Enhancing Supply Chain Risk Management: Utilizing Latent Dirichlet Allocation and Big Data Analytics to Identify Disruptions from Unstructured Data

Hadley Rax

ABSTRACT –

Addressing gaps in Supply Chain Risk Management's use of Big Data, this paper investigates the potential of Latent Dirichlet Allocation (LDA), Sentiment Analysis, and Big Data Analytics to identify Supply Chain Disruptions from unstructured data. Employing mixed-method analysis of social media, we developed a predictive framework for disruption detection. Findings reveal efficient topic modelling and the potential for quicker identification of risks. Future work will focus on integrating real-time data streams to broaden the framework's industry applicability and resilience against diverse disruption types.

INTRODUCTION

Supply Chain Risk Management (SCRM) is the process of identifying, assessing, managing, and mitigating risks within the supply chain. This involves an approach of identifying risk in the everyday sense, as well as the ability to look for exceptional risks, leading to the process being one of continuous assessment to reduce the potential vulnerabilities. The risks to a supply chain are wide in their variety, disruptions could be caused through natural disasters, supplier issues, logistical problems, regulatory changes, cybersecurity threats, among others. The risks are identified, and assessed in their likelihood and significance, before strategies are used to remove or mitigate them, traditional approaches relied heavily on internal data, where as with recent developments there is the opportunity to reach beyond this.

As Supply chains across the globe are getting more and more complex, there is growing opportunities for disruptive events to affect them from initial sourcing through to the point of delivery. A major and growing cause of disruptions is climate or political related disruptions, examples such as globally, the Covid 19 pandemic shut down the logistics and freighting completely in some instances and diminished supply chains from manufacturing to final delivery in some capacity for a period extending years. Through these disruptions there has been shortages of goods, raw materials, drop in consumer confidence, as well as rise in prices leading to wastage and customer dissatisfaction. Through a review of relevant literature, it has been identified that Supply Chain Risk Management has been quite slow to adopt Big Data Analytics and AI technologies, while there is huge potential in benefits from these adoptions. As the outsourcing of manufacturers has diversified the potential for disruptions to the chains has increased, but through this there is the opportunity to build better systems and software to assist decision makers in how act proactively and reactively when confronting disruptions to their supply chains. There is a lot of unstructured data that could feed into different parts of the growing supply chains to assist in decision making. My research investigates ways for companies to identify these disruptive events using unstructured data from social media and sentiment analysis. This paper investigates the use Latent Dirichlet Allocation and topic trends, with a focus on twitter tweets, looking to increased trends in certain topics to identify disruptions to supply chains. To further enhance the topics, sentiment analysis is used to indicate the users emotions, and thus allocate the topics to positive or negative externalities of the disruptions.

Without the development of adequate predictive analytics, Supply Chain Management Systems will be left relying on ancient practises to attempt to mitigate and remove risks, while having little insight into major disruptive events, that could be identified through the mass of unstructured data on offer. This report looks to cover the relevant literature associated to Supply Chain Risk Management and the utilization of unstructured data in it's recent developments, pose the research questions relating to this paper, before defining the research methods, and carrying out an evaluation of the methods and their application. This report will finish by concluding on the effectiveness of this process, and potential further research work in this space. The importance of Supply Chain Risk Management, and the ability to get on top of these issues quickly for industries couldn't be minimised, beyond the shortages of resources and deliveries, it can also be considered on the inflationary pressures that have recently plagued the globe being partially caused by the mess of current supply chains, still reeling from the pandemic, the increasing climate disruptions, as well as the geopolitical instances, such as the Russia-Ukraine War.

LITERATURE REVIEW

The traditional definition of supply chain risks is "any risks for the information, material and products flows from original supplier to the delivery of the final product for the end user" given from (Juttner et al., 2003)[1]. Where as within (Dadfar et al, 2012) [2], the writers consider the focus of risk management to identify potential risk events and their interdependencies. Dadfar et al speak to from their timeline that extensive work is still required on how to behave in the case of major disasters, speaking to how at the time, 2012, there wasn't yet appropriate modelling tools applied into decision making or supporting systems available in the instances of large disruptive events to a supply chain.

Supply Chain Risk management is a critical aspect of modern supply chain operations, with the increase complexity and vulnerability of supply chains, the rise in global sourcing, lean production, and reliance on information technology has increase the productivity and profitability of supply chain while also leaving them far more susceptible to risks, from material shortages to information disruptions. In (Tang et al 2011)[3], draws attention to vivid examples of such disruptions in the Taiwan Earthquake 2006, and Ericsson's Crisis 2000. The Taiwan earthquake disrupted information flows, and had a significant affect on shipping operations due to this reduction in communication abilities. The Ericssons' fire was only confined to a single factory, but being a key supplier lead to a serious shortage illustrating the companies reliance on single sourcing.

Through events such as these, SCRM moved from a reactive process to a more proactive process, attempting to identify such risks and implement strategies to mitigate them before the damage is done, this was a development on the traditional approach of SCRM, and has seen a huge shift with the recent developments of Big Data and Machine Learning. Speaking to the challenges of traditional Supply Chain Risk Management, (Dadfar et al, 2012) [2] speak to how as supply chains have globalized, they have become more interconnected and exposed to risks such as disasters, market fluctuations, and operational disruptions, it further speaks to how the interconnectedness of these chains can cause for cascading effects. This research paper, written in 2012, preceded the popularity and inclusion in Big Data Analytics into Supply Chain software's, and as such does a good job of identifying some of the issues that BDA and unstructured data methods have come to solve. (Fan et al, 2015)[4] considers the integration of BDA to enhance the resilience of supply chains against disruptions and ways to improve risk management. Through a framework leveraging big data through multistage stochastic optimization, incorporating real time monitoring and forecasting of risks. This directly addresses limitations previously seen around how to address large scale disruptive events through scenario specific analysis, leading to a more dynamic and informed decision-making approach to its practise. The Stochastic optimization is used to deal with the uncertainty of the data in the scenario-based analysis, within which the model creates multiple plausible scenarios of future events that could impact the supply chain, with each assigned probability and potential costs. By structuring the optimization in stages, it allows decisions at each stage to consider the outcomes of previous stages and as such allows the strategies to adapt and develop dynamically as new information is made available.

