Sentiment Analysis using KNIME: a Systematic Literature Review of Big Data Logistics

Gary Graham
Business School
Leeds University
Leeds, UK
e-mail:g.graham@leeds.ac.uk

Roy Meriton
Business School
Leeds University
Leeds, UK
cen5rfm@leeds.ac.uk

Abstract— **Text analytics and sentiment analysis can help researchers to derive potentially valuable thematic and narrative insights from text-based content such as industry reviews, leading OM and OR journal articles and government reports. The classification system described here analyses the aggregated opinions of the performance of various public and private, medical, manufacturing, service and retail organizations in integrating big data into their logistics. It explains methods of data collection and the sentiment analysis process for classifying big data logistics literature using KNIME. Finally, it explores the potential of text mining to build more rigorous and unbiased models of operations management.**

Keywords-Big data, logistics, sentiment analysis, KNIME, text analytics.

I. Introduction

Big data logistics can be defined as the modelling and analysis of (urban) transport and distribution systems through large data sets created by GPS, cell phone and transactional data of company operations, combined with human generated activity (i.e. social media, public transport) [1]. The demands and requirements are literally changing on a daily basis with the innovation in technologies with smart computing and big data. All types of organization whose logistics operation functions in a big data environment will have to adapt to changing customer demands. At the same time they will need to exploit the availability of big data technology to improve their process and operational capabilities. Big data requires firms to have more technical and technological supports to handle the five V's of Big Data and analytics that is "*Volume*", "*Variety*", *Veracity*", "*Value*" and "*Velocity*" [2]. However, with the growth of big data there is privacy surveillance and data misuse challenges [3]. Organizations also face challenges around quality, comprehensiveness, collection and the analysis of data from various sources. Furthermore, big data also needs to be robust, accessible, and interpretable if it is to provide organizations with meaningful opportunities and solutions.

The purpose of this paper therefore is to explore the risks and challenges of organizations implementing "*big data logistics*" into their operations. Secondly, to investigate the opportunities that big data provides organizations with, to improve their logistics performance. This will be achieved through the text processing of 552 records containing industry reviews, leading OM and OR journal articles and government reports. We will analyse the opinions of the performance of various public and private, manufacturing, medical, service and retail organizations in integrating big data (analytics) into their logistics.

II. KNIME METHOD

The KNIME text processing feature was designed and developed to read and process textual data [4][5], and transform it into numerical data (document and term vectors) in order to apply regular KNIME data mining nodes (for classification and clustering). This feature allows for the parsing of texts available in various formats (here we used .csv) as KNIME data cells stored in a data table. It is then possible to recognize and tag different kinds of named entities such as with positive and negative sentiment, thus enrichening the documents semantically. Furthermore, documents can be filtered (e.g. by the stop word or named entity filters), stemmed by stemmers for various languages pre-processed in many other ways. Frequencies of words can be computed, keywords extracted and documents can be visualized (e.g. tag clouds). To apply regular KNIME nodes to cluster or classify documents according to their sentiment, they can be transformed into numerical vectors.

Web of Science (WOS) and Scopus are powerful databases which provide different searching and browsing options [9]. The search options in both databases are the Standard Basic and Advanced. There are different searchable fields and several document types that permit the user to easily narrow their searching. Both databases sort the results by parameters such as; first author, cites, relevance and etc. The Refine Results section in both databases allows the user to quickly limit or exclude results by author, source, year, subject area, document type, institutions, countries, funding agencies and languages. The resulting documents provide a citation, abstract, and references at a minimum. Results may be printed, e-mailed, or exported to a citation manager. The results may also be reorganized according to the needs of the researcher by simply clicking on the headings of each column. Our search of “big data logistics” documents resulted in 552 records being retrieved from a ten year period from 2006 to 2016.

