

Machine Learning for Supply Chain's Big Data: State of the art and application to Social Networks' data

Radouane El-Khchine^{1,a}, Amine Amar², Zine Elabidine Guennoun², Charaf Bensouda¹ and Youness Laaroussi²

¹ Ibn Tofail University, Department of Mathematics, Kenitra, Morocco

² Mohamed V University, Department of Mathematics, Rabat, Morocco

Abstract. In the context of today's pattern of globalization and a huge amount of information, a smart supply management chain is required. Naturally, statistics and operations research are used for optimizing supply and demand objectives. However, the new context brings out new opportunities at descriptive, predictive and prescriptive levels for supply chain network design, logistics and distribution and strategic sourcing. The key question is still how to capture and to use information. One striking example can be taken from social media, where their use allow to gain insight into the perception of consumers and to capture a real time overview of consumer reactions, regarding one or more specific events. In this regard, different modern approaches, such as IoT or Quantum neural network, are developed. In the same line of thought, we propose an analytic approach, based on KNN, Logistic Regression and SVM with the use of Twitter data in chicken supply chain management. Results identify the main concerns related to chicken products and allow to the development of a consumer-centric supply chain. The proposed approach can be extended to other topics such as anomaly detection and codification of customer intelligence.

1 Introduction

Supply chain management is presented as the field which consists, in parts, to provide the right quantities of goods most efficiently at the right place in the right order within the right time. Meeting these demands requires planning the inbound logistics. The process of planning depends on frequently changing information of product development, assembly line planning and purchasing. Currently, a high amount of time is spent for gathering information during planning and existing knowledge from previous planning processes. Planning process can be separated into strategic (long-term) planning which generates an initial evaluation for feasibility of different plant and supplier locations to integrate new products into production network, tactical (mid-term) inbound logistics planning which focuses on the engineering of logistics process alternatives and their evaluation and operational (short-term) planning of logistics before start of production and where all preselected logistics process and resources will be continuously detailed and integrated into the production plant by pre-series processes during the ramp up.

Analytics in supply chain management is not a new task[1]. For a long time, supply chain management has use statistics and operation research for optimizing the objectives of matching supply and demand. Business analytics using information system support has a strong relationship to supply chain performance [2]. However, the development of big data indeed brings out new opportunities. The term supply chain analytics can be used to define the advanced big data analytics in supply chain management [3]. This analytics can be categorized into descriptive, predictive and prescriptive analytics [4].

Big data analytics permits users to capture, store, and analyse huge amount of data, from internal as well as external of the organization, from multiple sources such as corporate database, sensor-captured data such as RFID, mobile-phone records and locations, and internet in order to understand the meaningful insights [5]. Akter et al. [6] suggested that big data analytics has a big impact to enhance firm performance. By improving big data analytics capability, a firm could create new products and services, provides better customer service, increases sales and revenue, and expands into the new market. Zhong et al. [7] provided the discussion on the big data applications in various sectors such as financial services sector, healthcare, logistics, and manufacturing. However, present research reveals that there is a limited agreement regarding the performance of big data that support supply chain management, especially for strategic sourcing, supply chain network design, product design and development; demand planning, procurement, production, inventory, logistics and distribution, supply chain agility and sustainability.

Consequently, this article aims to explore the application of social media big data and its analysis in supply chain management, especially for chicken supply chain.

In this regard, we propose an analytic approach, based on a comparison between three machine learning models: K-Nearest Neighbours (KNN), Logistic Regression, and Support Vector Machine (SVM) with the use of Twitter data related to chicken supply chain. Results identify the main concerns for chicken products and allow to the development of a consumer-centric supply chain. The proposed approach can be extended to other topics such as anomaly detection and codification of customer

^a Corresponding author: radouane.elkhchine@gmail.com

intelligence. The rest of this paper is organized as follow: the next section expose several works related to the analyse of supply chain using big data. The third part explains preparation of data extracted from Twitter. In the fourth section, we present theoretical overview of methods and algorithms used, and the processing steps, when the fifth part presents the obtained results. This document ends with conclusion and an overview of further works.