Within (Li et al, 2015) [5] we see a different approach to using big data within Supply Chain Risk management though the use of pricing strategies to develop risk mitigation. The paper looks into the use of sensor data to monitor real time conditions of perishable goods, highlight real-time Big Data being used to dynamically adjust operations, mitigating the risks of product quality degradation. Expanding from purely internal structured data to public unstructured data within research paper (Singh et al, 2018) [16], using Twitter data to identify supply chain management issues within food, Signh et al utilized a Support Vector Machine and Hierarchical clustering with multiscale bootstrap resampling. This papers consider a beef supply chain across three weeks of data, using consumers views expressed on social media. This model worked well with smaller datasets, but a known issue with SVM are their high computational cost that struggles to scale up to meet growth.

In (Papadopoulos et al, 2017)[6] consider how unstructured data can be utilized in combination with traditional survey techniques to explore resilience in supply chains during disaster recovery. Through the use of large datasets of unstructured data from various social media platforms, in relation to the Nepal

earthquake of 2015, combined with survey data from managers involved in disaster relief. The researchers implement a mixed method approach, a unique method which integrates unstructured data analysis into a broader SCRM approach, allowing the analysis to aid in risk identification, assessment, and mitigation. A common theme in papers is just this, that unstructured data algorithms on their own lack insights and rely on the ability to work in conjunction with domain experts. This is another example of an innovative use and practical application of Big Data analytics within a real world disaster / disruption.
Within (Ganesh et al, 2022)[7] implement a text mining model using twitter data to identify risks, the study employs various sentiment analysis and association rule mining (ARM) to extract and analyse the Twitter data regarding supply chain disruptions. The shortcoming of this article along with many of the others comes down to the breadth of data, specifically the volume of tweets analysed, speaking to a larger dataset being able to provide a more robust analysis and insights from it. This article addresses a gap in the literature by incorporating ARM techniques not commonly used in SCRM, which could further enhance the understanding of risk factors and their interactions. This branches away from traditional reactive risk management approaches, positioning itself in a shift towards data-driven proactive SCRM strategies. With (Qomariyah et al, 2019)[8], (Yang et al, 2018)[9] we see considerations of Latent Dirichlet Allocation being used for topic modelling and sentiment analysis of twitter data, where the applications are of movie viewers sentiment [8], and the Media Centre of the Surabaya Government of Indonesia [9]. In both of these models we see them highlight the benefits of the model as powerful exploratory tools when used on large datasets, but do mention the limitations they personally experienced around only using plain text, English data and a low number of tweets [8].
Previous frameworks have been considered, such as with (Kara et al, 2023)[13] within this article the framework looks to utilize large amounts of data generated within supply chains to make informed proactive decisions around enchancing resilience, however this word within, speaks to collecting data from internally and is put into an example of the heavy machinery sector to validate their framework to assess and manage supplier risk. The conclusions of the papers speak to the need to continuous development of data mining, to keep pace with the dynamic risk factors faced within the industry. On the other side of the coin we see paper *(*Su et al, 2018)[15], developing a tool, Twitter Enabled Supplier Status Assessment, assisting companies in their global supplier selection process, built across four modules, Ontology Building, Extraction, Preprocessing and Classifier. This tool TESSA, assists companies in their decision making by narrowing down the potential suppliers through identifying their risks, and giving the selection team more knowledge around the potential suppliers.
Similar to this *(*Sadeek et al, 2023) [14] too considers data mining and further looks to twitter data relating to the Ukraine Russia War within February 2022, and delves into leveraging EMOLEX, a NRC Intensity Lexicon Library, which is a dictionary based NLP library, regularly updated with sentiments and emotions of specific English words. Again this article speaks to the size of their dataset being a little undersized, as well as the drawbacks of Sentiment analysis, mainly that of the binary set up used in this paper, where it is either positive or negative, leading to a compressed understanding of human expression.
(Schmidt et al, 2020)[18] takes a different angle, analysing supply chain glitches across firms and how Twitter responds to these glitches, as such over 2 Billion tweets were leveraged in the study, finding significant reactions on Twitter after Supply Chain glitches, while this differs from the questions of this paper, it does present an interesting approach, paired with an interesting approach to overcoming simultaneity bias, a bias which arises within a relationship when each variable influences the other simultaneously.
To be able to further the field of research around SCRM and being able to create insights from unstructured data, it is important to be able to measure and identify supply chain disruptions, there have only be a few true iterations of this, one of the most esteemed measures is the Global Supply Chain Pressure Index (GSCPI) built by Gianluca Benigno, Julian di Giovanni, Jan J.J. Groen, and Adam I. Noble of the Federal Reserve Bank of New York.Built in the aftermath of the Covid-19 pandemic this model was built to on the insights of how exogenous factors could influence supply chains while climate, and geopolitical developments could also cause major disruptions.
Within *"The GSCPI: A New Barometer of Global Supply Chain Pressures"*(Benigno et al, 2022)*[20]*, a process is established to capture pressures that arise at the global supply chain level, assessing it's capability and capacity to explain inflation, measuring the established framework against recent producer price inflation within the United States and euro area.
The GSCPI was developed as a gauge of supply constraints with respect to economic outcomes and has further developed the tool to show associations with goods and producer price inflation in the United States and the euro area. Metrics of the Index provide a comprehensive summary of potential disruptions and take the Baltic Dry Index (BDI) among other metrics into the development of the GSCPI, where the Baltic Dry Index is an index of average prices paid for

transportation of dry bulk materials. The Baltic Dry is considered a leading indicator of supply chain inflation, economic activity and is a purely supply and demand driven metric, however, it is extremely volatile and narrowly focused, thus the GSCPI expands beyond this to encompass Manufacturing indexes, delivery times and backlogs of orders, air and ocean freight costs, and inventory levels. Using real time data and geographical spread it integrates data from major economies, this leads to greater stability and less volatility than the Baltic Dry Index.