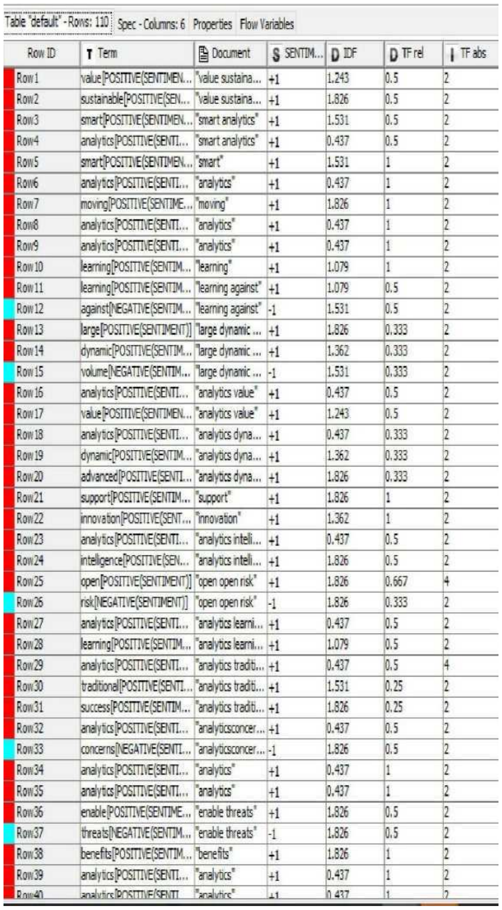
The described data was then loaded into KNIME with the File Reader node and processed. In this phase, only records in English language were collected. Language of the text is set to English and all texts that have different language values are filtered out, because English dictionary applied on reviews and posts written in other languages would not give results. Dictionary built for sentiment analysis of the phrase “*big data*” as it is used with respect to the term “*logistics*” was graded only as positive or negative. Scoring or sentiment analysis of the phrase “*big data logistics*” is done on the positive-negative level, therefore the phrase was analysed on the word level, giving each word associated with it a positive or negative polarity. For instance, efficiency would be scored positive whilst risks would be scored negatively.

For this task, publicly available MPQA subjectivity lexicon was used as a starting point for recognizing contextual polarity [7], this was expanded with a big data vocabulary built from the authors previous papers [3]. The existing dictionary containing of approximately 8000 words is expanded to fit the needs for sentiment analysis in a way that initial portion of sentences are collected, which are separated into single words with Bag of Words processing. Unnecessary words such as symbols or web URLs are filtered out and all useful, big data specific words are graded and added to the dictionary. For instance, “*veracity”*, “*value*”, “*volume*”, “*variety*” and “*velocity*”.

The records were analysed on the word level giving a positive or negative grade for a term connected to each phrase. Whilst text analytics of documents is usually accomplished simply with phrases counters and mean calculations, our analytics is frequency-driven. Two separate work flows were therefore built, one for calculating frequency based on a grade and category, and other one for positive-negative (sentiment) grading. These results are presented in Table 1.

TF*IDF (Term Frequency*Inverse Document Frequency) [7] method assigns non-binary weights related on a number of occurrences of a word. Weighting exploits counts from a background corpus, which is a large collection of documents; the background corpus serves as indication of how often a word may be expected to appear in an arbitrary text. TF*IDF calculation determines how relevant a given word is in a particular document.