2 Related works

Food products supply chain, such as chicken products pass through several steps, from production (farmer, abattoir and processor), to retailer and consumer. To have a constantly improved supply chain, continuous information should be analysed in every step, especially the one provided by consumer [8, 9]. Indeed, more and more industries are nowadays pursuing a consumer-driven supply chain [10, 11].

To meet aforementioned goals, researchers have employed social media information to retrieve valuable insights [12]. In fact, consumer opinion, given by comments on online social media provide management with unprecedented opportunities to leverage collective consumer intelligence for enhancing supply chain management and sales forecasting [13]. Several methods could be used for extraction of intelligence from tweets, such as Naïve Bayesian classifier, support vector machine, or artificial neural networks [10]. The sentiment analysis and social media are used to study several phenomena in different filed such as finance, medicine and others [14, 15].

N. Shukla et al. [16] propose a big data analytics based approach, which considers Twitter data for identifying supply chain management issues in beef supply chain. The results indicated that the proposed text analytic approach can be helpful to identify crucial customer feedback for supply chain management.

Other research works are conducted to investigate the effectiveness of e-retailers' logistics-related customer service interactions on Twitter with a view towards identifying effective and ineffective social media customer service strategies [17]

Social media have been used by many businesses, W. He et al. [18] apply text mining to analyze unstructured text content on Facebook and Twitter of three largest pizza chains. The results show the competitiveness of social media analysis and the power of text mining to extract business value from the vast amount of available social media data.

3 Data

To harvest Twitter posts, we developed a python script that uses Twitter API, we used a list of keywords to get data related to a specific topic. This list contains 22 keywords of the logistics and supply chain's jargon

combined with the word "chicken". We then got all the 10 days of the available historical data for each keyword.

The script send as much queries as needed to get all available tweets, each query returns a JSON file with 100 posts and a huge amount of related information we call metadata. Each file should be parsed to keep only needed information. In our case, we've been interested in variables like: Twitter post's ID so we can follow the historical line while harvesting data, the post itself is the more important variable, the place so we can do some geographical analysis, the retweeted count, the keyword used to get the tweet, date, time, and time zone when the tweet was posted.

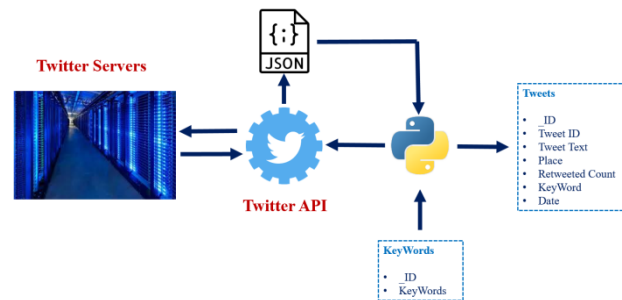


Figure 1.Data extraction process from Twitter servers to our databases.

The database where we store our data contains six tables: keywords, raw tweets table with different variables which will be cleaned and transformed to another table in order to train and test models, a backup table for raw tweets, a table for stop words, those are commonly used words we ignore, and a table of new tweets' extractions on which final classification model will be applied.

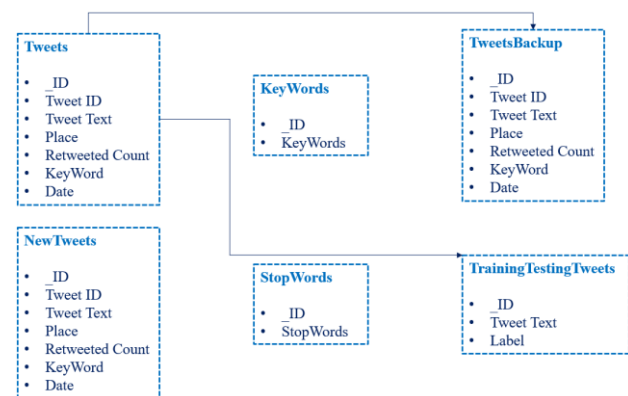


Figure 2.Database schema



Figure3.Keywords list from the “KeyWords” table

As demonstrated by Morstatter et al.[19], the available free data collected using Twitter API is a good and sufficient representations of the full data on Twitter. This allows us to generate results obtained to the general consumer present on Twitter.