4 articles on sentiment analysis

Looking into specific ways that sentiment analysis is combined with the topic modelling of tweets through Latent Dirichlet Allocation is found in

RESEARCH QUESTION

Within the context of Supply Chain Risk Management, how can unstructured data be utilized through Latent Dirichlet Allocation and Sentiment Analysis to identify disruptions to supply chains?

Through the literature review there was highlighted gaps in Supply Chain Risk Managements utilization of Big Data Analytics, this paper is looking specifically at ways to further it's ability to use unstructured data to identify disruptions. The aim of this research is to develop a framework to identify gaps using unstructured data, and looks to identify topic trends to show how this data can be used to more quickly identify disruptions to supply chains.

RESEARCH DESIGN / METHODS

Within this paper the area of focus was utilizing existing unstructured data from X(formally known as Twitter), and refining this to include data relating to supply chain related terminology.

I used applications of the following python libraries relating to data manipulation, natural language processing, statistical visualization, topic modelling.

Python Library	**Description**
Pandas	A powerful data manipulation and analysis library,
NLTK (Natural Language Toolkit)	A suite of libraries and programs used in classification, tokenization, parsing and semantic reasoning within statistical language processing.
Matplotlib	A plotting library used for static, interactive visualizations within python.
Seaborn	A statistical visualization library built on Matplotlib.
GenSim	A robust library designed for unsupervised topic modelling and natural language processing.
pyLDAvis	A library for the interpretation and visualization of topic modelling algorithms, often used in conjunction with GenSim

The research aims to see how supply chain disruptions over a period of the covid 19 pandemic can be identified effectively through the use of Latent Dirichlet Allocation of unstructured data. Within the python script the data is initially loaded, in from the csv, subset and renamed for the relevant purposes. Within this the subset is identified as a dataframe 'ss_df2', which filters to include the attributes user_location, date, text, hastags, removing other information not necessary to the processing of this information.

Keywords are created which relate to the supply chain topics of interest, and are used to filter the tweets to the topic of supply chain disruptions. In the wider context this is purely usedto reduce the computational expense, but this has the potential to reduce hidden relationships and insights.

Text Pre-processing - Within Natural language processing, preparation of the text data for analysis requires steps to ensure the text is more manageable and meaningful for algorithms, and as follows the use of the data for use within an Latent Dirichlet Allocation (LDA) model it must first be processed into a usable format for the model. Within this there is the removal of URLs from the text of the tweets, tokenization and lemmatization of the text and removal of stop words establish in the text.

Tokenization involves the splitting of a string of text into a list of tokens, which can be considered of as parts like words. This works by location word boundaries such as punctuations of whitespaces, this is useful in it allows the LDA model to work at a word / token level as opposed to at a sentence level, giving the opportunity to build sentiments based on this. An essential first step in processing the data, without tokenization, the text analysis would not be feasible. Lemmatization involves reducing a word to its base form of *lemma*. An example of this is run to running. This differs from stemming as it takes into consideration the morphological analysis of the word, whereas stemming merely chops the word at inflection. The process of Lemmatization in the instance of this project is used to reduce the words to their base, improve the models accuracy by consolidating different inflected forms of a word. Once Tokenization and Lemmatization are completed on the dataset, we can instigate stop word removal, which involves removing common words of little value to the analysis, examples of stop words would be “and”, “the”, “is”, “is. This is expanded to include additional highly repeated tokens that were found to be of no value within the initial model building.

Overall these three steps streamline the text data, reduce the computational complexity and improve the performance and accuracy of the Natural Language Processing models.

Analysis & Visualization - Through the analysis of the subset using NLTK’s FreqDist we can identify the most common words of the filtered tweets, as a check to whether the words are meeting our expectation and that there aren’t more fragments that should be added to the stop words list. Build Bag of words using Gensim ‘Dictionary’ created from the tokenized texts, and then creates a Bag of Words model. The bag of words is a often used method in natural language processing to represent text data. Within this model, the text tokens are represented as a bag of words of the text of the initial dataset. Bag of Words transforms text data into numerical features, by treating the text as unordered collections of words, it is effective at the document classification required in LDA. BOW also scales well with the size of the corpus, and as such makes it feasible as we move to processing larger and larger datasets.

Latent Dirichlet Allocation (LDA) - LDA is a probabilistic topic model, that assumes the documents are composed of topics that are distributions over words. A generative model, it posits a probabilistic procedure by which documents can be generated. Latent Dirichlet Allocation (Blei, 2003)[10] sets out the understanding of the LDA model that is adhered to in this paper. In our example, the Tweets are considered the documents, and the outline of the steps are as follows The Model follows the following steps:

1. Random Initialization – Assigning each word in each document a random topic, the initial assignment gives both topic representation of all documents and word distribution of the topics.
2. Iterative Refinement – Using Gibbs sampling to improve the assignments based on two principals. The words in the same document should share the same topics, and documents with similar topics should share similar distributions of words.
3. Final Distributions – After several iterations, the model arrives at a good approximation of the distributions of topics in each document, and the distribution of words in each topic.

The Logic and Formulation of Latent Dirichlet Allocation

Dirichlet Distribution for Topics in Documents

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1}$$

Where θ is the topic distribution of a document and α is the hyperparameter influencing the shape of the distribution.

Topic Assignment for each word.

$$p(z_n = k \mid \theta) = \theta_k$$

Where z_n represents the topic for the nth word, and k represents a specific topic.