Table 1 Big data logistics sentiments



	Row ID	Term	Document	SENTIM...	IDF	TF rel	TF abs
🟥	Row1	value[POSITIVE(SENTIMEN...	"value sustaina...	+1	1.243	0.5	2
🟥	Row2	sustainable[POSITIVE(SEN...	"value sustaina...	+1	1.826	0.5	2
🟥	Row3	smart[POSITIVE(SENTIMEN...	"smart analytics"	+1	1.531	0.5	2
🟥	Row4	analytics[POSITIVE(SENTI...	"smart analytics"	+1	0.437	0.5	2
🟥	Row5	smart[POSITIVE(SENTIMEN...	"smart"	+1	1.531	1	2
🟥	Row6	analytics[POSITIVE(SENTI...	"analytics"	+1	0.437	1	2
🟥	Row7	moving[POSITIVE(SENTIME...	"moving"	+1	1.826	1	2
🟥	Row8	analytics[POSITIVE(SENTI...	"analytics"	+1	0.437	1	2
🟥	Row9	analytics[POSITIVE(SENTI...	"analytics"	+1	0.437	1	2
🟥	Row10	learning[POSITIVE(SENTIM...	"learning"	+1	1.079	1	2
🟥	Row11	learning[POSITIVE(SENTIM...	"learning against"	+1	1.079	0.5	2
🟦	Row12	against[NEGATIVE(SENTIM...	"learning against"	-1	1.531	0.5	2
🟥	Row13	large[POSITIVE(SENTIMENT)]	"large dynamic ...	+1	1.826	0.333	2
🟥	Row14	dynamic[POSITIVE(SENTIM...	"large dynamic ...	+1	1.362	0.333	2
🟦	Row15	volume[NEGATIVE(SENTIM...	"large dynamic ...	-1	1.531	0.333	2
🟥	Row16	analytics[POSITIVE(SENTI...	"analytics value"	+1	0.437	0.5	2
🟥	Row17	value[POSITIVE(SENTIMEN...	"analytics value"	+1	1.243	0.5	2
🟥	Row18	analytics[POSITIVE(SENTI...	"analytics dyna...	+1	0.437	0.333	2
🟥	Row19	dynamic[POSITIVE(SENTIM...	"analytics dyna...	+1	1.362	0.333	2
🟥	Row20	advanced[POSITIVE(SENTI...	"analytics dyna...	+1	1.826	0.333	2
🟥	Row21	support[POSITIVE(SENTIM...	"support"	+1	1.826	1	2
🟥	Row22	innovation[POSITIVE(SENT...	"innovation"	+1	1.362	1	2
🟥	Row23	analytics[POSITIVE(SENTI...	"analytics intelli...	+1	0.437	0.5	2
🟥	Row24	intelligence[POSITIVE(SEN...	"analytics intelli...	+1	1.826	0.5	2
🟥	Row25	open[POSITIVE(SENTIMENT)]	"open open risk"	+1	1.826	0.667	4
🟦	Row26	risk[NEGATIVE(SENTIMENT)]	"open open risk"	-1	1.826	0.333	2
🟥	Row27	analytics[POSITIVE(SENTI...	"analytics learni...	+1	0.437	0.5	2
🟥	Row28	learning[POSITIVE(SENTIM...	"analytics learni...	+1	1.079	0.5	2
🟥	Row29	analytics[POSITIVE(SENTI...	"analytics tradit...	+1	0.437	0.5	4
🟥	Row30	traditional[POSITIVE(SENTI...	"analytics tradit...	+1	1.531	0.25	2
🟥	Row31	success[POSITIVE(SENTIM...	"analytics tradit...	+1	1.826	0.25	2
🟥	Row32	analytics[POSITIVE(SENTI...	"analyticsconcer...	+1	0.437	0.5	2
🟦	Row33	concerns[NEGATIVE(SENTI...	"analyticsconcer...	-1	1.826	0.5	2
🟥	Row34	analytics[POSITIVE(SENTI...	"analytics"	+1	0.437	1	2
🟥	Row35	analytics[POSITIVE(SENTI...	"analytics"	+1	0.437	1	2
🟥	Row36	enable[POSITIVE(SENTIME...	"enable threats"	+1	1.826	0.5	2
🟦	Row37	threats[NEGATIVE(SENTIM...	"enable threats"	-1	1.826	0.5	2
🟥	Row38	benefits[POSITIVE(SENTIM...	"benefits"	+1	1.826	1	2
🟥	Row39	analytics[POSITIVE(SENTI...	"analytics"	+1	0.437	1	2
🟥	Row40	analytics[POSITIVE(SENTI...	"analytics"	+1	0.437	1	2