4 Proposed Methodology

4.1 Overview of used Models

The extracted data will be analyzed based on three main machine learning classification algorithms, K-Nearest neighbors (KNN), support vector machine (SVM) and logistic regression.

4.1.1 K-Nearest neighbors (KNN)

KNN algorithm is one of the most basic classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. KNN is a non-parametric algorithm, i.e. it does not make any assumptions on the underlying data distribution and the model structure is completely determined from the data. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point, which will be classified according to the K (integer number) nearest points on the database [20].

4.1.2 Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper-plane, it belongs to the supervised learning field. The SVM was developed in the 1990s and was extremely popular around the time and continue to be the high-performing algorithm with little tuning [21].

Given labelled training data, the algorithm outputs an optimal hyper-plane, dividing a plane in two parts where each class lay in a side, the hyper-plane maximizes the margin between classes.

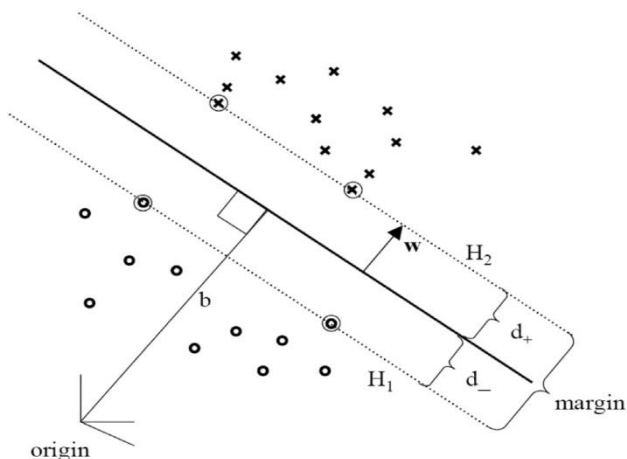


Figure4.Maximum margin hyper-plan for the SVM model

The margin is expressed by the following equations:

$$\begin{aligned} \vec{w}^T \vec{x}_i + b &\geq +1 & \text{for } y_i = +1 \\ \vec{w}^T \vec{x}_i + b &\leq -1 & \text{for } y_i = -1 \end{aligned}$$

Both constraints can be combined into one set of inequalities:

$$\vec{y}_i(\vec{w}^T \vec{x}_i + b) - 1 \geq 0 \quad \forall i$$

Maximize the margine is equivalente to:

$$\text{Max} \frac{2}{\|\vec{w}\|} \Rightarrow \text{Min}\|\vec{w}\| \Rightarrow \text{Min}\Phi(\vec{w}) = \frac{1}{2} \vec{w}^T \vec{w}$$

This is a classic optimization problem that could be resolved using Lagrangian method.

In some cases, data are not linearly separable, we use then a transformation to another space where it is easier to find the optimal hyper-plan.

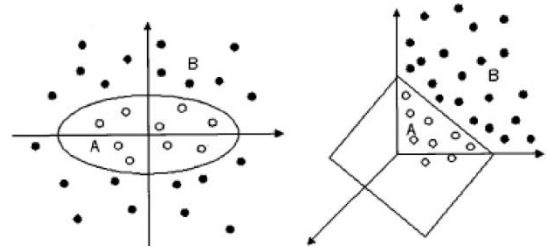


Figure5.Space transformation for non linear separable data

4.1.2 Logistic regression

Logistic regression is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome, called dependent variable, which is a binary variable. The goal of logistic regression is to find the best fitting model to describe the relationship between the dependent variable and a set of independent variables.