Word Distribution Given Topics

$$p(w_n \mid z_n = k, \beta) = \beta_{k,w_n}$$

Where β_{k,w_n} is the probability of word w_n under the topic k, as defined by the distribution parameterized by β

Following the above definitions the objective of a Latent Dirichlet Allocation is to compute the posterior distribution of hidden variables given an object

$$p(\theta, z \mid w, \alpha, \beta) = \frac{p(\theta, z, w \mid \alpha, \beta)}{p(w \mid \alpha, \beta)}$$

Latent Dirichlet Allocation models benefit from being able to produce highly interpretable topics depending of its hyperparameters and the coherence of the data. It allows for the discovery of potential hidden themes, which may otherwise not be apparent, and expands the potential of exploratory data analysis. Finally, I associate the tweets to topics, by assigning them to their most probable topic based on the LDA model, I am able to dictate trends over time and show over a time series how the different topics peak in different areas. The combination of the methods listed above provides a robust approach to analysis of complex textual data. Through the quantification of unstructured data as stated, we are able to build insights on big data.

With [8] and [9] we see considerations of Latent Dirichlet Allocation being used for topic modelling and sentiment analysis of twitter data, where the applications are of movie viewers sentiment [8], and the Media Centre of the Surabaya Government of Indonesia [9]. In both models we see them highlight the benefits of the model as powerful exploratory tools when used on large datasets but do mention the limitations they personally experienced around only using plain text, English data and a low number of tweets [8].

Sentiment Analysis - Following the Latent Dirichlet Allocation and analysis of topics, sentiment analysis can provide additional insights into the emotions or opinions expressed in tweets associated with the topics. Vader from NLTK is a pre-trained sentiment analysis model specifically tuned for social media text, and as such is an appropriate model to use in this circumstance. The Sentiment is analysed across the topics by calculating the average sentiment score per topic. While dealing with large amounts of public data, there are ethical considerations that must be considered. With topic modelling there can be the inadvertent perpetuation or amplification of biases through the data taken to train the model. This can be overcome through the awareness of the how the biases in the data can form, and thus through ensuring a diverse range of voices are involved in the data, so a solid mix of representation of different groups, opinions, sentiments.

Correlation & Regression Analysis - Following the Topic and sentiment analysis it's important to analyse the results, this is compared against the Baltic Dry Index Historical Data, in this case, the initial dataset requires pre-processing to convert the date to a time-series format, for the comparison, involving resampling against the weeks, allowing for the taming of the fluctuations of the initial index data. This data is merged against the topic trends, on which a Pearson correlation analysis and regression analysis is performed to provide understanding on how the textual narratives of the topic trends can correlate and potentially predict the economic indices of Supply Chain values.

APPLICATION OF RESEARCH METHODS

Initially the pilot test of this study was going to cover a specific event of disruption, ie the Auckland Floods of January 2023, a known localized event resulting in large scale disruption, effecting both sides of supply chains in terms of output from manufacturing and delivery to consumers. This was intended to be done through the ability to pull tweets from X's (formally known as Twitter) API 2, within which there would be some conditional filters on the hashtag relating to supply chains and deliveries, filters on the text to pick up on certain key words, filters on locations surrounding Auckland, as well as on the dates pushing either side of the event. However, beyond the payment to access X's API, there is further need to get academic permissions to pull historic tweets from X. This is a recent development, and beyond waiting the last couple months to get approval this has not occurred.

Thus the papers focus switched to the a similar large dataset , which happened to be a covid19_tweets.csv [19] containing 179,108 tweets, which contained data from a time period between 2020-07-24 and 2020-08-30, these covered multiple locations, but further refinement to a single area over a longer stretch of time, would be ideal in further iterations of the study. Through the integration of the above study, I did lose the ability to truly identify disruptions from events as the inability to generate data for the event in question and give it usable locations led to the reduction in ability of the data to identify supply chain disruptions in a larger scale. However, from the dataset, I was able to generate the Latent Dirichlet Allocation, assign the tweets to the most probable topics, and then build timeseries of the topics trends over time.

Below in the Tweet Length Distribution, this is occurring after the removal of URLs, the tokenisation and lemmatization of the tweets into a usable format. From this dictating the length of the tweets shows us the distribution approximately a bellcurve.

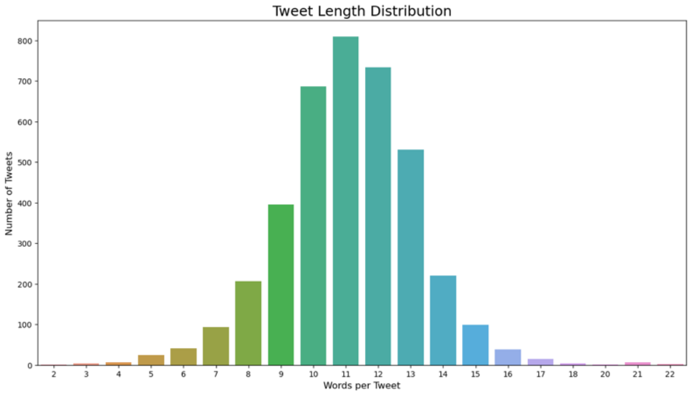


Figure One. Tweet length distribution from Covid19_tweets dataset.

The Intertopic Distance Map (from pyLDAvis library,) allows for the visualization of the dissimilarities between the different topics derived from the dataset. Within this Principal Component Analysis' dimensionality reduction technique is used reduce the variability of the data into a interpretable format.