Besides term frequency fw, d which equals the number of times word w appears in a document, size of the corpus D is also needed. Given a document collection, a word w and an individual document $d \in D$, TF*IDF value can be calculated:

$$TF * IDFw = fw, d * \log \frac{D}{f\ wd} \quad (1)$$

Total score for each word is given by multiplying TF*IDF value with attitude of a term. Attitude can have one of three values depending on the word polarity; -1 for word with negative polarity, +1 for word with positive polarity and 0 for neutral words. Final weights, which now represent attitude of each document , are grouped on the level of document and binned into three bins to give one of three final results for each term; positive, negative or neutral (Table. 2).

Table 2 TF-IDF Processing

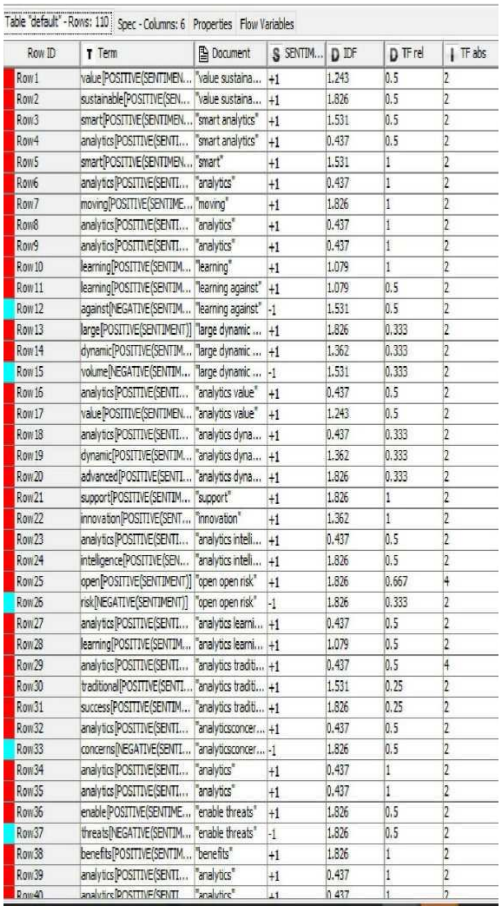


Table "default" - Rows: 110 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	T Term	Document	S SENTIM...	D IDF	D TF rel	I TF abs
🟥 Row1	value[POSITIVE(SENTIMEN...	"value sustaina...	+1	1.243	0.5	2
🟥 Row2	sustainable[POSITIVE(SEN...	"value sustaina...	+1	1.826	0.5	2
🟥 Row3	smart[POSITIVE(SENTIMEN...	"smart analytics"	+1	1.531	0.5	2
🟥 Row4	analytics[POSITIVE(SENTI...	"smart analytics"	+1	0.437	0.5	2
🟥 Row5	smart[POSITIVE(SENTIMEN...	"smart"	+1	1.531	1	2
🟥 Row6	analytics[POSITIVE(SENTI...	"analytics"	+1	0.437	1	2
🟥 Row7	moving[POSITIVE(SENTIME...	"moving"	+1	1.826	1	2
🟥 Row8	analytics[POSITIVE(SENTI...	"analytics"	+1	0.437	1	2
🟥 Row9	analytics[POSITIVE(SENTI...	"analytics"	+1	0.437	1	2
🟥 Row10	learning[POSITIVE(SENTIM...	"learning"	+1	1.079	1	2
🟥 Row11	learning[POSITIVE(SENTIM...	"learning against"	+1	1.079	0.5	2
🟦 Row12	against[NEGATIVE(SENTIM...	"learning against"	-1	1.531	0.5	2
🟥 Row13	large[POSITIVE(SENTIMENT)]	"large dynamic ...	+1	1.826	0.333	2
🟥 Row14	dynamic[POSITIVE(SENTIM...	"large dynamic ...	+1	1.362	0.333	2
🟦 Row15	volume[NEGATIVE(SENTIM...	"large dynamic ...	-1	1.531	0.333	2
🟥 Row16	analytics[POSITIVE(SENTI...	"analytics value"	+1	0.437	0.5	2
🟥 Row17	value[POSITIVE(SENTIMEN...	"analytics value"	+1	1.243	0.5	2
🟥 Row18	analytics[POSITIVE(SENTI...	"analytics dyna...	+1	0.437	0.333	2
🟥 Row19	dynamic[POSITIVE(SENTIM...	"analytics dyna...	+1	1.362	0.333	2
🟥 Row20	advanced[POSITIVE(SENTI...	"analytics dyna...	+1	1.826	0.333	2
🟥 Row21	support[POSITIVE(SENTIM...	"support"	+1	1.826	1	2
🟥 Row22	innovation[POSITIVE(SENT...	"innovation"	+1	1.362	1	2
