**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 operations management (OM) and operations research (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. Although our results show a promising high level of model accuracy, we also suggest caution that the performance of the solution should be compared in terms of the performance of other solutions. This work explains methods of data collection and the sentiment analysis process for classifying big data logistics literature using KNIME (Konstanz Information Miner). 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 global positioning systems (GPS), cell phone and transactional data of company operations, combined with human generated activity (i.e., social media, public transport) [1][3]. 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 [3]. 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 [2].

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 [4][7]. 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.

In Section 2, the KNIME method of text processing is presented including dictionary building, term and inter-document frequency calculations and pre-processing procedures for sentiment analysis. Section 3 reports the results including tag clouds and graphical representation of aggregated sentiments. Then Section 4 presents the key findings from a classification experiment conducted using decision tree analysis on ten of the most occurring positive and negative sentiment words towards big data logistics. Finally, in Section 5, our key conclusions and potential scientific contribution are outlined.

**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, 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, a publicly available MPQA (multiperspective answering) 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.

<table>
<thead>
<tr>
<th>TABLE 1 BIG DATA LOGISTICS SENTIMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF<em>IDF (Term Frequency</em>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<em>IDF calculation determines how relevant a given word is in a particular document. Besides term frequency $f_w, d$ 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</em>IDF value can be calculated:</td>
</tr>
<tr>
<td>$$TF*IDF_w = f_w, d * \log \frac{d}{f_w}$$</td>
</tr>
</tbody>
</table>

Total score for each word is given by multiplying TF*IDF value with attitude of a term (Table 2). 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. TF-IDF PROCESSING

<table>
<thead>
<tr>
<th>Term</th>
<th>Document</th>
<th>DF</th>
<th>TF relative</th>
<th>TF absolute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agile</td>
<td>1</td>
<td>2</td>
<td>1.29</td>
<td>0.5</td>
</tr>
<tr>
<td>Asset</td>
<td>1</td>
<td>2</td>
<td>1.00</td>
<td>0.5</td>
</tr>
<tr>
<td>Capability</td>
<td>1</td>
<td>2</td>
<td>1.29</td>
<td>0.5</td>
</tr>
<tr>
<td>Competitive</td>
<td>1</td>
<td>2</td>
<td>1.00</td>
<td>0.5</td>
</tr>
<tr>
<td>Effective</td>
<td>1</td>
<td>2</td>
<td>1.29</td>
<td>0.5</td>
</tr>
<tr>
<td>Enrichment</td>
<td>1</td>
<td>2</td>
<td>1.00</td>
<td>0.5</td>
</tr>
<tr>
<td>Optimization</td>
<td>1</td>
<td>2</td>
<td>1.29</td>
<td>0.5</td>
</tr>
<tr>
<td>Flexible</td>
<td>1</td>
<td>2</td>
<td>1.00</td>
<td>0.5</td>
</tr>
<tr>
<td>Intelligence</td>
<td>1</td>
<td>2</td>
<td>1.29</td>
<td>0.5</td>
</tr>
<tr>
<td>Sustainable</td>
<td>1</td>
<td>2</td>
<td>1.00</td>
<td>0.5</td>
</tr>
</tbody>
</table>

In Table 2, the sentiment polarity, IDF, TF relative and TF absolute are each presented.

### III. RESULTS

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

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 results of the sentiment analysis are presented in Fig. 2.

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

<table>
<thead>
<tr>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agile</td>
<td>Security</td>
</tr>
<tr>
<td>Asset</td>
<td>Inefficient</td>
</tr>
<tr>
<td>Capability</td>
<td>Confusing</td>
</tr>
<tr>
<td>Competitive</td>
<td>Dark</td>
</tr>
<tr>
<td>Effective</td>
<td>Challenges</td>
</tr>
<tr>
<td>Enrichment</td>
<td>Failures</td>
</tr>
<tr>
<td>Optimization</td>
<td>Culture</td>
</tr>
<tr>
<td>Flexible</td>
<td>Liability</td>
</tr>
<tr>
<td>Intelligence</td>
<td>Complex</td>
</tr>
<tr>
<td>Sustainable</td>
<td>Waste</td>
</tr>
</tbody>
</table>
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 tests are presented in Table 4.

![TABLE 4. CLASSIFICATION ACCURACY](image)

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 Fig. 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.

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]. 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