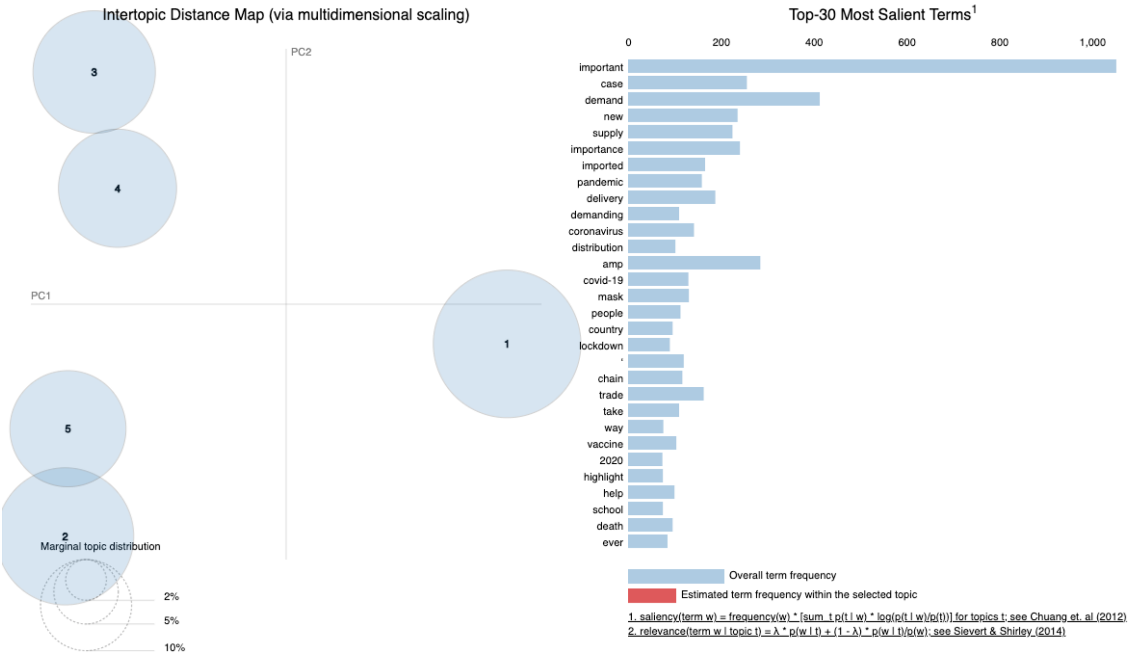


Figure Three. Intertopic Distance Map and Salient Terms table through pyLDAvis, with k=5

From the two settings of topics, k=5 is identified as a more succinct parameter to reduce the crossover of topics, giving us clearly defined between the attributes of the principal component analysis. We also get to see the best distinction of terms with k=5, from this we see the following relevant tokens within the topics:

[Topic 1] Global Trade & Covid Impact – covid, supply, imported, chain, trade, global, demand, etc.

[Topic 2] Public Health & Covid – covid19, health, people, etc.

[Topic 3] Covid's Economic Impact - demand, covid, delivery, trade, etc.

[Topic 4] Covid & Lockdown Effects - demand, importance, covid, lockdown, service, etc.

[Topic 5] Supply Chain Dynamics - distribution, shipping, supplying, important, covid19, etc.

As this dataset is closely related to covid 19 tweets, these topics are skewed by this attribute, but we can see differences between topics like two, that we imagine to be related to tweets regarding peoples health relating to Covid19., and five, which appears to be speaking quite clearly to supply chain issues regarding Covid19.

From this insight, when we build the trends below to reflect the number of tweets relating to the different topics, it means that within the insights we can identify on the below plots from the last plot, associated with Topic Five that there is a peak at approximately 18th August, before the trend against all the topics of heading to zero towards 30th August.

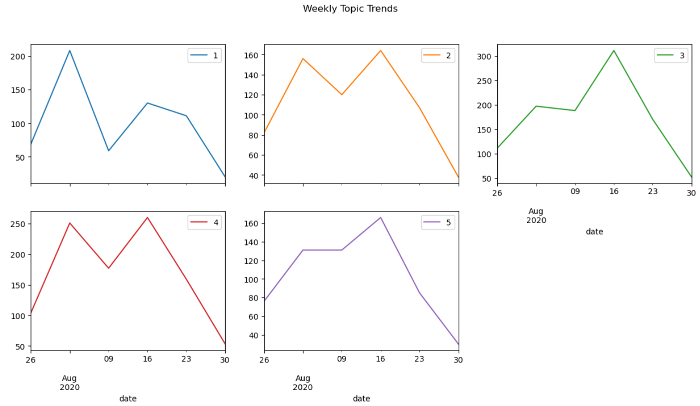


Figure Four Weekly Topic Trends across k=5 distribution

Thus from the pilot study we see peaks congregating on the first of August approximately and the 18th and 30th of August approximately, with the different topics having slightly different patterns against the dates.

In the calculation of the sentiment scores by topics we can see following allocation between the topics. Where VADER (Valence Aware Dictionary and Sentiment Reasoner) scores -1 as extremely negative, 0 as neutral and 1 as extremely positive. We can therefore see that topics 1, 2, and 3 sit pretty neutrally, where as topic 4 and 5 sit with a mean of around 0.25, indicating that they are slightly positive. Which given the context of the dataset, Covid19_tweets, and the context of the topics (Topic4 – Covid and Lockdown effects, Topic5- Supply Chain Dynamics) one wouldn't assume that the tweets at that period would be positive.