Logistic regression predicts the probability of occurrence of an event by fitting a logistic function to data, called also, the sigmoid function and defined by:

$$S(x) = \frac{1}{(1 + e^{-x})}$$

We can remark that the curve given by S has a finite limit of 0 as x approaches $-\infty$ and 1 as x approaches $+\infty$ and when $x=0$ is $S(x)=0.5$.

Thus, if the output is more than 0.5, we can classify the outcome as 1 (or YES) and if it is less than 0.5, we can classify it as 0(or NO) [22].

4.2 Processing Steps

The data on social media is highly unstructured. In fact, tweets contain text, URLs, hashtags, numbers, emoticons, and emoji. Therefore, appropriate text cleaning approach is required for effective knowledge gathering. Since there is no optimal data cleaning way, we develop our specific approach, depending on each case.

Indeed, after extracting the appropriate tweets that populate the main table, this table contains 13 693 entry. Since the models shouldn't be trained and tested on data with high similarity scores, the first step consists of deleting duplicates tweets and retweeted ones, next step is to perform a similarity test on the remaining tweets, so we reject tweets with similarity score higher than 80%, this operation took more than 5 hours, at this stage, we've got 4 906 tweets left. We used then a dictionary approach to label tweets; this generates labels between -5 and 5 depending on how many positive or negative dictionary words the tweet contains. Tweets with scores between -1 and 1 are dropped. Finally, we keep only 1 509 tweets with positive labels, and 1 509 tweets with negative labels. Those are the tweets populating the training and testing table of our database.

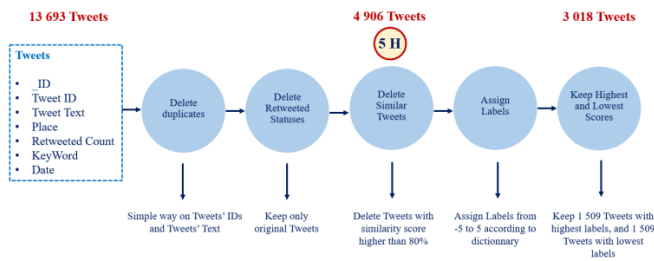


Figure6. Different steps for data cleaning process

The input for each used models should be under a matrix form, therefore, each tweet should pass through a “words to vectors” process. This consists in building a dictionary of all different words in all cleaned tweets, so each tweet could be represented by a vector.

First step is to remove “stop words”, then delete words with length less than 3 characters, next step is to transform the words into their basic forms using python library NLTK, then transforming capital letters to small ones, then it is necessary to add new words to the dictionary. At this stage, a vector with dimension of dictionary size could represent a tweet by putting in each word on the dictionary the number of times it appears on the tweet, the values should be standardized, finally, we assign labels, 0 to negative and 1 to positive tweets.

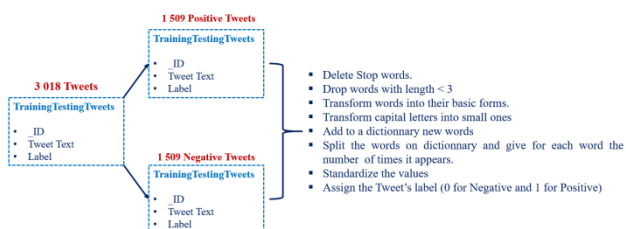


Figure7. Different steps of “Words to Vectors” process

	Word 1	Word 2	Word N	Label
Tweet 1	0	0.2	0.6	1
Tweet 2	0.23	0	0	1
Tweet 3	0.12	0	0.06	1
:	:	:	:	:
:	:	:	:	:
Tweet 1509	0.1	0.32	0.05	1

Tweet 1	0.3	0.15	0.1	0
Tweet 2	0.31	0	0	0
Tweet 3	0	0.01	0.02	0
:	:	:	:	:
:	:	:	:	:
Tweet 1509	0.32	0	0.08	0

Figure8. Final shape of input data after “Words to Vectors” process

In this study, three models are compared on same data for several times to measure which one performs the best to classify tweets. Before proceeding to this comparison, a previous study should be done to find best models' hyper-parameters to be used.

