & J. PONCE DE LEON[2]
[1]Decision Sciences Department, Bocconi University, Milan, Italy.
[2]Target Research, Milan, Italy.

ABSTRACT

Various researchers and analysts highlighted the potential of Big Data, and social networks in particular, to optimize demand forecasts in managerial decision processes in different sectors. Other authors focused the attention on the potential of Twitter data in particular to predict TV ratings. In this paper, the interactions between television audience and social networks have been analysed, especially considering Twitter data. In this experiment, about 2.5 million tweets were collected, for 14 USA TV series in a nine-week period through the use of an ad hoc crawler created for this purpose. Subsequently, tweets were classified according to their sentiment (positive, negative, neutral) using an original method based on the use of decision trees. A linear regression model was then used to analyse the data. To apply linear regression, TV series have been grouped in clusters; clustering is based on the average audience for the individual series and their coefficient of variability. The conclusions show and explain the existence of a significant relationship between audience and tweets.
Keywords: Audience, Forecasting, Regression, Sentiment, TV, Twitter.

1 INTRODUCTION

Many authors in these years highlighted the potential of Big Data in general to support better quantitative forecasting approaches to economic data. In particular, a recent and very interesting work from Hassani and Silva [1] summarized the problems and potential of Big Data forecasting. In short, the main issues concern skills (need for both information system and statistical competences), signal and noise (increasing noise to signal ratio), hardware and software (system crashes handling the increased data input), statistical significance (ability to distinguish between randomness and statistically significant outcomes), architecture of algorithms (data loaded in memory vs. distributed data) and Big Data itself (difficulty to transform unstructured data into structured and rapid evolution and change of data over time). Nevertheless, the application of Big Data in forecasting shows a clearly increasing success, with major results in economics (macroeconomic indexes and monetary policy), population dynamics, crime, energy supply, environment (weather), biomedical science and media. We can find examples of operational application in the retail sector with companies such as Edited, using Big Data for forecasting future fashion collecting data from social media, to the airline sector (improving traffic forecasting and pricing management), to the media sector, for instance, considering the well-known Netflix use of Big Data forecasts for decision making prior to producing the TV show 'House of Cards'. Other contributors, such as Marr [2], highlighted the increasing role of Big Data in the retail sector, in particular, in predicting trends, forecasting demand (Ozon.ru case history), optimizing pricing (Walmart and Stage Stores), identifying profitable customer segments (Macy's) and increasing the conversion rates of online stores (Swedish e-commerce platform Klarna).

WIT CONFERENCES
This paper is part of the Proceedings of the International Conference on Big Data (Big Data 2016)
www.witconferences.com

ISSN: 1755-7437 (paper format), ISSN: 1755-7445 (online), http://www.witpress.com/journals
DOI: 10.2495/DNE-V11-N3-220-229

Focusing the attention specifically to forecasting based on social network data, Arias, Arratia and Xuriguera [3] highlight that exploiting the wealth of information coming from the huge repositories generated by Twitter and Facebook has become of strategic importance for companies. The authors considered in particular Twitter data: as widely known, in this microblogging site, users form social networks with other users that allow them to broadcast short messages of text called 'tweets'. Tweets may concern any topic and users are absolutely free to broadcast what they want; they are publicly available through Twitter's API and their free availability has made them a privileged source of information for academic researchers, companies and institutions. As well explained in the paper, by downloading huge number of tweets and using appropriate natural language and sentiment analysis techniques, it is possible to get an idea of the general mood about a specific topic of interest, in a given place and time. Based on this, many researchers in various fields built a sentiment index from tweets and tried to correlate it with a target time series, for instance presidential polls [4], stock prices [5–7], box office revenues [8] and also TV ratings [9]. This last paper and the results of a recent Nielsen research [10] inspired our work.