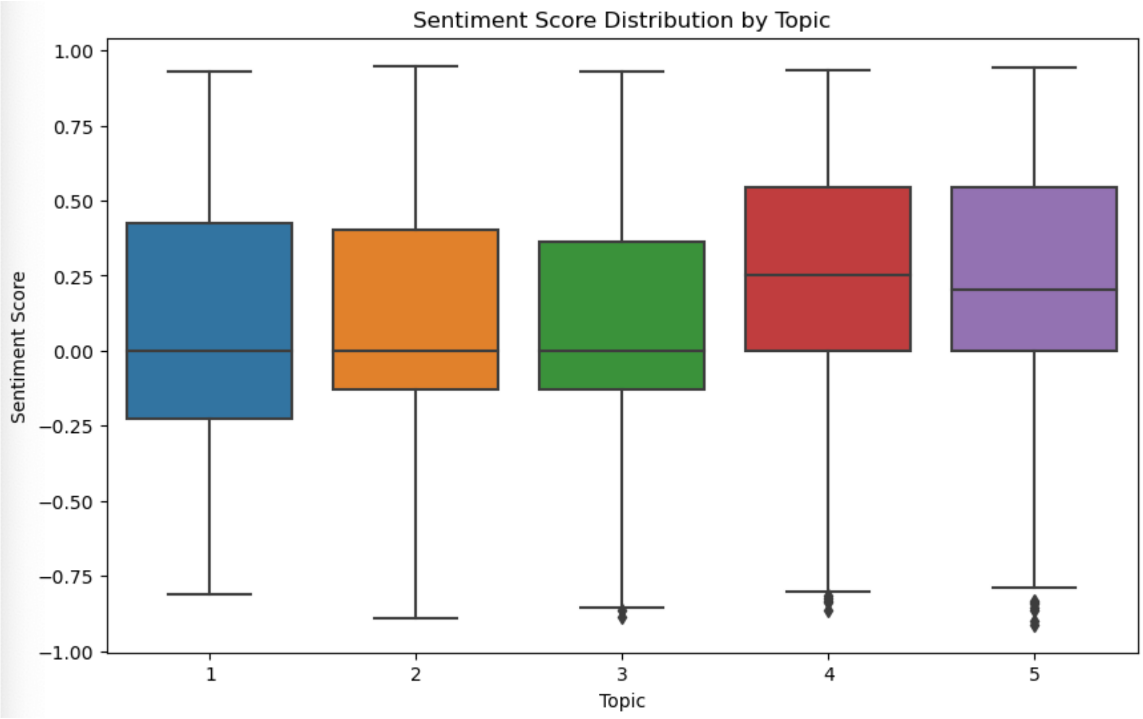


Figure Five. Sentiment Score Distribution by Topic, with k=5.

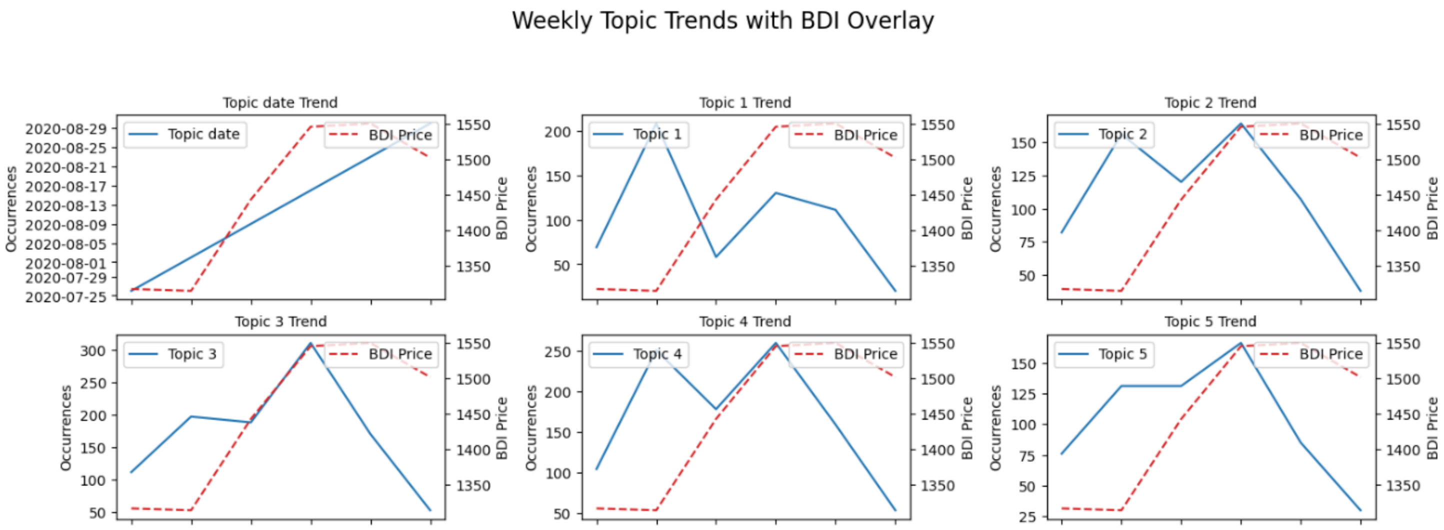
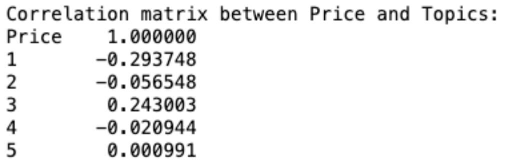
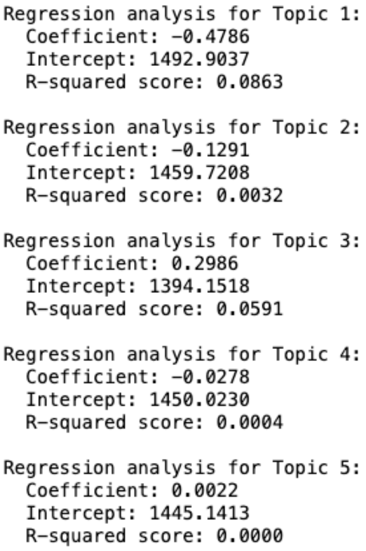


Figure Six. Plotting Topic Trend Occurrences against BDI Weekly Price Aggregation.



```
Correlation matrix between Price and Topics:
Price    1.000000
1       -0.293748
2       -0.056548
3        0.243003
4       -0.020944
5        0.000991
```