For SVM model, we used a grid search on several values of “Gamma” and “C” hyper parameters, for each combination, a corresponding SVM model is trained on data and tested three times after a random shuffle. Best scores are obtained for Gamma = 1 and C = 10, we choose a radian basis function as kernel.

We obtain the best hyper-parameter for KNN model by testing several values of K = 3, 5, 7, 10 on data. The best scores are shown for K = 5. The logistic regression model we used doesn't have any hyper-parameters.

Once all hyper-parameters known, we run the comparison of the three models on data. We first shuffle the matrix 20 times randomly, we split the data into training set of size 2 518 tweets and a testing set of 500 tweets, the three models are trained and tested on same data, and the scores are stored. This process is repeated 500 times, and took more than 15 hours.

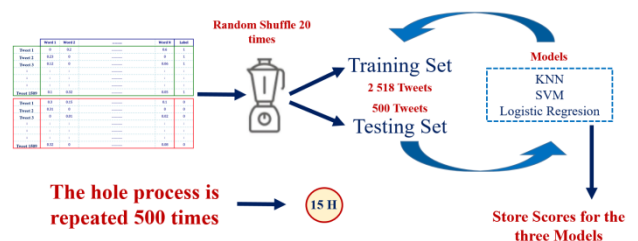


Figure9. Process for training and testing models to perform comparison between SVM, KNN, and Logistic regression

At this level, we could compare the three models and choose the best one.

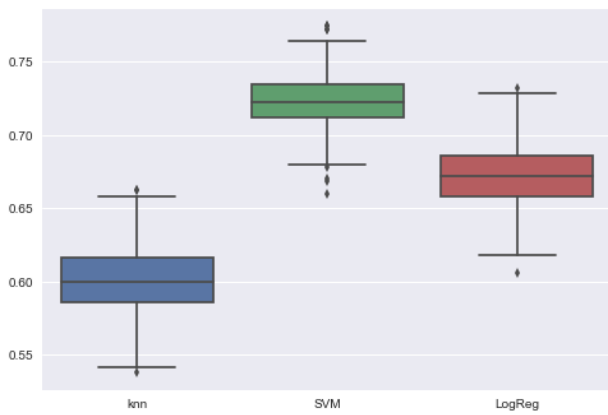


Figure10.Results of comparison between SVM, KNN, and Logistic regression

The KNN model made in average a score of 60% and may reach 66%, the logistic regression model records in average a score of 67% on testing data and can reach 73%, the SVM model scores on testing set in average around 72% and reaches 78%. Logistic regression model got the best score in 0.8% of cases and SVM is the best model in 99.2% of cases.

The SVM model with Gamma = 1, C = 10, and a radian basis function kernel is the one we'll be using next. We fit this model to the training data as much as needed until we got the SVM parameters that reach at least a score of 78% on testing data. The final SVM model made a score of 78.8% on testing set.

5 Results

After performing the benchmark study which compares performance of KNN, Logistic regression, and SVM models, we do know the model that works the best on our kind of data. At this step, we perform a second massive extraction of Twitter data, respecting same keywords as before, during another 10 days. We obtained 11 032 of raw tweets, after a cleaning process (removing exact duplicates, removing URLs, removing emoticons and emoji, ...), we have a number of 6 417 tweets left on which our SVM model will be run.

Results of classification show sentiments related to the chicken supply chain based on consumer feedbacks on Twitter.



Figure11.Results of the SVM classifier on the 6 417 new tweets

Each point on the data visualization above is a one of the 6 417 tweets, 49% of customers on Twitter are expressing

positive opinion about chicken supply chain. On the other side, 51% of feedbacks are negative.

It will be interesting to explore a geographical distribution of positive and negative feedbacks and how it is different from a country to another.