To be precise, the efforts of Wakamiya, Lee and Sumiya were more dedicated to find a complement or an alternative to the traditional TV ratings developed by Nielsen Media Research [10] than trying to anticipate TV ratings using Twitter data. The authors start from the consideration that the Set Meter (electronic device monitoring homes viewing) or People Meter (remote controller allowing to recognize specific family members watching TV sets) approaches are rather limited due to the fact that they are necessarily sample based (with a sample quite large considering the overall cost, but relatively small considering specific areas or targets) and they can hardly take into account new viewing behaviours (mobility and nonlinear viewing). The authors developed a system to gather tweets for a region designated for the analysis, building a tweet database including fields like time, location by latitude and longitude and texts containing a reference to a specific TV programme chosen from an Electronic Program Guide (EPG) database; lastly, they built a synthetic rating for each TV programme based on a triple relevance measure (textual, spatial and temporal) filtering out also irrelevant tweets.

Nielsen Media Research analysis is more similar to our investigation; the researchers considered some fall premieres of brand new USA TV programmes and they tried to evaluate how Twitter TV activity, tracked on a 24/7 basis ahead of the premieres, could have been used to anticipate the sizes of the audiences that watched the premiere episodes of those programmes. Twitter TV activity did prove out as an additional signal that could be used together with other factors to anticipate premiere audience sizes.

2 TWEET DATASET SETUP

2.1 Choice of TV programmes: relevance of US TV market and TV Series

The market for television content in the United States was worth $176 billion in 2014. The market value includes all stations that broadcast television content, regardless of whether they do it via satellite or terrestrial, in analogue or digital mode; the value is defined as the amount of money obtained by the television networks in the form of advertising sales, subscriptions or public funds.

This market has had a moderate growth in the past 5 years and is expected to have low growth for the next 4.

The higher-value segment of this market is that of subscriptions to paid services with a value of $102 billion, followed by that of advertising sales with $73 billion.

Although the market value of the subscriptions to pay-TV broadcasting services is significantly higher than that of free television, we decided to focus on the latter because the audience for free channels is much higher (and so probably Twitter activity related to these programmes).

At the beginning, we focused the attention on 'The Five Majors' NBC, CBS, ABC, FOX and The CW, broadcasting across the whole country and covering more than 95% of the population. Within the free TV schedules, as TV series is one of the most successful programme category we decided to focus on the audience and the Twitter activity related to 14 TV prime time series, during a nine-week period (from March 22nd to May 10th, 2015). The choice of the number of TV series and weeks is related to the opportunity to get enough observations for the analysis (considering changing of the schedules due to Easter holidays and some discontinuity in the weekly broadcasting).

To choose the specific TV series to analyse, we used four different criteria, which, in order of relevance, are the following:

1. Number of episodes remaining before the end of the season (since the information is not directly available from the broadcaster, as a predictor we used the average of the total number of chapters of the previous seasons)
2. Average ratings of the previous season, in order to choose the series most viewed
3. The network, in order to differentiate the champion on several television networks
4. The day of the week, to get a sample distributed throughout the week in order to monitor different audiences

Table 1 reports the TV programmess included in the analysis, together with the broadcasting network, the average ratings in the previous season and the scheduled day of the week.

FOX and The CW networks have been discarded at the end because the ratings of their TV series were considerably lower compared to those of the other majors. As a choice criterion we preferred

Table 1: TV series selected for the analysis.

TV series	Network	Ratings (previous season)	Day
The Big Bang Theory	CBS	19.96	Thursday
NCIS	CBS	17.9	Tuesday
Madam Secretary	CBS	11.9	Sunday
Blue Bloods	CBS	10.88	Friday
Criminal Minds	CBS	10.8	Wednesday
CSI	CBS	10.46	Wednesday
The Good Wife	CBS	10.3	Wednesday
Hawaii Five-0	CBS	10.1	Friday
The Blacklist	NBC	10.1	Thursday
Chicago Fire	NBC	9.7	Tuesday
Grey's Anatomy	ABC	8.8	Thursday
Revenge	ABC	8.44	Sunday
2 Broke Girls	CBS	8.43	Monday
American Crime	ABC	8.3	Monday

to give greater importance to the audience than to the equilibrium in the sample distribution of broadcasting networks.

2.2 Twitter data collection

To collect the tweets of the selected TV series, we needed to develop a specific script in the Python programming language, able to interact with the Twitter Streaming API that allows for real-time downloading of tweets containing certain keywords (the tweets collected in this way represent more or less 1% of the total amount of tweets). The selected keywords were the official hashtag of the TV series, as well as some possible hashtag derivatives, and the official account of the television programme. From an operational standpoint, an efficient Twitter crawler needs a stable internet connection, as every disconnection interrupts the flow of tweets. For this reason, we decided to rent a Canadian server with a guaranteed 99% 'up time'. In any case, it was appropriate also to set up a local crawling backup to limit the risk to lose relevant information.