Figure Seven. Correlation Matrix between BDI Weekly Price Aggregation and Topic Trend Occurrences



```
Regression analysis for Topic 1:
  Coefficient: -0.4786
  Intercept: 1492.9037
  R-squared score: 0.0863

Regression analysis for Topic 2:
  Coefficient: -0.1291
  Intercept: 1459.7208
  R-squared score: 0.0032

Regression analysis for Topic 3:
  Coefficient: 0.2986
  Intercept: 1394.1518
  R-squared score: 0.0591

Regression analysis for Topic 4:
  Coefficient: -0.0278
  Intercept: 1450.0230
  R-squared score: 0.0004

Regression analysis for Topic 5:
  Coefficient: 0.0022
  Intercept: 1445.1413
  R-squared score: 0.0000
```

Figure Eight. Regression Analysis between BDI Weekly Price Aggregation and Topic Trend Occurrences

Through the correlation matrix and topics, it can be observed a negative correlation between BDI price and topics 1, 2, and 4, a weak positive correlation for BDI price and topic 3 and almost no correlation between topic 5 and BDI price. Within topics 2 and 3 have weak correlations, where topic one is a moderate correlation. Indicating that with Topic 1 as the topic occurrence increase, the BDI price tends to decrease, and is the strongest relationship of the five topics.

The regression analysis attempts to model the relationship between the Topics and the BDI Price Aggregation, from the output

Topic 1 shows a moderate negative impact on BDI Price, with a small but more significant proportion of variability explained, (8.63% of the variability in Price.

Topic 2 and Topic 4 show minimal impact on Price, with very little of the variation in Price explained by these topics.

Topic 3 has a modest positive impact on Price, explaining about 5.91% of the variability.

Topic 5 despite having a positive coefficient, essentially shows no real impact on Price, with no variability explained.

This suggests Topic 1 has the most significant influence on the BDI Price among the topics, albeit still moderate.

EVALUATION OF RESEARCH METHODS

In evaluation of the research method and the application onto the pilot of the study, we can identify shortcomings in the output. The Covid19_tweets data doesn't trend into periods of disruptions beyond the peaks of 18th and 30th August, this can be attributed to the dataset being too small and too short in it's collection. This paired with the location being broad means that these peaks attribute to global phenomena of supply chain – covid19 related phenomena. But it must be noted that these peaks have occurred once the data has been filtered against key words, as such these keywords apply the shape to the distribution of the data.

The above shortcomings come from a lack of decent datasets relating to this, without gaining access to X's API to pull data, the only option was to try and find the closest fit for the pilot study. But to step away from the quality of the data, further work could be research could be completed to expand the data from just twitter to include a selection of reputable global newspapers, to adjust the extraction, prioritization, and tokenization, potentially using just the headlines of articles could be enough to scrape to provide insights on upcoming disruptions.

Outside of the quality of the data, the effectiveness of this approach to identify potential disruptions and map these through Latent Dirichlet Allocation is still evident. This method is well suited to addressing the complex dynamics of supply chain disruptions influenced by external events such as Covid-19 highlighted in the write up above. This method can capture a broad spectrum of impacts from disruptions and is able to give the user insights in the groupings of the topics, highlighting the potential for users to gain insights into groupings of topics that they may have otherwise missed. However, while the method provided valuable insights into the public reactions to Covid 19 and the filtered dataset relating to supply chain disruptions, it encountered limitations in depth due to the variability and the limit in the availability of relevant unstructured data.

Through the incorporation of sentiment analysis to the topic model, we are able to further understand the connotations of the users' emotions when they enter these tweets, being able to separate the positive supply chain related events to the negatives. This further speaks to another avenue of research which would be looking into the division of tweets into delivery related supply chain tweets and how this can be separated from the broader context of supply chain related information, relating to climate, resource scarcity and geopolitical actions.

Through the correlation and regression analysis, Topic 1 points to the most significant influence, while also highlighting the expectation of a negative influence, as expected through the

Compared to traditional methods mentioned in the literature review, this approach allows for much more dynamic situations if the user has the time to adjust the keywords to fit their opportunity, in fact with larger datasets and a more established bank of keywords, this updating would be a far less regular occurrence. It must be noted that the single source of data, X, for this model means an introduction of bias towards not only internet-active demographics, but also events where there is access to internet, thus in the case of a power cut, this method in its current form would miss the opportunity to identify the disruption. The limitation mentioned above, do not detract from the significance of the study, which has the potential to identify critical points of failure in supply chains, further research should consider expanding the data sources to include more data sources from news media, as well as potential inputs from major supply chain participants. Further refinements to data filtering techniques to better manage the scale and variability of unstructured data could also assist in the ability to identify these disruptions.

CONCLUSION

This study has demonstrated the significant potential of integrating Latent Dirichlet Allocation (LDA) and Big Data Analytics to enhance Supply Chain Risk Management (SCRM) by utilizing unstructured data from social media platforms. Through meticulous analysis of data derived from Twitter, the research has established a robust predictive framework capable of identifying disruptions in supply chains with greater speed and efficiency. This approach not only leverages technological advancements but also addresses critical gaps in traditional SCRM practices, which have typically relied more on internal data and less on the dynamic, real-time insights that social media can offer. The findings underscore the importance of incorporating advanced data analytics into SCRM to better anticipate and react to disruptions. Future research should aim to expand the data sources and perhaps integrate real-time data streams from various social media platforms to enhance the framework's applicability across different industries. Additionally, there is a need to refine the models used to manage the diversity and complexity of data, which can further improve the accuracy and effectiveness of disruption detection.

By advancing these methodologies, businesses can not only safeguard against potential risks more effectively but also gain a strategic advantage in managing their supply chains in an increasingly volatile global market. The continued development and refinement of these analytical tools will play a pivotal role in shaping resilient, agile, and responsive supply chains for the future.