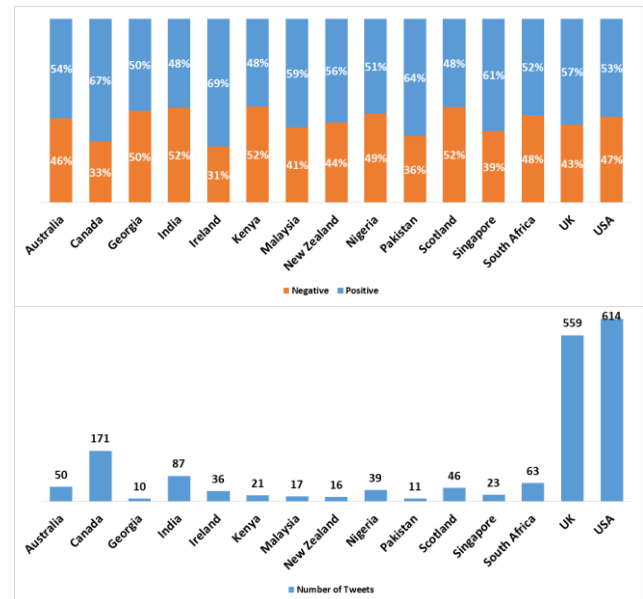
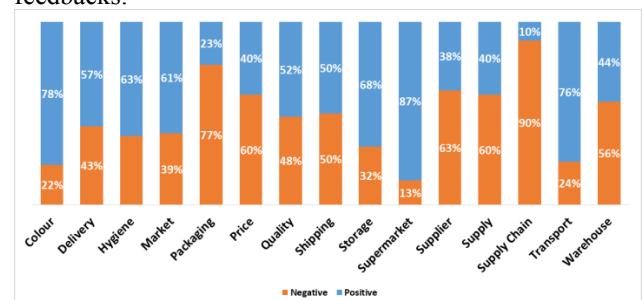


Figure12.Results of the SVM classifier on some chosen countries

Higher proportions of positive opinions related to chicken supply chain are shown in Ireland with 69%, followed by Canada where 67% of customers are happy with the quality of chicken supply chain, then United Kingdom with 57% of satisfied people on Twitter.

On the other hand, higher percentages of negative feedbacks arises from India and Scotland with 52%, then 48% for South Africa.

Another way to analyse results could be spreading the 6 417 tweets by keywords to figure out which one cumulate more positives or negatives consumers' feedbacks.



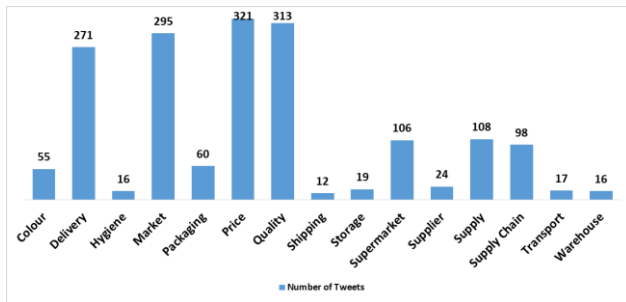


Figure13. Results of the SVM classifier on some chosen keywords

Higher parts of positive feedbacks are shown for keyword “Flavour” (90%), followed by the keyword “Supermarket” with 87%, and keyword “Colour” (78%), then 57% of opinions related to keyword “Delivery” are positives.

For negative opinions, 77% of consumers reacting on Twitter about “Packaging” are not happy, keyword “Price” has a proportion of 60%, then 57% of people speaking about “Meat” have a negative opinions.

6 Conclusion and further works

Analysis of the given information from social media, is used to understand the chicken supply chain, based on consumer feedback on Twitter. The study of such type of data conduct to understand reasons for positive and negative sentiments, identify communication nature and remarks of Twitter users discussing about several aspects related to chicken supply chain.

Consumers are using social media like Twitter to express their point of view on chicken meet quality (color, taste, price, packaging, ...), which generates plenty of useful available information to develop and improve supply chain strategy. This information is big in size considering its volume, variety and velocity and it is vague and unstructured in nature. In the proposed methodology, customers’ tweets associated with chicken supply chain are being extracted, sorted and organized into tables with different variables which will be cleaned and transformed in order to train and test models. A new table of tweets is extracted on which final classification model will be applied.