🟥 Row23	analytics[POSITIVE(SENTI...	"analytics intelli...	+1	0.437	0.5	2
🟥 Row24	intelligence[POSITIVE(SEN...	"analytics intelli...	+1	1.826	0.5	2
🟥 Row25	open[POSITIVE(SENTIMENT)]	"open open risk"	+1	1.826	0.667	4
🟦 Row26	risk[NEGATIVE(SENTIMENT)]	"open open risk"	-1	1.826	0.333	2
🟥 Row27	analytics[POSITIVE(SENTI...	"analytics learni...	+1	0.437	0.5	2
🟥 Row28	learning[POSITIVE(SENTIM...	"analytics learni...	+1	1.079	0.5	2
🟥 Row29	analytics[POSITIVE(SENTI...	"analytics traditi...	+1	0.437	0.5	4
🟥 Row30	traditional[POSITIVE(SENTI...	"analytics traditi...	+1	1.531	0.25	2
🟥 Row31	success[POSITIVE(SENTIM...	"analytics traditi...	+1	1.826	0.25	2
🟥 Row32	analytics[POSITIVE(SENTI...	"analyticsconcer...	+1	0.437	0.5	2
🟦 Row33	concerns[NEGATIVE(SENTI...	"analyticsconcer...	-1	1.826	0.5	2
🟥 Row34	analytics[POSITIVE(SENTI...	"analytics"	+1	0.437	1	2
🟥 Row35	analytics[POSITIVE(SENTI...	"analytics"	+1	0.437	1	2
🟥 Row36	enable[POSITIVE(SENTIME...	"enable threats"	+1	1.826	0.5	2
🟦 Row37	threats[NEGATIVE(SENTIM...	"enable threats"	-1	1.826	0.5	2
🟥 Row38	benefits[POSITIVE(SENTIM...	"benefits"	+1	1.826	1	2
🟥 Row39	analytics[POSITIVE(SENTI...	"analytics"	+1	0.437	1	2
🟥 Row40	analytics[POSITIVE(SENTI...	"analytics"	+1	0.437	1	2

III. RESULTS

Tag clouds were initially used to visualise our initial findings. A simple tag cloud presented in Figure 1 gives the most used words in the positive (left hand cloud) and negative used words (right hand cloud).

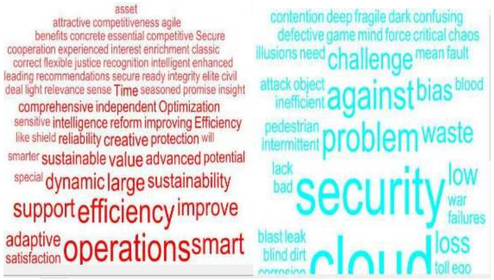

asset
attractive competitiveness agile
benefits concrete essential competitive Secure
cooperation experienced interest enrichment classic
correct flexible justice recognition intelligent enhanced
leading recommendations secure ready integrity elite civil
deal light relevance sense Time seasoned promise insight
comprehensive independent Optimization
sensitive intelligence reform improving Efficiency
like shield reliability creative protection will
smarter sustainable value advanced potential
special dynamic large sustainability
support efficiency improve
adaptive satisfaction operations smart
contention deep fragile dark confusing
defective game mind force critical chaos
Illusions need challenge mean fault
attack object inefficient against bias blood
pedestrian intermittent problem waste
lack bad security low war failures
blast leak blind dirt corrosion cloud loss toll ego