There is much room for further research in this area with it being such a new field of research and the clear lack of previous work already done in this space. Getting access to larger and more relevant twitter datasets would be a great first step, and from that being able to expand to the different news media from specific areas to try and get a more granular approach to this model. In it's current iteration it suffered from being unable to get specific data, and as such relied on the largest data set it could find, which unfortunately lacked the proper level of specificity to region as well as topic. Another avenue of interest, would be once better data is collected, being able to link it to the Federal Bank of New York's' Global Supply Chain Pressure Index, and even looking into ways that this index could be localised to smaller areas, in this way, the pressure index could look to a model such as this, creating supply chain insights from unstructured data, as an early indicator of potential pressures and stressors on the supply chain pressure index.

REFERENCES

[1] Jüttner, Uta & Peck, Helen & Christopher, Martin. (2003). Supply Chain Risk Management: Outlining an Agenda for Future Research. International Journal of Logistics : Research & Applications. 6. 197-210. 10.1080/13675560310001627016.

[2] Dadfar, D., F. Schwartz, and Stefan Voss. "Risk management in global supply chains–Hedging for the big bang." In: Mak, H.Y.,Lo,H. (eds.) Transportation and Logistics Management Proceedings of the 17th International HKSTS Conference (HKSTS 2012), pp. 159-166. Hong Kong (2012), https://www.researchgate.net/publication/259150219_Risk_management_in_global_supply_chains_-_Hedging_for_the_big_bang

[3] Ou Tang, S. Nurmaya Musa, Identifying risk issues and research advancements in supply chain risk management, International Journal of Production Economics, Volume 133, Issue 1, 2011, Pages 25-34, ISSN 0925-5273, https://doi.org/10.1016/j.ijpe.2010.06.013.

[4] Fan, Y., Heilig, L., Voß, S. (2015). Supply Chain Risk Management in the Era of Big Data. In: Marcus, A. (eds) Design, User Experience, and Usability: Design Discourse. Lecture Notes in Computer Science(), vol 9186. Springer, Cham. https://doi.org/10.1007/978-3-319-20886-2_27

[5] Li, D., & Wang, X. (2015). Dynamic supply chain decisions based on networked sensor data: an application in the chilled food retail chain. International Journal of Production Research, 55(17), 5127–5141. https://doi.org/10.1080/00207543.2015.1047976

[6] Thanos Papadopoulos, Angappa Gunasekaran, Rameshwar Dubey, Nezih Altay, Stephen J. Childe, Samuel Fosso-Wamba, The role of Big Data in explaining disaster resilience in supply chains for sustainability, Journal of Cleaner Production, Volume 142, Part 2, 2017, Pages 1108-1118, ISSN 0959-6526, https://doi.org/10.1016/j.jclepro.2016.03.059.

[7] Deiva Ganesh, A. and Kalpana, P. (2022), "Supply chain risk identification: a real-time data-mining approach", Industrial Management & Data Systems, Vol. 122 No. 5, pp. 1333-1354. https://doi.org/10.1108/IMDS-11-2021-0719

[8] Siti Qomariyah, Nur Iriawan, Kartika Fithriasari; Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis. AIP Conf. Proc. 18 December 2019; 2194 (1): 020093. https://doi.org/10.1063/1.5139825

[9] Yang, Sidi, and Haiyi Zhang. "Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis." International Journal of Computer and Information Engineering 12.7 (2018): 525-529.https://www.researchgate.net/publication/335106801_Text_Mining_of_Twitter_Data_Using_a_Latent_Dirichlet_Allocation_Topic_Model_and_Sentiment_Analysis

[10] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research. 2003; 3(4–5): 993–1022.

https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=http://githubhelp.com

[11] Wiegmann, Matti & Kersten, Jens & Klan, Friederike & Potthast, Martin & Stein, Benno. (2020). Analysis of Detection Models for Disaster-Related Tweets. 10.5281/zenodo.3713920.

[12] Merve Er Kara, Seniye Ümit Oktay Fırat, Abhijeet Ghadge, A data mining-based framework for supply chain risk management, Computers & Industrial Engineering, Volume 139, 2020, 105570, ISSN 0360-8352, https://doi.org/10.1016/j.cie.2018.12.017.

[13] Merve Er Kara, Seniye Ümit Oktay Fırat, Abhijeet Ghadge, A data mining-based framework for supply chain risk management, Computers & Industrial Engineering, Volume 139, 2020, 105570, ISSN 0360-8352, https://doi.org/10.1016/j.cie.2018.12.017.

[14] Sadeek, S.N., Hanaoka, S. Assessment of text-generated supply chain risks considering news and social media during disruptive events. Soc. Netw. Anal. Min. 13, 96 (2023). https://doi.org/10.1007/s13278-023-01100-0

[15] Chuan-Jun Su, Yin-An Chen, Risk assessment for global supplier selection using text mining, Computers & Electrical Engineering, Volume 68, 2018, Pages 140-155, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2018.03.042.

[16] Akshit Singh, Nagesh Shukla, Nishikant Mishra, Social media data analytics to improve supply chain management in food industries, Transportation Research Part E: Logistics and Transportation Review, Volume 114, 2018, Pages 398-415, ISSN 1366-5545, https://doi.org/10.1016/j.tre.2017.05.008.

[17] Bongsug (Kevin) Chae, Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research, International Journal of Production Economics, Volume 165, 2015, Pages 247-259, ISSN 0925-5273, https://doi.org/10.1016/j.ijpe.2014.12.037.

[18] Christoph G. Schmidt, David A. Wuttke, George P. Ball, Hans Sebastian Heese, Does social media elevate supply chain importance? An empirical examination of supply chain glitches, Twitter reactions, and stock market returns, Journal of Operations Management, March 2020, 10.1002/joom.1087

[19] Covid19_tweets.csv, 68.71 MB, https://www.kaggle.com/datasets/gpreda/covid19-tweets

[20] Benigno, Gianluca and Benigno, Gianluca and di Giovanni, Julian and Groen, Jan J. and Noble, Adam I, The GSCPI: A New Barometer of Global Supply Chain Pressures (May 2022). FRB of New York Staff Report No. 1017, Available at SSRN: https://ssrn.com/abstract=4114973