View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Loughborough University Institutional Repository

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

Gary Graham[1], *Royston Meriton, Patrick Hennelly*

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 opinions of the performance of various public and private, manufacturing, medical, 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 then gives an overview of the differences and explores future possibilities in sentiment analysis for investigating different industrial sectors and data sources.

1. 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 the firm implementing "*big data logistics*" into their operations. Secondly, to investigate the opportunities that big data provides the firm 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. DATA MINING TOOLS

Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories. Data mining has many application fields such as marketing, business, science and engineering, economics, games and bioinformatics.

Text mining or sentiment analysis [4] is the analysis of data contained in a natural language text, which deals with the computation of opinion, sentiment and subjectivity in text. Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information from the text documents. The basic task of sentiment analysis is to determine the polarity of a given texts.

Currently, many data mining and knowledge discovery tools and software are available for every one and different usage such as the Waikato Environment for Knowledge Analysis (WEKA) [5] RapidMiner [4] and Orange [6]. These tools and software[1] provide a set of methods and algorithms that help in the better utilization of data and information available to users; including methods and algorithms for data analysis, cluster analysis, genetic algorithms, nearest neighbour, data visualization, regression analysis, decision trees, predictive analytics and, text mining.

Text mining tools, are not without their limitations. They tend to use a relatively simple dictionary approach to identify associations. This means they cannot identify novel or newly named big data science phrases and terms. Another

[1] Corresponding author, Leeds University Business School, Maurice Keyworth Building, University of Leeds, Leeds, LS2 9JT. T: +44 (0) 113 343 8557, E: g.graham@leeds.ac.uk.

[1] Jovic et al [8] provide a detailed review through a comparative study of diverse collection of data mining tools.

limitation lies in its inability to extract context or meaning from sentences or terms. Methods that use artificial intelligence (AI), word context or machine learning (ML) methods could potentially improve the current term identification system [10]. In spite of these limitations we suggest methodologically that text mining is a useful starting point in building a neutral coding system for theoretical framing and concept development [11].

The tasks of dictionary making and sentiment analysis are done in this paper by the means of KNIME [7], which we found to be a user-friendly graphical workbench capable of entire analysis process. KNIME uses six different steps to process texts: reading and parsing documents, named entity recognition, filtering and manipulation, word counting and keyword extraction, transformation and visualization. The choice of KNIME came because of its straightforward graphical user interface for data pre-processing, its intuitive data flow user interface and from its powerful data mining elements. However, there are also several limitations including its high memory requirements, the complexity of Lab nodes and its syncretic analysis of text as there are limited semantic and pragmatic analysis functions available.

III. KNIME METHOD

The KNIME text processing feature was designed and developed to read and process textual data, 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 recognise 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 visualised (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.

IV. DATA COLLECTION

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.

V. DICTIONARY BUILDING

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*".

VI. DATA SCORING

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 of 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.

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

Figure 2 Big data logistics sentiments

A. Big Data Record Grading

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. 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 $d \epsilon D$, 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 (Fig. 3).

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

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 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

Figure 3 TF-IDF Processing

VII RESULTS

Tag clouds were initially used to visualise our initial findings. A simple tag cloud presented in Figure 4 gives the most used words in the positive (left hand cloud) and negative used words (right hand cloud).

Figure 4 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 5 it can be seen that sentiments are far more positive (245) than negative (95).

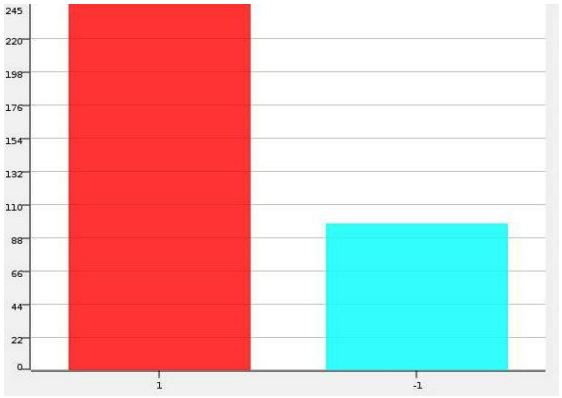

Figure 5 Aggregated sentiments

VIII 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 Figure 6 below).

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

Figure 6 Most occurring words

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 results are presented in Figure 7.

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		

Figure 7 Classification accuracy

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 8 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 8 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.

VIX CONCLUSIONS

In this paper, we studied the classification of opinions towards "*big data logistics*". We 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].

However this model is only a prototype to highlight the text processing potential of KNIME. In the future, we intend to build comparisons between a range of industrial and retailing 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.

Theoretically, it is evident from our initial text processing of the big data literature, that our work confirms that of Foss Wamba et al., [2] and Mehmood et al., [3]. We can confirm there is an evolution taking place from strategic analysis to operational implementation.

Thematic patterns and framework categories need building from the extracted terms. Then, linkages and co-occurrences need exploring to establish a grounded approach for building theory from KNIME and other data mining tools [4]. As well as the positive sentiments that dominate the big data modelling landscape, theoreticians need to factor in more negative and risk constructs to enable more robust and accurate model development.

For practitioners, the advantages of big data to operations is identified to be: "*optimization*", "*intelligence*", "*flexible*" and "*efficiency*". Whilst the risks and challenges to operations from big data are identified as: "*security"*, "*culture*", "*inefficient*", "*waste*".

More in-depth analysis and more discrete modelling are clearly needed to assist in the implementation of big data initiatives [2]. Some of the changes that operations and their connected logistics face are revolutionary and this requires careful consideration from both a practical and theoretical point of view. The description of the new models and the rich context in which these new models are embedded will provide a deep insight to researchers and practitioners in exploring similar opportunities and challenges in their own domains.

References

1. Blanco, E.E. and Fransoo, J.C., (2013) "Reaching 50 million nanostores: retail distribution in emerging megacities". TUE WorkingPaper – 404. January. Available at: http://cms.ieis.tue.nl/Beta/Files/WorkingPapers/wp_404.pdf.

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., 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., 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. The WEKA data mining software: an update, SIGKDD Explorations, vol. 11, no. 1, pp. 10–18, 2009.

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

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), Springer Berlin Heidelberg, pp. 319–326, 2008.
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), 2014 37th International Convention on. IEEE, 2014.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: 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, : 344-362.
11. Glaser, B. G., and Strauss, A. L. (2009) The discovery of grounded theory: Strategies for qualitative research. Transaction publishers, 2009.