Figure 1 Tag clouds of positive/negative sentiment

The attitudes towards big data were classified as "positive", "neutral" and "negative". Neutral grades can be avoided, and we accomplished this by removing grade bins and removing a bin for neutral grade. The positive and negative grades were aggregated for all terms associated with big data. In Figure 2 it can be seen that sentiments are far more positive (245) than negative (95).

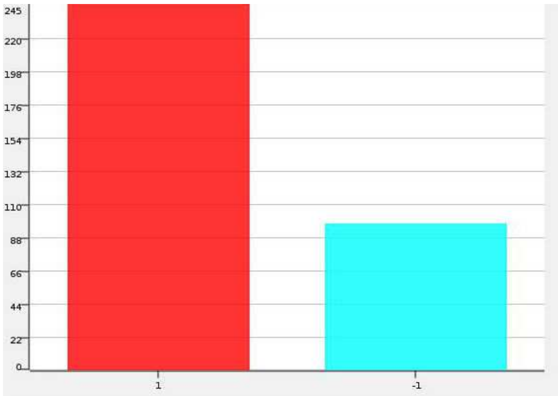


Figure 2 Aggregated sentiments

IV. CLASSIFICATION EXPERIMENT

In order to test the validity of the TF*IDF classification model we ran a prototype experiment with the ten most common words extracted (i.e. those with the highest TF*IDF scores) (see Table 3 below).

Table 3 Most occurring words

Positive	Negative
Agile	Security
Asset	Inefficient
Capability	Confusing
Competitive	Dark
Effective	Challenges
Enrichment	Failures
Optimization	Culture
Flexible	Liability
Intelligence	Complex
Sustainable	Waste

Then using the TF*IDF decision tree learner/predictor approach we tested the accuracy of the classification system (that we had adopted in differentiating the big data logistics sentiments). Our rests are presented in Table 4.

Table 4 Classification accuracy

Classification	TruePo	FalsePo	TrueNe	FalseNe	False No	Recall	Precision	Sensitivity	Specificity	F Measure	Accuracy	Cohen Kappa
Analytics	13	31	12	0		1	0.295	1	0.279	0.456		
Unspecified errors	2	10	44	0		1	0.167	1	0.815	0.286		
											0.268	0.096
	Mean	SD	Skew	Kourtsis								
FalsePo	0.9318	4.871	5.9587	35.7322								
TruePo	0.3409	1.9759	6.4517	41.8415								
TrueNeg	0.7955	6.708	-5.9538	37.1936								
FalseNeg	0.9138	0.8436	0.6156	-0.8109								
Recall	0.0645	0.2497	3.7281	12.717								
Precision	0.2311	0.0911										
Sensitivity	0.0645	0.2479	3.7281	12.717								
Specificity	0.9794	0.1116	-6.0956	38.4034								
F Measure	0.3709	0.1205										
Accuracy	0.8779	0										
Cohen's Kappa	0.0961	0										

Our model shows a predictive accuracy of 88% in classifying the textual data. We then tested using the hierarchical classification function in Knime the ability of the classification model to deal with the addition of features. From Figure 3 we can see by feature 4 that the model peaked at 100% accuracy and then maintained this level of accuracy as features kept being added to it.