In total, we collected 2,417,430 tweets monitoring 75 different episodes of the 14 selected TV series; the overall average audience was 8.91 million viewers per episode (see Table 2 for details).

We can observe large differences in terms of the quantity of tweets between a TV series and another and a substantial lack of correlation among tweets and ratings. The majority of the difference in Twitter activity is probably due to the relevant differences in the TV series content and in the audience socio-demographic characteristics (for instance: Grey's Anatomy percentage of 18–49 viewers is equal to 33.6% against the 16.1% we get for NCIS). The outlier value reported for Grey's

Table 2: Collected tweets and average ratings.

TV series	Network	Cluster	Episodes num.	Aver. Viewers (million)	CV Viewers (%)	Total Tweets
Blue bloods	CBS	1	5	10,76	6,9	25,348
Madam Secretary	CBS	1	4	10,57	5,3	19,639
Criminal Minds	CBS	1	4	10,36	2,1	122,298
Hawaii Five-0	CBS	1	4	8,63	2,5	60,001
The Good Wife	CBS	1	6	8,50	4,5	49,636
CSI	CBS	1	6	7,94	6,5	39,937
The Blacklist	NBC	1	5	7,34	8,3	108,833
Chicago Fire	CBS	1	6	7,00	6,0	117,747
2 Broke Girls	NBC	1	3	6,78	1,0	13,846
Grey's Anatomy	ABC	2	7	7,98	11,9	1,341,113
American Crime	ABC	2	7	4,41	13,7	59,544
Revenge	ABC	2	6	4,28	8,6	226,024
The Big Bang Theory	CBS	3	6	15,81	6,0	85,142
NCIS	CBS	3	6	14,44	3,1	148,422
		Total/ Av.	75	8,91		2,417,530

Anatomy is mainly due to one episode (April 23rd), with one of the main characters definitely (and definitively) leaving the scene.

For the subsequent analyses we decided to separate the TV series into three qualitative clusters, looking at the average and the variability data of the audience: one (cluster 3) is composed of NCIS and the Big Bang Theory, the two TV series getting the highest average number of viewer (14–16 million), another one (cluster 2) of Grey's Anatomy, Revenge and American Crime, TV series with a medium-low number of viewers (4–8 million) and with the largest relative variation in the audience, with the remaining 9 TV series in the last group (cluster 1 – 7–10 million viewers on average).

2.3 Sentiment analysis on tweets

Usually, sentiment analysis is carried out by applying a set of more or less complex rules to the text to be analysed. The method most used consists in the creation of specific dictionaries that assign a numerical value to each word according to their significance; then words with a positive meaning, as 'beautiful', 'good', etc. will have a value greater than zero, while those with a negative connotation, like 'hate', 'ugly', etc. will get a value less than zero. The neutral words simply are not counted. Considering the arithmetic sum of all values encountered within the text, a rule must be chosen to define a text 'positive', 'negative' or 'neutral'. Though it is possible to apply this approach also to tweets, these kinds of systems have been widely studied and well applied to long and relatively clear texts such as reviews or blogs; but, taking into account short texts steeped in sarcasm or slang, like most tweets, it becomes less useful [11].

For this reason, we decided to implement an approach to tweets categorization based on Knime Decision Trees algorithm [12]. Instead, in order to use a set of predefined dictionary and rules to classify tweets into 'positive', 'negative' or 'neutral' categories, we manually classified a set of 14,000 tweets (1,000 per programme) carefully reading their contents and evaluating from a qualitative point of view not only words meaning but also words association and the impact of emoticons and slang. For instance, we classified as 'positive' tweets containing reference to happiness, joy, love, admiration, friendship, affinity, enthusiasm, and the rest of the positive feelings; as 'neutral' spam, advertising, aseptic citations, and all the other tweets that do not have an intrinsic sentiment; as 'negative', sadness, pain, hate, irritation, frustration, boredom, anger, rancour, and the rest of the negative feelings. Then, after an accurate pre-processing phase to 'clean' the tweets, and the necessary tweets vectorization to build the explanatory variables set (presence/absence of specific words in every tweet), we launched the algorithm obtaining an average precision of the reclassification equal to 65% (around 70% considering only neutral/ intense reclassification). Finally, applying the tree rules to the whole tweets database, all the tweets were classified into the three selected categories, generating the distribution for each TV series shown in Fig. 1.