Based on this analysis, results of classification showed interesting insights on how consumers are reacting on Twitter about topics related to chicken supply chain. Twitter provides several metadata that could be used for more advanced analysis like geographical data, date and time for time series analysis on how opinions polarity progresses, users, etc.

For businesses that adopt a consumer-centric strategy especially in supply chain management, this kind of studies could be very helpful to understand more and more consumers’ needs.

In future, an enhanced list of keywords could be used for further analysis of the issue. Twitter analytics could be employed for longer time duration and, like time series analysis on tweets classifications could be performed. We also intend to use machine learning models to generate scores rather than binary classifications and using more advanced machine learning models such as deep artificial neural networks on larger datasets.

References

1. G.C. Souza, Supply chain analytics. *Business Horizon*, **57**, 595–605 (2014).
2. P. Trkman, K. McCormack, M.P.V. de Oliveira, and M.B. Ladeira, *Decision Support System*, **49**, 318–327 (2010).
3. G. Wang, A. Gunasekaran, E. W. Ngai, and T. Papadopoulos, *International Journal of Production Economics*, **176**, 98-110 (2016).
4. S. Tiwaria, H.M. Weeb, and Y. Daryanto, *Computers & Industrial Engineering*, **115**, 319-330 (2018).
5. J. Zakir, T. Seymour, K. Berg. *Big data analytics. Issues in Information Systems*, **16** (2), 81–90 (2015).
6. S. Akter, S.F. Wamba, A. Gunasekaran, R. Dubey and S.J. Childe, *International Journal of Production Economics*, **182**, 113-131 (2016).
7. R.Y. Zhong, S.T. Newman, G.Q. Huang, and S. Lan, *Computers & Industrial Engineering*, **101**, 572-591 (2016).
8. R. Handfield, T. Linton. *The LIVING Supply Chain: The Evolving Imperative of Operating in Real Time*; Wiley, ISBN: 978-1-119-30625-2, (2017).
9. M. Hugos. *Essentials of Supply Chain Management*. Wiley, ISBN: 978-0-470-94218-5 (2011).
10. J. Yoon, R. Narasimhan and M. K. Kim, *International Journal of Production* (2017).
11. B. Schulze-Ehlers and S. Anders, *Renewable Agriculture and Food Systems*, **33**, 73-85 (2018).
12. C. Dhaou. *Social media sentiment analysis: lexicon versus machine learning*, *Journal of Consumer Marketing*, **34** (6), (2107).
13. R. Y. K. Lau , W. Zhang, W. Xu. *Parallel Aspect-Oriented Sentiment Analysis for Sales Forecasting with Big Data*, Wiley Online Library (2017).
14. S. Krishnamoorthy. *Sentiment analysis of financial news articles using performance indicators*, *Knowledge and Information Systems*, **56** (2), 373-394 (2018).
15. J. Wei, X. Liao, H. Zheng, G. Chen, X. Cheng. *Learning from context: A mutual reinforcement model for Chinese microblog opinion retrieval*, *Frontiers of Computer Science*, **12** (4), 714-724 (2018).
16. A. Singha, N. Shukla, and N. Mishra, *Transportation Research Part E: Logistics and Transportation Review*, **114**, 398-415 (2018).
17. N. Shukla, N. Mishra and A. Singh, *IMMM* (2017).

18. J. Bhattacharjya, A. Ellison and S. Tripathi, International Journal of Physical Distribution & Logistics Management, **46**, 659-680 (2016).
19. W. He, S. Zha, L. Li, International Journal of Information Management, **33**, 464-472 (2013).
20. F. Morstatter, J. Pfeffer, H. Liu and K. M. Carley, In Proceedings the Seventh International AAAI Conference on Weblogs and Social Media, Boston, MA (2013).
21. O. Sutton, Introduction to k Nearest Neighbor Classification and Condensed Nearest neighbor Data Reduction, (2012).
22. S. R. Gunn, Technical report, University of Southampton, Faculty of Engineering, Science and Mathematics; School of Electronics and Computer Science (1998).