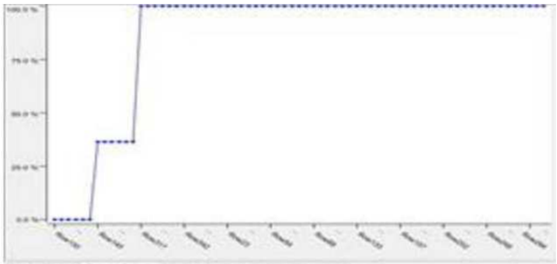


Figure 3 Model features accuracy

So this initial test prototype of the model seems to have a high degree of accuracy and validity in dealing with sentiment classification. However, this is only a prototype of the decision model, so more robust testing will be needed in the future. Specifically, this will provide more stringent MPLA testing for variance.

V CONCLUSIONS

In this paper, we have presented a novel approach to extracting key words and predicting "positive" and "negative" sentiments. We proved the validity of our approach by examining different classifiers that utilized twenty features extracted from the TF*IDF processing [7]. This model is only a prototype to highlight the text processing potential of KNIME [6][8]. In the future, we intend to build comparisons between a range of industrial and retail sectors. We see the role of KNIME potentially as an important mediating step in the framing and building of theoretical frameworks. Furthermore, it could be adopted to build much more grounded and unbiased coding systems of qualitative data. Our work confirms that of Foss Wamba et al., [2] and Mehmood et al., [3], that is, we can confirm there is a growth in opinion on big data, not only at strategic and policy levels, but also with respect to its operational implementation. Thematic patterns and framework categories need building from our extracted key terms. Then, linkages and co-occurrences need exploring to establish a grounded approach for building theory from KNIME and other data mining tools [4][10][11]. As well as positive sentiments theoreticians need to factor in more negative and risk constructs to enable more robust and accurate model development. More in-depth analysis and more discrete modelling are clearly needed to assist in the implementation of big data initiatives [2].

REFERENCES

[1] Blanco, E.E. and Fransoo, J. C. (2013). "Reaching 50 million nanostores: retail distribution in emerging megacities". TUE Working Paper – 404. January.

[2] Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G. and Gnanzou, D. (2015). "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study". International Journal of Production Economics, 34, 2, pp. 77-84.

[3] Mehmood, R., Meriton, R., Graham, G., Hennelly, P. and Kumar, M. (2016). "Exploring the influence of big data on city transport operations: a Markovian approach". International Journal of Operations and Production Management. Forthcoming.

[4] Hofmann, M. and Klinkenberg, R. (2013). "RapidMiner: data mining use cases and business analytics applications". Boca Raton: CRC Press.

[5] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009). "The WEKA data mining software: an update". SIGKDD Explorations, 11, 1, pp. 10–18.

[6] Demšar, J., Curk, T. and Erjavec, A. (2013). "Orange: data mining toolbox in Python". Journal of Machine Learning Research, 14, pp. 2349−2353,

[7] Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T. and Meinl, T. (2008). "KNIME: the Konstanz information miner, in data analysis, machine learning and applications (studies in classification, data analysis, and knowledge organization). Berlin: Springer.

[8] Jović, A., Brkić, K. and Bogunović, N. (2014) "An overview of free software tools for general data mining". Information and Communication Technology, Electronics and Microelectronics (MIPRO), 37th International Convention on IEEE. Available at: http://bib.irb.hr/datoteka/699127.MIPRO_2014_final.pdf

[9] Archambault, É., Campbell, D. and Gingras, Y. (2009). "Comparing bibliometric statistics obtained from the Web of Science and Scopus". Journal of the American Society for Information Science and Technology, 60, 7, pp. 1320-1326.

[10] Lau, K-N., Kam-Hon L. and Ho, Y. (2005). "Text mining for the hotel industry". Cornell Hotel and Restaurant Administration Quarterly, 46, 3, pp. 344-362.

[11] Glaser, B. G. and Strauss, A. L. (2009) The discovery of grounded theory: Strategies for qualitative research. New Jersey: Transaction Publishers. G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*