2.4 Final tweet database

On the basis of the proposed tweet sentiment analysis, a dataset was built containing as observations every single monitored episode of each TV series, and as fields, the number of viewers per each episode (checking carefully for outliers and eliminating a very small number of them – 4 episodes in three different TV series), the total number of tweets and the number of positive, neutral and negative tweets considering the following time-windows:

1. Within 3 hours of the start of the episode
2. Within 24 hours of the start of the episode

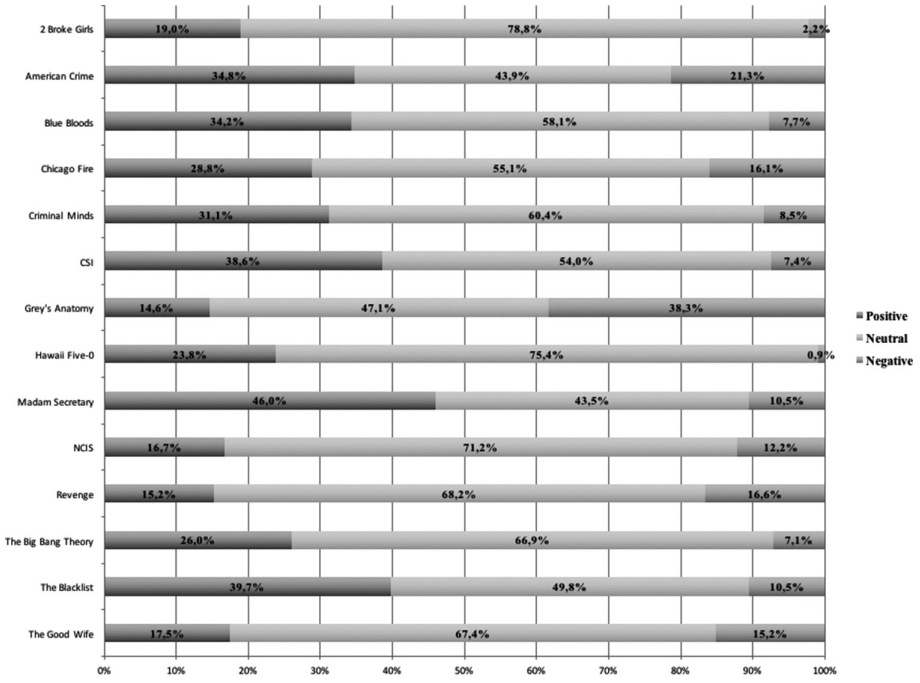

Figure 1: Tweets classification per series.

3. Within 7 days of the start of the episode
4. Since the end of the previous episode to the beginning of the episode
5. During the previous episode

To remove the size-effect and to concentrate more on the potential influence of the tweets intensity we built also the same variables on tweets considering the percentage of positive, negative and neutral over the total.

3 ANALYSIS RESULTS – LINEAR REGRESSION MODEL

We applied the linear ordinary least squares model to the obtained dataset, using *number of viewers* as the dependent variable and *tweets sentiment classification* with *different time windows* as explanatory variables. We produced separated models per each cluster of TV series and we took into account where necessary the different level in the audience of TV series belonging to the same cluster using level dummy variables. To select among the potential predictors, we used a forward selection approach with enter p-value set at a 5% level. Detailed results for each cluster follow below.

3.1 Cluster 1 results

This cluster is the largest one, including 9 series and 43 observations; due to the selected variables to enter (in particular *percentage of negative tweets from last episode*), we lose 8 observations for missing values and so 35 observations are used in the estimation. Figure 2 shows the regression output.

The obtained R^2 is very high (94.1%) and so test F significance (more than 99.9%). Apart from the TV series specific level dummy variables (with a t-test significance higher than 99.9%), we

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,970[a]	,941	,923	,42844

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	76,116	8	9,515	51,834	,000[b]
	Residual	4,773	26	,184		
	Total	80,889	34			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6,560	,237		27,652	,000
	MadameSecretary	2,431	,266	,509	9,126	,000
	BlueBloods	2,927	,277	,539	10,576	,000
	Hawaii5O	1,564	,299	,288	5,225	,000
	Blacklist	-1,148	,283	-,240	-4,059	,000
	Criminal Minds	2,899	,284	,534	10,197	,000
	Chicago	-1,743	,294	-,321	-5,922	,000
	% NEG Tweets since last ep	13,472	2,533	,335	5,319	,000
	% INT Tweets 24 before	1,606	,754	,136	2,129	,043

Figure 2: Regression output for TV series cluster 1.

found a relevant impact of the *percentage of negative tweets since last episode* (for every percentage point more we see an increment of the viewers on average equal to 134,720 people – significance 99.9%) and of the *percentage of intense tweets since 24 hours before the episode* (for every increasing percentage point we see an increment of the viewers on average equal to 16,060 people – significance higher than 95%).

3.2 Cluster 2 results

The cluster contains three TV series with a relatively large audience variability in the sample period, with a total of 20 available observations; as in the previous case, due to the selected variables to enter (*percentage of negative tweets from last episode*), we lose 3 observations for missing values and so 17 observations are used in the estimation. Figure 3 shows the regression output.

In addition, the obtained R^2 is very high (88.1%) in this case and so test F significance (more than 99.9%). We found again a relevant impact of the *percentage of intense tweets since last episode* (for every percentage point more we see an increment of the viewers on average equal to 72,190 people – significance 99.7%) and of the *total tweets since 24 hours before the episode* (for every 10,000 tweets more we see an increment of the viewers on average approximately equal to 15,000 people – significance higher than 99.9%). So, for this cluster, the 'quantity' of tweets and not only the prevalent sentiment seems relevant.

3.3 Cluster 3 results

The cluster contains only two TV series and in particular the ones with the largest audience in the selected sample time window; anyway we have a minimum number of available observations (12)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,939[a]	,881	,864	,71641

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	53,104	2	26,552	51,734	,000[b]
	Residual	7,185	14	,513		
	Total	60,289	16			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,46433	,541		4,553	,000
	% INT Tweets since last ep	7,21900	2,042	,372	3,536	,003
	TOT Tweets 24 before	,00015	,000	,702	6,673	,000

Figure 3: Regression output for TV series cluster 2.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,856[a]	,733	,674	,57561

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8,197	2	4,098	12,369	,003[b]
	Residual	2,982	9	,331		
	Total	11,179	11			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	12,931	,507		25,529	,000
	Final_NCIS	-1,778	,515	-,686	-3,452	,007
	% NEG Tweets 7 days before	28,715	5,949	,960	4,827	,001

Figure 4: Regression output for TV series cluster 3.

to try to build a simple regression model. Figure 4 shows the regression output for this small cluster.

The obtained R^2 is good (73.3%) and test F significance is again very high (more than 99.9%). We needed to introduce a dummy variable (Final_NCIS – significance 99.3%) to take into account a clear drop in audience in the last two episodes included in the sample time window (this drop was

present also in the previous seasons). In this case, the *percentage of negative tweets 7 days before the episode* shows a high significance of the *t*-test (99.9%): for every increase of a percentage point, the two considered TV series get on average an increase of viewers equal more or less to 287,150.

4 CONCLUSION AND SUGGESTIONS FOR FUTURE RESEARCH

4.1 Managerial implications

The provided analysis suggests a significant relationship between the percentage of negative or intense tweets (in particular in the week before the airing) and the audience of an episode of a TV series. As is well known, correlation does not mean causation and so it is obvious that the main determinants of the success of a TV series have to be found in factors different from the activity of viewers on social networks. However, if Twitter TV activity did prove out as an additional signal that could be used together with other factors to anticipate audience sizes, confirming Nielsen Research first studies and conclusions, we can draw at least three simple managerial indications:

a. Advertising agencies could fine-tune their advertising plans before the episodes
b. Networks could identify potential strengths and weaknesses of their TV shows earlier to maximize ad sales and course-correct marketing activities (and possibly the content of the shows themselves).
c. To the extent that social media leads people to become aware of new shows, networks could leverage Twitter TV activity to better reach their intended audiences.
d. So, finally, if the findings do not necessarily mean that Twitter TV activity causes larger audience sizes, the results we showed, together with the results of previous studies, increasingly indicate that Twitter TV activity and reach data can help networks and agencies to make superior, data-driven advertising and program marketing decisions.

4.2 Limits and suggestion for future research

The analyses carried out for this paper have methodological limitations dictated by the need to reduce time and costs. So, this experiment should be repeated taking into account a number of important elements to make up for all the possible methodological limitations. Among the others, we suggest to:

a. increase the length of the period of data collection: from 8–9 weeks to a full year in order to collect tweets and audience for a series for the whole duration of a season;
b. raise the number of series monitored: the number of observations used in this paper seems minimal to obtain significant results, and obviously it is advisable to significantly increase the sample size;
c. increase the training sample for the sentiment analysis: probably this will improve the performance of the decision trees classification;
d. add additional points of polarity for the sentiment analysis: this analysis was carried out using three points polarity (positive, negative, neutral). It would be interesting to measure the effects on predictive models by adding two more points polarity (strongly positive and strongly negative);
e. check if the presence of certain words within the tweet has effects on the prediction of the audiences.

REFERENCES

[1] Hassani, H. & Silva, E.S., Forecasting with big data: a review. *Annals of Data Science*, **2**(1), pp. 5–19, 2015.
http://dx.doi.org/10.1007/s40745-015-0029-9

[2] Marr, B., Big Data: A Game Changer in the Retail Sector, *Forbes*, available at http://www.forbes.com/sites/bernardmarr/2015/11/10/big-data-a-game-changer-in-the-retail-sector/, 2015.

[3] Arias, M., Arratia, A. & Xuriguera, R., Forecasting with twitter data, special issue on social web mining. *ACM Transactions on Intelligent Systems and Technology (TIST)*, **51**, 2013.

[4] O'Connor, B., Balasubramanyan, R., Routledgex, B.R. & Smith, N.A., From tweets to polls: linking text sentiment to public opinion time series. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Washington, pp. 122–129, 2010.

[5] Zhang, X., Fuehres, H. & Gloor, P.A., Predicting stock market indicators through twitter "I hope it is not as bad as I fear". *Procedia-Social and Behavioral Sciences*, **26**, pp. 55–62, 2011.
http://dx.doi.org/10.1016/j.sbspro.2011.10.562

[6] Wolfram, M.S.A., Modelling the stock market using twitter. *M.S. Thesis, School of Informatics, University of Edinburgh*, 2010.

[7] Bollen, J., Mao, H. & Zeng, X., Twitter mood predicts the stock market. *Journal of Computational Science*, **2**(1), pp. 1–8, 2011.
http://dx.doi.org/10.1016/j.jocs.2010.12.007

[8] Mishne, G. & Glance, N., Predicting movie sales from blogger sentiment. *Proceedings of AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs* (AAAI-CAAW), pp. 155–158, 2006.

[9] Wakamiya, S., Lee, R., Kawai, Y. & Sumiya, K., Crowd- powered TV viewing rates: measuring relevancy between tweets and TV. *Database Systems for Advanced Applications,* pp. 390–401, 2011.

[10] Nielsen Media Research, Must see TV: how twitter activity ahead of fall season premieres could indicate success, available at http://www.nielsen.com/us/en/insights/news/2015/must-see-tv-how-twitter-activity-ahead-of-fall-season-premieres-could-indicate-success.html, 2015.

[11] Mihanović, A., Gabelica, H. & Krstić, Z., Big data and sentiment analysis using KNIME: Online Reviews vs. Social Media. *Information and Communication Technology, Electronics and Microelectronics* (MIPRO), Opatija, 2014.

[12] Wakade, S., Shekar, C., Liszka, K.J. & Chan, C., Text mining for sentiment analysis of twitter data. *International Conference on Information and Knowledge Engineering*, (IKE'12), Las Vegas, pp. 109–114, 2012.